A Zero Trust Framework for Realization and Defense Against Generative AI Attacks in Power Grid

Md. Shirajum Munir[1], Sravanthi Proddatoori[2], Manjushree Muralidhara[2], Walid Saad[3], Zhu Han[4], and Sachin Shetty[5]
[1]School of Cybarsecurity, [2]Dept. of CS, [5]Dept. of ECE, Old Dominion University, Norfolk, VA 23529, USA
[3]Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA
[4]Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA
Email: mmunir@odu.edu; sprod002@odu.edu; mmura001@odu.edu; walids@vt.edu; hanzhu22@gmail.com; sshetty@odu.edu

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622574

***Abstract*—Understanding the potential of generative AI (GenAI)-based attacks on the power grid is a fundamental challenge that must be addressed in order to protect the power grid by realizing and validating risk in new attack vectors. In this paper, a novel zero trust framework for a power grid supply chain (PGSC) is proposed. This framework facilitates early detection of potential GenAI-driven attack vectors (e.g., replay and protocol-type attacks), assessment of tail risk-based stability measures, and mitigation of such threats. First, a new zero trust system model of PGSC is designed and formulated as a zero-trust problem that seeks to guarantee for a stable PGSC by realizing and defending against GenAI-driven cyber attacks. Second, in which a domain-specific generative adversarial networks (GAN)-based attack generation mechanism is developed to create a new vulnerability cyberspace for further understanding that threat. Third, tail-based risk realization metrics are developed and implemented for quantifying the extreme risk of a potential attack while leveraging a trust measurement approach for continuous validation. Fourth, an ensemble learning-based bootstrap aggregation scheme is devised to detect the attacks that are generating synthetic identities with convincing user and distributed energy resources device profiles. Experimental results show the efficacy of the proposed zero trust framework that achieves an accuracy of 95.7% on attack vector generation, a risk measure of 9.61% for a 95% stable PGSC, and a 99% confidence in defense against GenAI-driven attack.**

I. Introduction

Power grid supply chain (PGSC) cybersecurity is necessary to the infrastructure that provides electrical power to homes, businesses, and critical facilities. The PGSC infrastructure is expected to deploy around 30-40 billion distributed energy resource (DER) devices such as renewable energy sources, consumers, prosumers, generators, electric vehicles (EV), EV charging stations, and so on by 2025 to meet an envisioned 40% energy cost reduction by 2050 [1]–[4]. The rigorous expansion of diversified DERs brings indispensable cyber challenges for power grid operations [1]–[3], [5] by creating a large surface. Additionally, artificial intelligence (AI) can induce adversarial attacks on PGSC [6]–[8].

Generative artificial intelligence (GenAI) models such as generative adversarial networks (GAN) [9]–[12] offer significant benefits in data augmentation and reconstruction. Therefore, GANs can expand the of cyber attack vectors in the power grid by generating synthetic identities with convincing user and DER device profiles [6], [8]. In particular, GenAI can create new attack vectors for launching replay attacks by generating observed control message parameters such as the reaction time of participants, nominal power consumed, price elasticity coefficient [13] and their pattern from the trusted DERs. GenAI can also imitate the broadcast data distribution of DERs such as communication data packet, packet size, IP, port, demand-response energy data, and so on for introducing protocol-type attacks in PGSC [14]. These types of attack vectors have not been included in DER security standard IEEE 1547 [15]. Clearly, advances in GenAI can lead to *novel attack surfaces* that, in turn, introduce new vulnerabilities and risks to the power grid, which can potentially lead to *1) unauthorized parties gaining access due to de-synchronized control and communication messages by protocol attack, and 2) power outage, energy theft, and money fraud are caused by replay attacks on nominal power consumed and price elasticity coefficient of DERs.*

In order to defend against these new vulnerabilities, it is essential to address several unique challenges that include:

- Generation of potential attacks that can be created by GenAI in order to understand the potential vulnerabilities in advance.
- Design of tail and risk-based reliability measure and trust metrics to analyze the worst-case vulnerabilities of various energy DERs control and communication messages for low latency recovery, and adaptation of energy grid behavior changes.
- Moving from classical *trust and verify approaches* into a zero-trust regime built on the paradigm of *never trust and always verify* which effectively identify, explain, and defend any disrupted events carried on by GenAI in PGSC.

The main contribution of this paper is to address the above

This work is supported in part by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory, National Science Foundation under Grant No. 2219742 and Grant No. 2131001, the Office of Naval Research (ONR) MURI Grant N00014-19-1-2621, VIRGINIA INNOVATION PARTNERSHIP CORPORATION Grant No 230849, the Commonwealth Cyber Initiative under contract number HC-3Q24-049, an investment in the advancement of cyber R&D, innovation, and workforce development.

TABLE I: Summary of notations.

Notation	Description
$\mathcal{I}$	A set of DRE
$q_i(t)$	Power (i.e., +ve for generator, -ve for consumption)
$\boldsymbol{x}_{it}$	Control/status message
Θ_i	Rotor angle
β_i	Damping constraint
α_{ij}	Coupling strength between i and j
$\hat{q}_i$	Power
Φ_i	PGSC market elasticity
τ_i	Response delay
G_θ	Generator
D_ϕ	Discriminator
$\eta \in (0,1)$	CVaR significant probability
ξ	CVaR confidence level

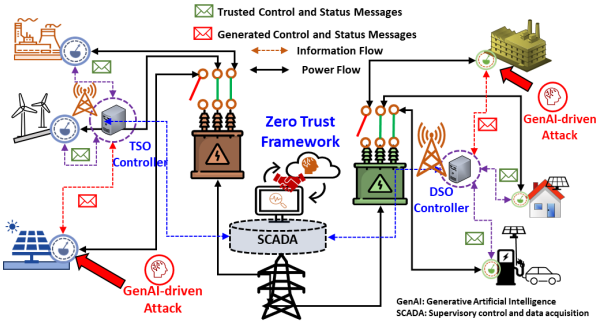


Fig. 1: A system model of a zero trust framework for risk realization and defense against GenAI attacks on the PGSC.

