A Bayesian Belief Network Framework for Protecting Generative AI-driven Attack in Smart Grid Communication

Md Shirajum Munir[1], Sravanthi Proddatoori[2], Manjushree Muralidhara[2], and Sachin Shetty[3]
[1]School of Computing, Analytics, and Modeling, Univesity of West Georgia, Carrollton, GA 30118, USA.
[2]Dept. of Computer Science, Old Dominion University, Norfolk, VA 23529, USA.
[3]Dept. of ECE, Old Dominion University, Norfolk, VA 23529, USA.
Email: mmunir@westga.edu; sprod002@odu.edu; mmura001@odu.edu; sshetty@odu.edu

***Abstract*—In the era of artificial intelligence (AI), native generative AI (GenAI) has become prominent in generating intelligent cyber-attacks in a smart grid environment by imitating operational parameters and metrics. As a result, it is imperative to understand the attack generation capabilities of native GenAI models, such as Generative Adversarial Networks (GANs) and Autoencoders in a smart grid environment. This research aims to explore the potential of GenAI models to accurately imitate and analyze the behavior of cyberattacks within a smart grid environment. The goal is to understand the risks posed by these intelligent attacks while developing strategies to protect against disruptive Distributed Energy Resources (DER) events, such as incorrect load shifting, imbalanced demand supply, and unstable price forecasting. First, this work leverages two customized native GenAI models, namely GAN and Autoencoder, to synthesize DERs control message parameters such as nominal power consumption, price elasticity coefficients, and communication data packets, among others, to introduce highly unpredictable new intelligent attack vectors. Second, this research proposes and develops a new Bayesian Belief Networks (BBNs) framework by creating correlation dependency nodes among DER control message parameters to understand the risks of intelligent attack vectors better while monitoring and mitigating their impact on smart grid operations. In particular, by leveraging the customized BBN, the system can observe and analyze the uncertain behaviors in DER operations that GenAI introduces, and the proposed framework can help mitigate grid vulnerabilities by understanding the poisonous parameters through mutual information. Finally, the experiment results show that the Autoencoder outperformed GAN in reproducing intelligent attacks with an MSE about 98.984% lower than that of GANs. Additionally, BBN can explain prominent parameters of the intelligent attacks' by quantifying dependencies through mutual information.**

***Index Terms*—Intelligent Cyber Attack, Generative AI, GAN, Autoencoder, Bayesian Belief Networks, Cyber Risk.**

This work is supported in part by a professional development grant from Dr. James 'Earl' Perry College of Mathematics, Computing, and Sciences, University of West Georgia, and the Commonwealth Cyber Initiative under grant HC-3Q24-049, an investment in the advancement of cyber R&D, innovation, and workforce development.

I. Introduction

In today's digitally connected world, protecting smart grid systems from intelligent cyber attacks is crucial [1]–[4]. It undergoes the cybersecurity challenges faced by smart grid systems, highlighting the need for advanced detection methods [5]. The growing use of generative artificial intelligence (GenAI) in energy systems, including the creation of synthetic data and forecasting, makes this assessment crucial [1]–[4], [6]–[9]. Cybersecurity researchers can benefit greatly by leveraging GenAI models because of their exceptional capacity to produce synthetic data and mimic intricate behaviors. Several studies [7], [9] utilized Variational Autoencoder (VAE) and GAN models for synthetic data generation in smart homes. However, there is a lack of investigation on GenAI models in terms of intelligent attack generation capabilities and analyzing such new attack vectors in a smart grid environment.

In this work, our main focus is on investigating the capacity of GenAI models such as GAN and Autoencoder to reproduce intelligent attacks in a smart grid environment. Then investigate how serious these attacks are and develop a new Bayesian Belief Network (BBN) framework to understand the effect of such attacks for enabling cybersecurity strategies for smart grid systems. We summarized our key contributions as follows:

- First, we have designed two native generative AI models, such as GAN and Autoencoder to reproduce intelligent attacks in a smart grid environment while investigating the accurate generation capability among these models.
- Second, we have developed a Bayesian Belief Network (BBN)-based risk explanation and protection framework to enable the understanding capabilities of new intelligently generated cyber attacks such as replay attacks by native GenAI. In particular, we utilize Pearson correlation coefficient [10] among the smart grid operational parameters to build the BBN while each parameter is considered as a node.
- Third, we calculate the mutual information among the BBN node (i.e., smart grid operational parameters) of the BBN to determine the prominent attack features of a particular native GenAI-driven attack for protecting the smart grid operation from attacks.
- Finally, we have considered an open-source intelligent control system (ICS) dataset WUSTL-IIOT-2018 [11] to analyze the capability of GenAI-driven attacks and evaluate the proposed framework. Experiment results show Autoencoder has more potential to generate intelligent attacks in smart grid operation than the GAN while the proposed framework can pinpoint the parameters/features of attacks by quantifying mutual information of BBN node.