technical challenges by proposing a *zero trust framework* for risk measuring and defense against GenAI-driven attacks on the PGSC. Towards developing this framework, we make the following key contributions:

- We design a new zero trust system model of PGSC and formulate a joint optimization problem for generating novel attack surfaces, measuring risk, and defense against the generated control/status message of DERs.
- We develop a domain-specific GAN mechanism for potential vulnerability creation. Here, the main novelty is the capability of generating new attack vectors for further understanding by modeling generative adversarial networks for generating synthetic identities that convincingly mimic the device profiles of legitimate users and DER device profiles.
- We develop tail-based reliability metrics for realizing the risk of potential attack. Then, we propose a trust quantification approach for continuous validation on understanding the underlying risk of DERs'.
- We devise a defense strategy for GenAI-driven attacks on PGSC by leveraging an ensemble learning method (i.e., a bootstrap aggregation (bagging) mechanism) for solving a random forests (RF) regression problem.
- The performance of the developed zero trust framework is validated by leveraging two state-of-the-art PGSC datasets. Our experimental analysis shows that the proposed zero trust framework can successfully generate control/status (about 95.7%), quantify extreme risk (around 9.61%) for PGSC stability parameters with a 95% confidence (trust), and achieve around 99% accuracy for GenAI-driven attacks detection on PGSC.

II. System Model for Realizing GenAI-Driven Attacks in PGSC

We consider a power grid supply chain equipped with a set $\mathcal{I}$ of I DERs such as generators, consumers, and prosumers (as seen in Figure 1). In our system, we consider finite, continuous time, such that each time slot $t \in (0, T)$. Therefore, at time slot t, each DER $i \in \mathcal{I}$ can generate $q_i(t)$ (i.e., $q_i(t)$ is a positive value) or consume $q_i(t)$ (i.e., $q_i(t)$ is negative value) power. In this PGSC, supervisory control and data acquisition (SCADA) systems monitor and orchestrate the power grid operation while the transmission system operator (TSO) and distribution system operator (DSO) assist in transferring and distributing energy, respectively. In particular, TSO brings energy from production to the main grid while DSO distributes it to the end users such as consumers.

At the time t, each DER $i \in \mathcal{I}$ can exchange SCADA control and status message $\boldsymbol{x}_{it}$ with the SCADA system. Consequently, DER $i \in \mathcal{I}$ can send and receive control message vector $\boldsymbol{x}_{it} = (a_{it}, b_{it}, c_{it}, d_{it}, e_{it})$ to execute operational command (e.g., energy supply, grid health maintenance, connect/disconnect from the main grid, etc). Each message $\boldsymbol{x}_{it}$ contains send packet a_{it}, send packet size b_{it}, number of packets source to destination c_{it}, number of packet destination to source d_{it}, and total received packets e_{it}. Fake or generated control messages create a major risk for cyber vulnerabilities by executing protocol and replay attacks in PGSC. In particular, the reconstruction capability of GenAI introduces a high risk of protocol and replay attacks in PGSC. Thus, PGSC is potentially under the high risk of cyber vulnerabilities that may create power outages, grid health, information theft, unstable market, and so on. We will hence introduce a novel system model for identifying the cyber vulnerabilities risk of the potential GenAI-driven cyber attacks for assuring a stable PGSC.

A. Power Grid Supply Chain Stability Model

In our model, each DER i can transfer energy to other DERs in set $\mathcal{I}$. All DERs in $\mathcal{I}$ are equipped with oscillators. Now, for transferring energy from DER i to DER $j \in \mathcal{I}, i \neq j$, we define a coupling strength α_{ij}, a rotor angle Θ_i, and a damping constraint β_i for DER i. We can now define the dynamics of power transmission by an oscillator model [16] in PGSC,

$$\frac{d^2\Theta_i}{dt^2} = q_i - \beta_i \frac{d\Theta_i}{dt} + \sum_{j=1, j\neq i}^{I} \alpha_{ij} \sin(\Theta_j - \Theta_i), \quad (1)$$

where q_i represents the power. Therefore, the power transfer between DER i to DER $\forall j \in \mathcal{I}, i \neq j$ is relay on the time derivative. Therefore, for produced/supply power $\hat{q}_i$, the oscillator model (1) can be presented as follows [17]:

$$\hat{q}_i(t) = q_i - \Phi_i \frac{d\Theta_i}{dt}(t), \quad (2)$$

where Φ_i is the elasticity of DER i and Φ_i is proportional to energy market elasticity [17]. Further, the rotation reference of angular frequency deviation $\frac{d\Theta_i}{dt}$ depends on the power grid architecture such as $2\pi \times 50$ Hz or $2\pi \times 60$ Hz. Consequently, the potential supply chain instability is induced by a response delay τ_i of each DER $i \in \mathcal{I}$ (i.e., generator and consumer in PGSC). Then, we can present transmission power $\hat{q}_i(t)$ as $\hat{q}_i(t-\tau)$, where τ is the response delay. We can now derive a new oscillator model using (1) and (2):

$$\frac{d^2\Theta_i}{dt^2} = q_i - \beta_i \frac{d\Theta_i}{dt} + \sum_{j=1}^{I} \alpha_{ij} \sin(\Theta_j - \Theta_i) - \Theta_i \frac{d\Theta_i}{dt}(t-\tau). \tag{3}$$