The paper is organized as follows. A description of GenAI-

driven intelligent attack generation is presented in Section II. The proposed BBN-based protection framework in smart grid is described in Section III. We provide a detailed discussion on experimental analysis in Section IV. Finally, we conclude our discussion in Section V.

II. GenAI-Driven Intelligent Attack Generation

We consider two native GenAI models such as GAN and Autoencoder, as the intelligent attack generation agents in a smart grid environment, as seen in Figure 1.

A. GAN-based agent for penetration capability analysis

We leverage the GAN architecture that was first introduced by Ian Goodfellow [12]. The GAN architecture consists of a generator G and a discriminator D. The Generator creates synthetic data from random noise, and the Discriminator distinguishes between real and generated samples. The training involves iterative refinement of both components to achieve realistic data generation. [13]. In particular, a GAN consists of a generator and discriminator. The generator takes random noise z and transforms it into a fake sample $G(z)$. The discriminator evaluates whether a sample is real or fake, outputting the $D(G(z))$ for generated samples and $D(x)$ for real samples. The GAN trains to minimize the cross entropy loss between the real and generated distribution. The mathematical expression of vanilla GAN based on the cross-entropy between real and generated distributions is defined as follows:

$$\min_{\theta_G}\max_{\theta_D} V(\theta_D,\theta_G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D_{\theta_D}(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\text{latent}}(\boldsymbol{z})}[\log(1-D_{\theta_D}(G_{\theta_G}(\boldsymbol{z})))]. \quad (1)$$

B. Autoencoder-based agent for penetration capability analysis

The Autoencoder architecture includes an Encoder, a Decoder, and a latent feature representation. We use an Autoencoder to reproduce the original input. Simultaneously, it should create a meaningful latent representation. It captures all the features from the original data samples [14]. The Encoder compresses data into a latent space while the Decoder reconstructs the data from this compressed form. This process helps in learning critical features from the data.

The Autoencoder-based model for penetration capability analysis involves a system architecture with an encoder g_ϕ, a decoder f_Θ, and a bottleneck layer representing input data in a low-dimensional form. The encoder compresses the input data into a feature vector, which is then passed through the bottleneck, while the decoder reconstructs the original data dimensions from the compressed representation. The effectiveness of the model is determined by minimizing the difference between the input data $\boldsymbol{x}$ and the reconstructed output $\boldsymbol{x'}$, with the Mean Absolute Error (MAE) serving as the cost function to evaluate this reconstruction accuracy.

This model is implemented using neural networks, where the weights and biases are defined by parameters Θ and ϕ. By feeding the input feature vector into the Autoencoder, the model compresses and then decompresses the data, aiming to reconstruct the original input as closely as possible. Minimizing the reconstruction error is crucial, as it indicates the model's ability to regenerate data accurately, which is essential for effective penetration capability analysis.

TABLE I: Cost Function Notations.

Notation	Description
$\mathcal{L}(\theta,\phi$	Cost function of attack generation
n	Number of samples
x^i	Input sample i^{th}
$f(\theta)$	Encoder function
$g(\phi)$	Decoder function
x'	Generated attack

The cost function of the considered Autoencoder is as follows [15]:

$$L(\theta,\phi) = \frac{1}{n}\sum_{i=1}^{n}|x^i - f_\theta(g_\phi(x^i))|. \quad (2)$$

In the cost function 2, $f_\theta(g_\phi(x^i)$ represents the generated control messages because the x^i goes through the $g(\phi)$ first, which is the encoder and then it goes through the function $f(\theta)$ which is a decoder. This mathematical representation represents the data flow from input to output. The equation 2 represents that we are summing up the features from $i=1$ to n and considering the difference between the original input data and the generated data.