Clearly, the stability of PGSC $s_i(t) \approx \frac{d^2\Theta_i}{dt^2}$ relies on the physical behavior and control message $\boldsymbol{x}_{it}$ of each DER $i \in \mathcal{I}$. For measuring the PGSC stability in a finite time interval length of T, we can write grid stability as follows:

$$s_i(t) \approx \frac{d^2\Theta_i}{dt^2} = q_i - \beta_i \frac{d\Theta_i}{dt} + \sum_{j=1}^{I} \alpha_{ij} \sin(\Theta_j - \Theta_i) - \frac{\Phi_i}{T} \int_{t-T}^{t} \frac{d\Theta_i}{dt}(t'-\tau)dt'. \tag{4}$$

$s_i(t)$ can be used to assess whether the PGSC is stable or not, For instance, a positive value of $s_i(t)$ means the PGSC is linearly unstable. Therefore, GenAI can manipulate and create fake parameters (e.g., rotor angle Θ_i, damping constraint β_i, elasticity Φ_i, etc) of a control message $\boldsymbol{x}_{it}$. To assess the risk of GenAI-driven attacks, in our model, we use GAN to analyze the capability of a new attack surface in PGSC.

B. GAN for Identifying GenAI-driven Attack Vectors on PGSC

We use GAN [9] to uncover the new attack surface on PGSC. We specifically leverage GAN to reproduce the PGSC control and status messages $\boldsymbol{x}_{it}$ to examine the risk of cyber vulnerabilities and power grid instability. Considering a likelihood-free generator G_θ can generate operational control message $\boldsymbol{x}_{it}$, where θ denotes learning parameters. We introduce a discriminator D_ϕ with parameters ϕ. Therefore, generator G_θ can generate control message $\boldsymbol{x}_{it}$ from sample $\boldsymbol{z}_{it}$ based on some latent variables, where intuitively, $\boldsymbol{z}_{it}$ is a noise vector. We define y_{it} as a decision variable that discriminator D_ϕ uses to predict whether $\boldsymbol{x}_{it}$ is a generated control message or not. Consequently, a control message generator G_θ minimizes the residual between two sample distribution $P_{\mathbb{X}} \approx P_\theta$ while discriminator D_ϕ maximizes the distance distribution of $P_{\mathbb{X}}$ and P_θ, where $\mathbb{X}$ is a given distribution of the DER control message. We can write the GAN model as follows [9]:

$$\min_\theta \max_\phi U(G_\theta, D_\phi) = \min_\theta \max_\phi \mathbb{E}_{\boldsymbol{x}_{it} \sim P_{\mathbb{X}}} \big[\log D_\phi(\boldsymbol{x}_{it})\big] + \mathbb{E}_{\boldsymbol{z}_{it} \sim P_{\boldsymbol{z}_{it}}} \big[\log(1 - D_\phi(G_\theta(\boldsymbol{z}_{it})))\big]. \tag{5}$$

In (5), for a given generator G_θ, the discriminator D_ϕ is maximizing the objective with respect to parameters ϕ. The discriminator D_ϕ then performs the role of a binary classification decision y_{it} (i.e., whether the control message is original or fake) on $\boldsymbol{x}_{it} \sim P_{\mathbb{X}}$. We define $P_{\mathbb{X}}(\boldsymbol{x}_{it})$ and $P_G(\boldsymbol{x}_{it})$ as, respectively, the probability of an actual and generated control message. Hence, the discriminator D_ϕ can be written as follows:

$$\hat{D}_\phi(\boldsymbol{x}_{it}|G_\theta) = \frac{P_{\mathbb{X}}(\boldsymbol{x}_{it})}{P_{\mathbb{X}}(\boldsymbol{x}_{it}) + P_G(\boldsymbol{x}_{it})}. \tag{6}$$

We can observe the probability of generated control message of DER i at time t by estimating (6). Therefore, a generated control message $\boldsymbol{x}_{it}$ has significantly increased the risk of cyber vulnerability and energy market instability in the PGSC. The generated control message can execute a replay and protocol attack in PGSC. In particular, the GAN can reproduce a copy of a DER control message such as send packet a_{it}, send packet size b_{it}, number of packets source to destination c_{it}, number of packet destination to source d_{it}, and total received packets e_{it} while capable of manipulating rotor angle Θ_i, damping constraint β_i, elasticity Φ_i, and so on.

In this work, we develop a *Zero trust framework for risk realization and defense against GenAI-driven cyber attacks* in the PGSC. Therefore, we consider extreme value theory such as conditional-value-at-risk (CVaR) [18]–[20] to realize AI-driven cyber vulnerabilities in PGSC.

III. GenAI-driven Vulnerability Risk Assessment Problem Formulation of PGSC

Next, we formulate a zero trust risk assessment problem to understand GenAI-driven cyber vulnerability on PGSC. We quantify the tail risk of cyber attacks by leveraging the concept of CVaR [18], [20], [21]. In particular, we formulate a residual minimization problem for quantifying tail risk of a AI-generated control message $\boldsymbol{x}_{it}$ at DER $i \in \mathcal{I}$ while satisfying CVaR confidence level ξ. We consider $h(\boldsymbol{x}_{it}, \xi)$ is a probability distribution of trustworthy control message while ξ can be a cut-off point of a risk deviation function $\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z})$, where $\boldsymbol{z}$ represents latent variables of GAN (see detailed in section II-B). Thus, for a CVaR confidence ξ, a cumulative distribution function (CDF) can be calculated as follows [21]:

$$h(\boldsymbol{x}_{it}, \xi) = \int_{\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z}) \le \xi} P(\boldsymbol{z}) d\boldsymbol{z}, \tag{7}$$

where ξ is inversely proportional to $\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z})$. In (7), $h(\boldsymbol{x}_{it}, \xi)$ becomes a nondecreasing and continuous function [20], [21] because ξ satisfies $\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z}) \le \xi$. For a CVaR significant probability $\eta \in (0, 1)$, we can define a random variable $\Psi_\eta(\boldsymbol{x}_{it})$ of control message $\boldsymbol{x}_{it}$. Therefore, we can define a value-at-risk quantification function $\xi_\eta(\boldsymbol{x}_{it})$ of control message $\boldsymbol{x}_{it}$ as follows:

$$\xi_\eta(\boldsymbol{x}_{it}) = \min_{\xi \in \mathbb{R}} h(\boldsymbol{x}_{it}, \xi) \ge \eta. \tag{8}$$

We can estimate ξ in (8) by satisfying $h(\boldsymbol{x}_{it}, \xi) \ge \eta$ and $\xi_\eta(\boldsymbol{x}_{it})$ becomes an upper-bound of tail risk on control message $\boldsymbol{x}_{it}$. Therefore, we can capture a conditional expectation of CVaR $\Psi_\eta(\boldsymbol{x}_{it})$ of AI generated control message $\boldsymbol{x}_{it}$ as follows:

$$\min_{\xi \in \mathbb{R}} \frac{1}{(1-\eta)} \int_{P(\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z})) \ge \xi_\eta(\boldsymbol{x}_{it})} \Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z}) P(\boldsymbol{z}) d\boldsymbol{z}, \tag{9}$$

where $P(\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z})) \geq \xi_\eta(\boldsymbol{x}_{it}) = (1-\eta)$. Therefore, we can define the tail-risk realization objective $\Lambda_\eta(\boldsymbol{x}_{it}, \xi)$ as follows:

$$\min_{\xi \in \mathbb{R}} \xi + \frac{1}{(1-\eta)} \int_{h(\boldsymbol{x}_{it},\xi) \geq \xi} [h(\boldsymbol{x}_{it}, \xi) - \xi]^+ P(\boldsymbol{z}) d\boldsymbol{z}. \quad (10)$$

In CVaR formulation (10), $[h(\boldsymbol{x}_{it}, \xi) - \xi]^+$ is positive and continuous since $h(\boldsymbol{x}_{it}, \xi)$ is a continuous function in (7). An approximate function of CVaR in (10) will be::

$$\hat{\Lambda}_\eta(\boldsymbol{x}_{it}, \xi) = \min_{\xi, \boldsymbol{x}_{it}, y_{it}} \xi + \frac{1}{(1-\eta)} \frac{1}{|\mathcal{I}|T} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{I}|} \Delta_{it}, \quad (11)$$

where $\Delta_{it} \geq (h(\boldsymbol{x}_{it}, \xi) - \xi)$ and $\Delta_{it} \geq 0$. Therefore, we formulate the risk-realization problem of GenAI-driven control message in PGSC as follows:

$$\min_{\xi, \boldsymbol{x}_{it}, y_{it}} \quad \xi + \frac{1}{(1-\eta)} \frac{1}{|\mathcal{I}|T} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{I}|} \Delta_{it}, \quad (12)$$

$$\text{s.t.} \quad \Delta_{it} \geq (h(\boldsymbol{x}_{it}, \xi) - \xi), \Delta_{it} \geq 0, \quad (12a)$$

$$\hat{D}_\phi(\boldsymbol{x}_{it}|G_\theta) \geq \frac{P_{\mathbb{X}}(\boldsymbol{x}_{it})}{P_{\mathbb{X}}(\boldsymbol{x}_{it}) + P_G(\boldsymbol{x}_{it})}, \quad (12b)$$

$$h(\boldsymbol{x}_{it}, \xi) \geq \eta, \eta \in (0,1), \quad (12c)$$

$$s_i(t) \leq 0, s_i(t) \in \boldsymbol{x}_{it}, s_i(t) \in (-1,1), \quad (12d)$$

$$y_{it} \geq \omega_0 + \omega_1 z_{1i} + \cdots + \omega_N z_{Ni}, \forall z_{Ni} \in \boldsymbol{z}_{it}, \quad (12e)$$

$$y_{it} \in \{0,1\}, y_{it} \in \boldsymbol{y}, \forall i \in \mathcal{I}. \quad (12f)$$

The objective of (12) is to minimize the expected shortfall (i.e., mean-variance) with a given significant label of risk η on a generated AI-driven control message in PGSC. Therefore, in (12), we have three decision variables, CVaR cut-off point in long-tail distribution ξ, generated control message $\boldsymbol{x}_{it}$ of DER $i \in \mathcal{I}$, and binary decision variable $y_{it} \in \boldsymbol{y}$ to determine whether the control messages become fake or real. Constraint (12a) provides to an upper-bounded equivalent function of original objective (11). Constraint (12c) assigns a probability for determining an actual and generated control message of DER $i \in \mathcal{I}$ during the GAN fake message generation. Then, constraint (12c) ensures a certain significant level $\eta \in (0,1)$ (e.g., 0.95) of tail risk for a generated AI-driven control message $\boldsymbol{x}_{it}$. Constraint (12d) establishes a connection among the grid stability parameters such as rotor angle Θ_i, damping constraint β_i, elasticity Φ_i of oscillator model (4) to transfer energy. Constraint (12d) assures a stable PGSC by restricting $s_i(t)$ to negative values. Constraint (12e) establishes a relationship between the GAN's latent variables $\boldsymbol{z}$ and a regression weight ω for distinguishing $y_{it} \in \boldsymbol{y}$ among generated and original control message. Finally, constraint (12f) assures y_{it} as a binary variable for each control message $\boldsymbol{x}_{it}$.

The formulated zero trust problem (12) is to a combinatorial optimization problem due to the relationship among the corresponding constraints. Further, decision variables of the formulated problem (12) belong to both time and space domains while they are correlated. As a result, the formulated zero trust problem (12) is hard to solve in polynomial time complexity. Therefore, we propose a zero trust framework for extreme risk realization and defense against generated-AI driven attacks on PGSC. In particular, the proposed zero trust framework consists of 1) a domain-specific GAN model that can generate fake control/status messages, and 2) a probabilistic linear model with regression mechanism to realize risk and defense against attack surface on PGSC.