III. Proposed BBN-based Protection Framework in Smart Grid

A. Bayesian Belief Network (BBN)-based Framework for Penetration Capability Analysis

In this section, the proposed framework leverages a BBN to analyze penetration capabilities within a smart grid system. The proposed framework incorporates covariance and mutual information calculations to model the probabilistic relationships between system parameters. This approach allows the system to detect and explain attack patterns introduced by native GenAI models.

To illustrate the proposed framework, firstly, the system uses the feature set of smart grid operational parameters, denoted by $\boldsymbol{x'}$, as the input to the BBN, where each node in the BBN represents a control parameter of the Distributed Energy Resources (DER), such as nominal power consumption or communication data packets.

Secondly, we have considered the covariance correlation coefficients between the parameters to quantify the linear dependencies between two continuous variables. This coefficient is a covariance measure that tells you how strongly two variables are related. The Pearson correlation coefficient between two random variables such as X and Y is defined as follows [10], [16]:

$$g(X,Y) = \frac{E[X^2]-(E[X])^2}{\sqrt{E[Y^2]-(E[Y])^2}\cdot\sqrt{E[XY]-E[X]E[Y]}}. \quad (3)$$

In 3, the $\mathbb{E}[XY]$ is the expected value of the product of X and Y, and $\mathbb{E}[X]$, $\mathbb{E}[Y]$ are the expected values of X and Y, respectively. The correlation coefficient $g(X,Y)$ defines how closely the parameters are related, which is critical for constructing the edges in the BBN. Nodes in the BBN with a higher covariance are more likely to influence each other. This helps the system understand the strength of relationships among DER parameters, providing a foundation for analyzing

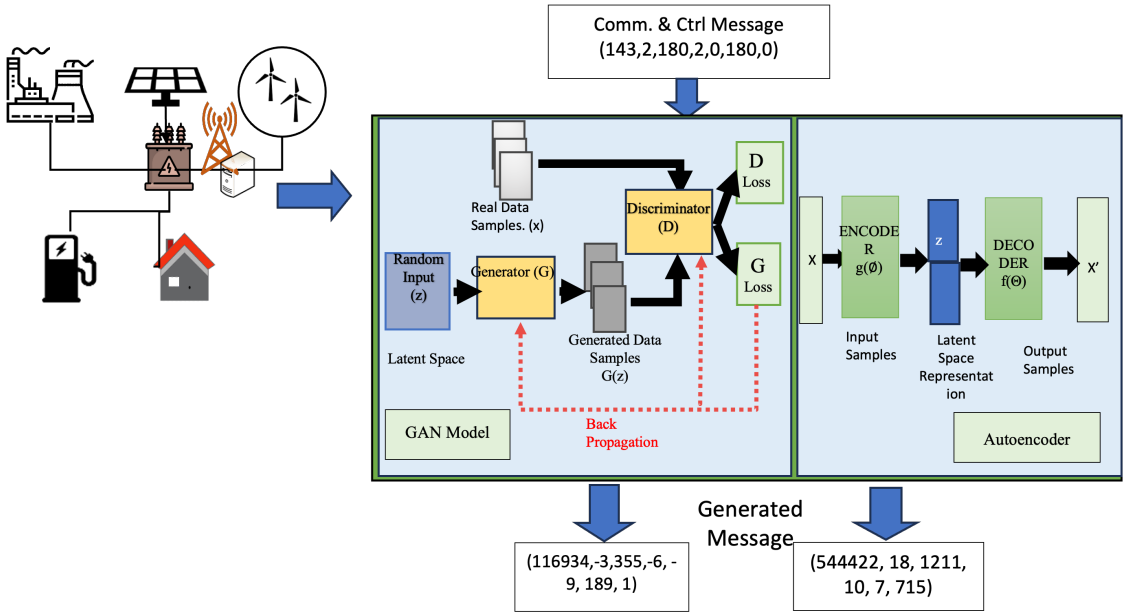


Fig. 1: Native GenAI models for analyzing the penetration capability in communication and control messages in smart grid.