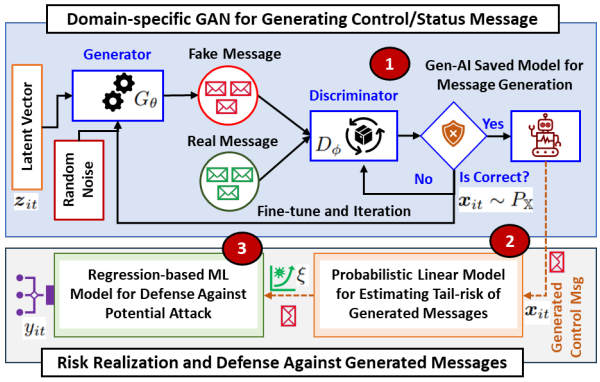


Fig. 2: Proposed zero trust framework for risk realization and defense against GenAI-driven attacks on the PGSC.

Algorithm 1 GAN-based Training Algorithm for Control/Status Messages Generation in PGSC

Input: $\mathcal{I}$, $\mathbb{X}$
Output: $\forall \boldsymbol{x}_{it} \in \mathcal{I}$
Initialization: G_θ, D_ϕ, θ, ϕ, $\boldsymbol{z}$
for Until max epoch: $n \geq N$ **do**
  Mini batch: $\mathbb{X}$, $\boldsymbol{z}$
  Gradient decent θ**:** $\bigtriangledown_\theta U(G_\theta, D_\phi)$ in (13)
  Gradient ascent ϕ**:** $\bigtriangledown_\phi U(G_\theta, D_\phi)$ in (14)
  Execute: $\hat{D}_\phi(\boldsymbol{x}_{it}|G_\theta) \geq \frac{P_{\mathbb{X}}(\boldsymbol{x}_{it})}{P_{\mathbb{X}}(\boldsymbol{x}_{it}) + P_G(\boldsymbol{x}_{it})}$, in (12b)
end for
Trained model saved as $h5$ **file**
return θ, ϕ, $\boldsymbol{x}_{it}$

IV. Zero Trust Framework Design

We solve the formulated zero trust risk realization problem (12) by designing an analytical framework (as seen in Figure 2) that can generate fake control/status messages, capable of quantifying extreme risk on generated messages, and protects the PGSC by autonomously detecting fake messages. In particular, we develop a domain-specific GAN mechanism to create the new attack vector by generating control/status messages $\boldsymbol{x}_{it}$ of DERs $\forall i \in \mathcal{I}$. We determine conditional-value-at-risk confidence level ξ of the GenAI-driven attack vector by solving a probabilistic model while a regression-based machine learning (ML) model is devised to detect the fake y_{it} control message to protect PGSC.

A. A GAN for Producing New Attack Vector on PGSC

Algorithm 1 illustrates the proposed GAN-based training mechanism for producing new attack vectors by generating DERs control/status messages on PGSC. We initialize a generator G_θ, discriminator D_ϕ, noise vector $\boldsymbol{z}$, learning parameters

Algorithm 2 Probabilistic and Regression-based Algorithm for Realizing Risk and Defense Against GenAI Attacks

Input: $\mathcal{I}$, η, $\boldsymbol{x}_{it}$, θ, ϕ, trained model (h5)
Output: ξ, y_{it}
Initialization: Θ_i, β_i, Φ_i, η, θ, ϕ
1: **for** $t \geq T$ **do**
2: **for** $P(\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z})) \geq \xi_\eta(\boldsymbol{x}_{it})$ **do**
3: **Estimate:** σ, μ: $g(\boldsymbol{x}_{it}) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{(\boldsymbol{x}_{it}-\mu)^2}{2\sigma^2}$
4: **Estimate:** $\xi_\eta(\boldsymbol{x}_{it}) = \Gamma(1-\eta) * \sigma - \mu$ for (8)
5: **Estimate:** $\Psi_\eta(\boldsymbol{x}_{it}) = \frac{1}{(1-\eta)} * \Omega(\xi_\eta(\boldsymbol{x}_{it})) * (\sigma - \mu)$ for (9)
6: **Check:** Constraints (12a), (12c), (12d) and **Estimate:** $s_i(t)$ using (4)
7: **Estimate:** $\Lambda_\eta(\boldsymbol{x}, \xi)$ for (11)
8: **for** $i \geq |\mathcal{I}|$ && l **do**
9: **Estimate:** $y_{it} = \omega_0 + \omega_1 z_{1i} + \cdots + \omega_N z_{Ni}, \forall z_{Ni} \in \boldsymbol{z}_{it}$ for (12e) using bagging [22]
10: **Check:** Constraint (12f)
11: **end for**
12: **end for**
13: **end for**
14: **return** ξ, y_{it}

θ and ϕ at the beginning of Algorithm 1. Algorithm 1 is designed for offline training, and thus, line 1 determines the maximum number of epochs N and line 2 represents a high-level step to usage of mini batch during training. In line 3 of Algorithm 1, we execute a gradient decent $\bigtriangledown_\theta U(G_\theta, D_\phi)$ mechanism to determine the learning parameters θ for the generator G_θ and evaluating a control message generation loss. The gradient decent of generator G_θ is given by:

$$\bigtriangledown_\theta U(G_\theta, D_\phi) = \frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \bigtriangledown_\theta \log(1 - D_\phi(G_\theta(\boldsymbol{z}_{it}))). \quad (13)$$

Then, Algorithm 1 executes gradient ascent $\bigtriangledown_\phi U(G_\theta, D_\phi)$ to determine the learning parameters ϕ of discriminator D_ϕ in line 4. The gradient ascent D_ϕ of the discriminator is given by:

$$\frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \bigtriangledown_\phi \big[\log D_\phi(\boldsymbol{x}_{it}) + \log(1 - D_\phi(G_\theta(\boldsymbol{z}_{it}))) \big]. \quad (14)$$

In line 5 of Algorithm 1, the discriminator D_ϕ assigns the probability of generated message and evaluating for fine tuning. Finally, a trained GAN model is saved to realizing generated AI-driven attacks on PGSC.

B. Probabilistic and Regression-based Extreme Risk Realization and Defense Mechanism

We develop a probabilistic and regression-based mechanism for realizing extreme risk and defense against GenAI-driven control message attacks in PGSC. Algorithm 2 presents the overall solution procedure to analyze the CVaR and defense mechanism for new attack vectors on the PGSC generated by Algorithm 1. Therefore, Algorithm 2 receives a trained model as an input from Algorithm 1 and generated control message $\boldsymbol{x}_{it}$. Line 2 of Algorithm 2 ensures the iterative process continues until $P(\Upsilon(\boldsymbol{x}_{it}, \boldsymbol{z})) \geq \xi_\eta(\boldsymbol{x}_{it})$ while line 3 estimates mean μ and standard deviation σ for measuring the reconstruction capabilities of generated control message $\boldsymbol{x}_{it}$. We derive a probability point function (PPF) $\Gamma(1-\eta)$ and estimate $\xi_\eta(\boldsymbol{x}_{it})$ the distribution of generated control message risk (8) as follows (line 4 in Algorithm 2):

$$\xi_\eta(\boldsymbol{x}_{it}) = \Gamma(1-\eta)(\sigma - \mu), \quad (15)$$

where $\Gamma(1-\eta)$ is a probability point function and $\eta \in (0, 1)$. Then, we construct a probability density function (PDF) Ω in line 5 of Algorithm 2 and capture the conditional expectation of CVaR for the AI generated controlled message $\boldsymbol{x}_{it}$. Thus, line 5 of Algorithm 2 execute the following function,

$$\Psi_\eta(\boldsymbol{x}_{it}) = \frac{1}{(1-\eta)} * \Omega(\xi_\eta(\boldsymbol{x}_{it}))\sigma - \mu, \quad (16)$$

where $\Omega(\xi_\eta(\boldsymbol{x}_{it}))$ is a PDF of generated controlled message $\boldsymbol{x}_{it}$. In Algorithm 2, line 6 executes constraints (12a), (12c), and (12d) and estimates PGSC stability index $s_i(t)$ using (4). Line 7 calculates the extreme risk (i.e., CVaR confidence level) cut-off point ξ of the AI generated attack vector $\boldsymbol{x}_{it}$. Finally, lines 8 to 11 are responsible to distinguish between real $y_{it} = 0$ and generated $y_{it} = 1$ control messages x_{it} to protect the PGSC from generated AI-driven attacks. The above solution provides a sub-optimal solution and performance relies on the parameter $\eta \in (0, 1)$.

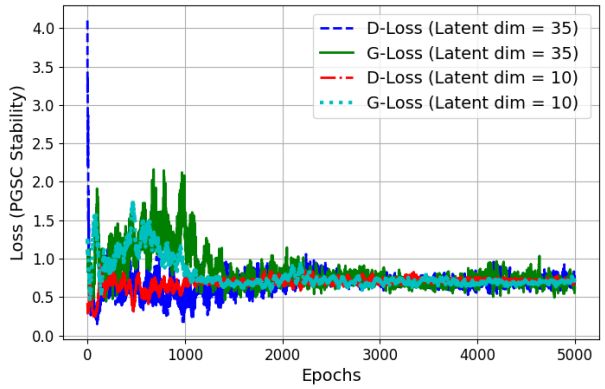
The complexity of the proposed zero trust risk realization and defense framework on PGSC completely depends on the complexity of Algorithm 2 since Algorithm 2 will be deployed in SCADA and being up and running. On the other hand, Algorithm 1 is used for offline training to train an AI model for generating fake DER control/status messages while a trained model is being used by Algorithm 2. Therefore, the complexity of Algorithm 1 can be ignored for the proposed zero trust framework on PGSC. Then, the complexity of Algorithm 2 includes the complexity of two base problems: 1) a probabilistic linear model for extreme risk realization, and 2) a bagged-based random forest scheme for defense mechanism. Hence, the complexity of the probabilistic linear model-based risk realization becomes $\mathcal{O}(|\mathcal{I}|^2)$ [20], where $|\mathcal{I}|$ is the number of generated control messages of DERs $\forall i \in \mathcal{I}$. Now, we define l as the number of bagged trees, where each message $\boldsymbol{x}_{it}$ consists of $|\boldsymbol{x}_{it}|$ features with the weight points ω during the regression learning for detecting AI generated control message. For a given number of bagged trees l, the overall complexity (i.e., time and space) of the defense mechanism belongs to $\mathcal{O}(l|\boldsymbol{x}_{it}|^2|\boldsymbol{\omega}|^2 \log(|\boldsymbol{\omega}|))$, where $\mathcal{O}(l|\boldsymbol{x}_{it}||\boldsymbol{\omega}|^2 \log(|\boldsymbol{\omega}|))$ is the time complexity. As a result, the total complexity of the proposed zero trust framework for PGSC leads to $\mathcal{O}(|\mathcal{I}|^2 + l|\boldsymbol{x}_{it}|^2|\boldsymbol{\omega}|^2 \log(|\boldsymbol{\omega}|))$.