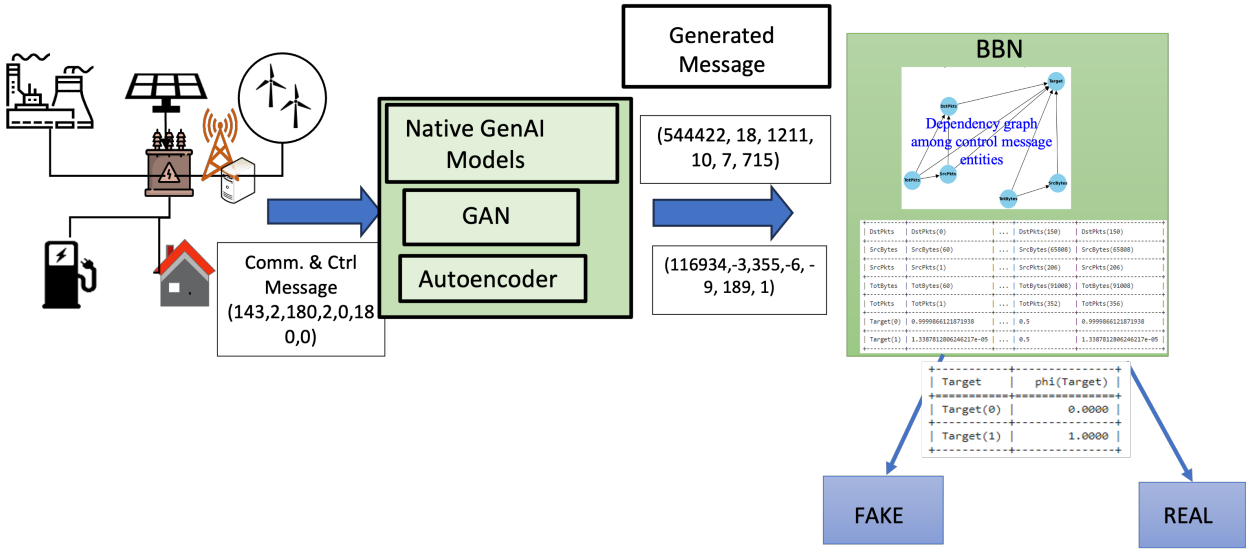


Fig. 2: Proposed native BBN-based protection framework.

dependencies in the network. Once the dependencies between BBN nodes are established using Pearson's correlation, the framework computes the mutual information between each pair of nodes to capture more complex, non-linear dependencies. Mutual information quantifies how much information a DER parameter provides about another DER parameter beyond simple linear relationships. Mutual information is computed as follows [17]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in B} P(x,y) \log \left(\frac{P(x)P(y)}{P(x,y)} \right), \quad (4)$$

where $P(x,y)$ represents the joint probability of parameters A and B while $P(x)$ and $P(y)$ are their individual probabilities. The mutual information $I(X;Y)$ captures the total dependency between X and Y in the BBN.

By analyzing the mutual information, the framework identifies the most critical parameters and their influence on intelligent and reply attacks in the smart grid. The mutual information-based analysis helps to explain and rank the dependencies among parameters, allowing for a deeper understanding of its impact.

B. Algorithmic Procedure of the Proposed Protection Framework

We present our proposed native GenAI-driven intelligent replay attack generation and BBN-based protection framework algorithmic procedure in Algorithm 1. In Algorithm 1, lines from 1 to 13 present the two native GenAI agent training procedure for producing intelligent attacks in a smart grid environment. In particular, lines from 1 to 6 describe the GAN training pseudo description while lines from 7 to 13 represent the Autoencoder training procedure in Algorithm 1. Lines from 14 to 22 of Algorithm 1 illustrate the steps of BBN-based protection framework for smart grid. The Pearson correlation is first used to initialize BBN nodes, and mutual information is computed to establish dependencies between

Algorithm 1 Native GenAI-driven Intelligent Attack Generation and BBN-based Protection Framework

Input: Dataset $\mathcal{D}$, including real samples $x \in \mathcal{D}$, noise vector z, and BBN parameters Θ

Output: Intelligent attack sample x', Anomalies a

Initialization: Initialize GAN parameters θ_G and θ_D, Autoencoder parameters ϕ and Θ, and BBN structure

1: **for** epochs **do**
2: **Train GAN:**
3: Generate fake samples: $x' \leftarrow G(z; \theta_G)$
4: Discriminate: $D(G(z; \theta_G))$ and $D(x)$
5: Update G and D using loss function (1)
6: **end for**
7: **for** epochs **do**
8: **Train Autoencoder:**
9: Encode: $z \leftarrow g_\phi(x)$
10: Decode: $x' \leftarrow f_\Theta(z)$
11: Compute loss: $L(\theta, \phi)$ using (2)
12: Update g_ϕ and f_Θ to minimize $L(\theta, \phi)$
13: **end for**
14: **Construct BBN:**
15: Compute Pearson correlation coefficient to initialize BBN nodes using 3
16: Build directed acyclic graph with nodes and edges based on mutual information
17: Estimate conditional probabilities and BBN parameters Θ
18: Compute mutual information between features using 4
19: **Intelligent Attack Detection and Explanation:**
20: Perform probabilistic inference on generated samples x' using BBN
21: Identify anomalies a based on the inferred probabilities
22: **return** Intelligent attack sample x', Anomalies a

TABLE II: Experimental Setup.