V. Experimental Results and Analysis

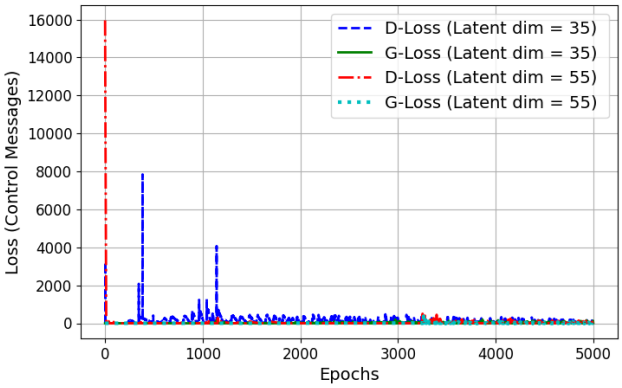
The developed zero trust framework is one of the first work that attempts to realize and defense against GenAI-driven

TABLE II: Summary of Experimental Setup.

Description	Values
Generator	Sequential, 64 units, ReLu (dense), Binary Cross-Entropy, Adam, LR: 0.02, Latent Space: 35, Epoch: 5000
Discriminator	Sequential, 64 units, LeakyReLU (0.2) (dense), Sigmoid, Binary Cross-Entropy, Adam, LR: 0.02, Latent Space: 35, Epoch: 5000
RF bagging	estimators: [50, 100, 200], max features: [auto, sqrt, log2], max depth: [2,4,5,6,7,8], criterion: [gini, entropy]
CVaR	$\eta = \{0.9, 0.95, 0.99\}$



(a) PGSC stability parameters.



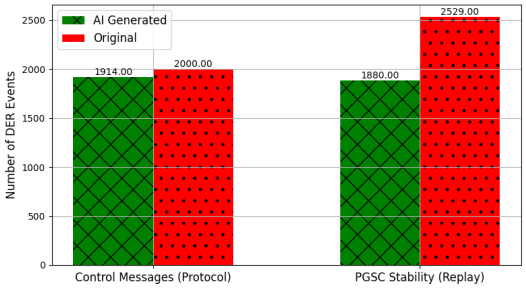
(b) DERs control messages.

Fig. 3: Generation and discrimination loss comparison of the proposed GAN-based model in PGSC.

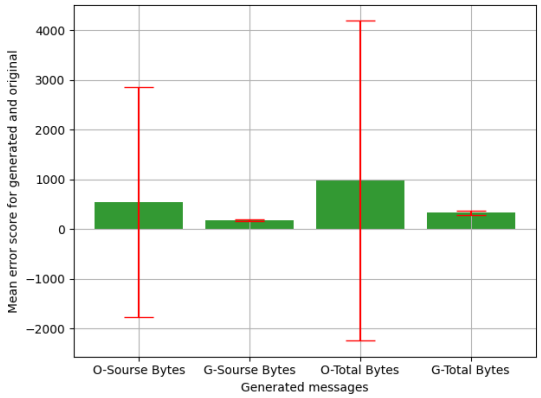
TABLE III: Performance analysis (0-1) for AI generated message detection among several regression-based models.

Methods	Precision	Recall	f1-score	Accuracy
RF (Bagging)	1.0	1	1	1.0
KNN	0.99	1	1	0.99
SVM	1.0	1	1	1.0
Logistic Regression	1.0	1	1	1.0

attacks on PGSC. Therefore, to the best of our knowledge, there are no prior works that can serve as a baseline. Therefore, we compare the proposed zero trust framework using two state-of-the-art datasets, 1) power grid stab stability [13], and 2) SCADA control message [14] to justify the efficacy. We summarize the important parameters of our experimental setup in Table II.



(a) Protocol and replay attacks on PGSC by GenAI.



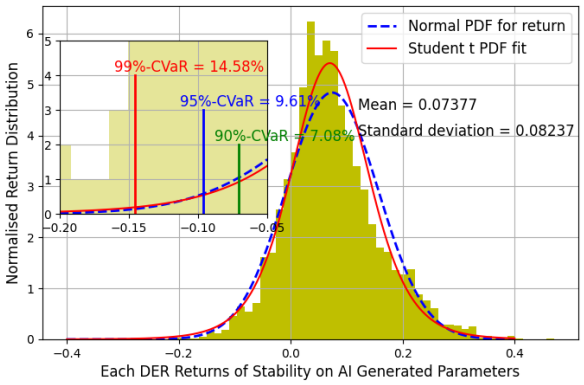
(b) Error analysis of AI generated control message on PGSC.

Fig. 4: Capability of GenAI to create the attack vector on PGSC.

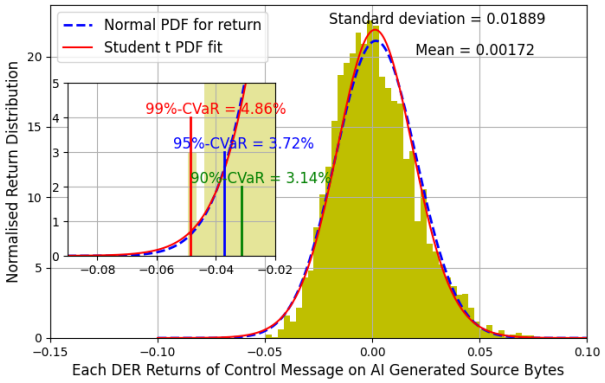
In Figure 3, we assess the convergence, generator loss, and discriminator loss of the proposed GAN-based training Algorithm 1 for two datasets under different latent variables. We choose latent variable length as 35 for both datasets (as seen in Figure 3a for PGSC stability parameters and Figure 3b for control message generation) due to smooth convergence. Then, we analyze the capability of creating new attack vector for both PGSC stability parameters and DER control message in Figure 4, where we achieve around 95.7% accuracy for protocol attack generation (in Figure 4b) and about 74.3% accuracy on replay attack generation (in Figure 4a).

In Figure 5, we assess the extreme risk of the GenAI-driven protocol and replay attacks on the developed zero-trust framework. Figure 5a illustrates that the proposed framework can quantify the extreme risk 7.08%, 9.61%, and 14.58% of GenAI-driven replay attacks for 90%, 95%, and 99% confidence, respectively. Further, Figure 5b demonstrates the extreme risk of GenAI-driven protocol attacks in PGSC, where the proposed framework can find 3.14%, 3.72% and 4.86% risk for 90%, 95%, and 99% confidence, respectively.

in Table III, we analyze the performance of the proposed bagging-based defense mechanism on zero trust framework over several regression-based methods. The results of Table III clearly show that the proposed zero trust framework can effectively detect the GenAI-driven replay and protocol attacks on PGSC.



(a) Replay attacks of PGSC.



(b) Protocol attacks of PGSC.

Fig. 5: Risk realization of AI-generated protocol and replay attacks on PGSC.

VI. Conclusion

In this paper, we have introduced a novel zero-trust framework for the power grid to extreme risk realization and defense against generative AI-driven attacks such as protocol type and replay attacks on PGSC. In particular, we have designed the first approach to investigating GenAI-driven cyber attacks (i.e., protocol and replay) in PGSC, and created a novel zero-trust framework to realize and defend against GenAI attacks for PGSC. The proposed zero trust brings a state-of-the-art cybersecurity framework in the domain of critical power grid supply chains to protect the systems from AI-driven cyber attacks by continuously validating the trust of monitored DERs and their control messages. Experimental results demonstrate the efficiency of the proposed zero trust framework, achieving an accuracy of 95.7% in attack vector generation, a risk realization of 9.61% for a 95% stable PGSC, and a 99% confidence level in defense against Generative AI-driven attacks. In the future, we will further investigate the authentication of each DER to verify the data against being forged.

References

[1] A. A. Habib, M. K. Hasan, A. Alkhayyat, S. Islam, R. Sharma, and L. M. Alkwai, "False data injection attack in smart grid cyber physical system: Issues, challenges, and future direction," *Comput. Electr. Eng.*, vol. 107, no. C, pp. 1–16, April 2023.

[2] M. N. Nafees, N. Saxena, A. Cardenas, S. Grijalva, and P. Burnap, "Smart grid cyber-physical situational awareness of complex operational technology attacks: A review," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–36, February 2023.

[3] A. Vishnoi and V. Verma, "The analysis on impact of cyber security threats on smart grids," in *Security and Risk Analysis for Intelligent Edge Computing*. Springer International Publishing, June 2023, pp. 111–118.

[4] "Cyber resilience," https://www.iea.org/reports/power-systems-in-transition/cyber-resilience., accessed: September, 2023.

[5] M. S. Munir, S. Shetty, and D. B. Rawat, "Trustworthy artificial intelligence framework for proactive detection and risk explanation of cyber attacks in smart grid," in *2023 Winter Simulation Conference (WSC)*, 2023, pp. 636–647.

[6] S. P. Dash and K. V. Khandeparkar, "A false data injection attack on data-driven strategies in smart grid using gan," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2023, pp. 313–324.

[7] M. Ben Driss, E. Sabir, H. Elbiaze, and W. Saad, "Federated learning for 6g: Paradigms, taxonomy, recent advances and insights," *arXiv e-prints*, pp. arXiv–2312, 2023.

[8] M. Ravinder and V. Kulkarni, "A review on cyber security and anomaly detection perspectives of smart grid," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2023, pp. 692–697.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, November 2020.

[10] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," *arXiv preprint arXiv:2211.14343*, 2022.

[11] Q. Zhang, A. Ferdowsi, W. Saad, and M. Bennis, "Distributed conditional generative adversarial networks (gans) for data-driven millimeter wave communications in uav networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1438–1452, 2022.

[12] A. Ferdowsi and W. Saad, "Brainstorming generative adversarial network (bgan): Toward multiagent generative models with distributed data sets," *IEEE Internet of Things Journal*, vol. 11, no. 5, pp. 7828–7840, 2024.

[13] U. M. L. Repository, "Electrical grid stability simulated data set," https://archive.ics.uci.edu/ml/datasets, accessed: September, 2023.

[14] M. Z. M. A. Teixeira and R. Jain, "Wustl-IIOT-2018 dataset for ICS (SCADA) cybersecurity research," https://ieee-dataport.org/open-access/wustl-iiot-2018, accessed: September, 2023.

[15] D. G. Photovoltaics and E. Storage, "Ieee standard for interconnection and interoperability of distributed energy resources with associated electric power systems interfaces," *IEEE Std*, vol. 1547, pp. 1547–2018, 2018.

[16] B. Schäfer, C. Grabow, S. Auer, J. Kurths, D. Witthaut, and M. Timme, "Taming instabilities in power grid networks by decentralized control," *The European Physical Journal Special Topics*, vol. 225, pp. 569–582, May 2016.

[17] B. Schäfer, M. Matthiae, M. Timme, and D. Witthaut, "Decentral smart grid control," *New journal of physics*, vol. 17, no. 1, pp. 015 002–015 016, January 2015.

[18] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can terahertz provide high-rate reliable low-latency communications for wireless vr?" *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9712–9729, 2022.

[19] M. S. Munir, S. F. Abedin, N. H. Tran, Z. Han, E.-N. Huh, and C. S. Hong, "Risk-aware energy scheduling for edge computing with microgrid: A multi-agent deep reinforcement learning approach," *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 3476–3497, 2021.

[20] M. S. Munir, D. H. Kim, A. K. Bairagi, and C. S. Hong, "When cvar meets with bluetooth pan: A physical distancing system for covid-19 proactive safety," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 858–13 869, June 2021.

[21] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, May 2002.

[22] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, August 1996.