Component	Configuration
GAN Generator	Input layer with latent vectors Hidden layers: 64, 128, 256 neurons with ReLU activation Output layer: 7 neurons with linear activation
GAN Discriminator	Input layer: 7 neurons Hidden layers: 64 neurons (Leaky ReLU, alpha = 0.2), 128 neurons Output layer: 1 neuron with sigmoid activation
Training Objective	Minimize binary cross-entropy loss for classification accuracy between real and generated data
Autoencoder Encoder	Layers: 128, 64, 32, 8 neurons (ReLU activation)
Autoencoder Decoder	Layers: 8, 32, 64, 128 neurons Output layer: 6 neurons
Autoencoder Loss Function	Minimize reconstruction error using Mean Absolute Error (MAE)

smart grid parameters. The BBN then estimates conditional probabilities and performs probabilistic inference on the generated attack samples, enabling the system to identify anomalies. This approach allows for the detection and explanation of intelligent attack vectors by analyzing abnormal behavior in the data. Then the mutual information between nodes (i.e., dependent parameters) is calculated via scikit-learn's mutual info regression function in line 18 to capture dependencies that influence the severity of an intelligent attack. The complexity of inference depends on the number of features and leads to $O(|\boldsymbol{x}|)^n$.

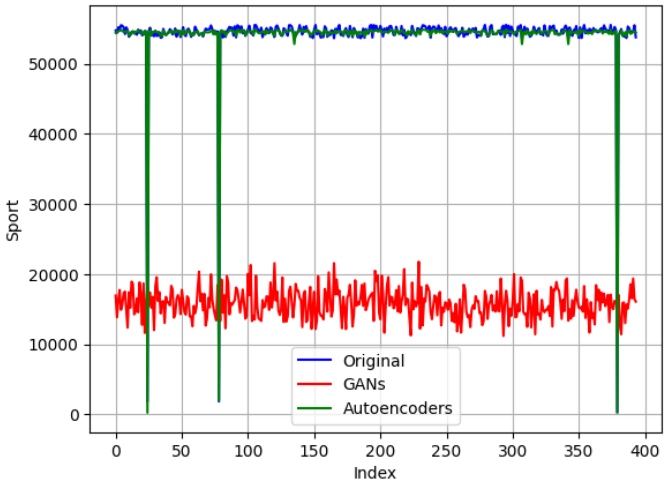


Fig. 3: The penetration performance comparison between the device port number and generated port number by the proposed framework.

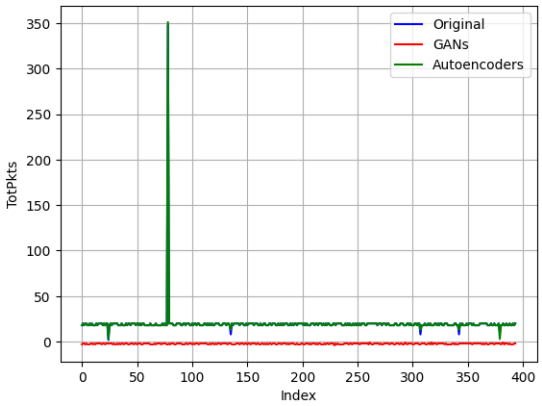


Fig. 4: The penetration performance comparison between the original message and generated message by the proposed framework.

IV. Experimental Results & Discussion

In this section, we discuss the key findings of our experimental results. We analyze the performance of native GenAI models and BBN models in generating intelligent replay attacks and their ability to replicate original control messages. The native GenAI models were trained on the WUSTL-IIOT-2018 dataset [11] to generate synthetic data mimicking attack vectors in the smart grid context. Table II contains comprehensive details on the architecture, design, and fine-tuning of native GenAI models and BBN Network. In particular, Table II includes all relevant parameters and configurations used in models, providing an in-depth reference for understanding their architectural patterns and optimization processes.

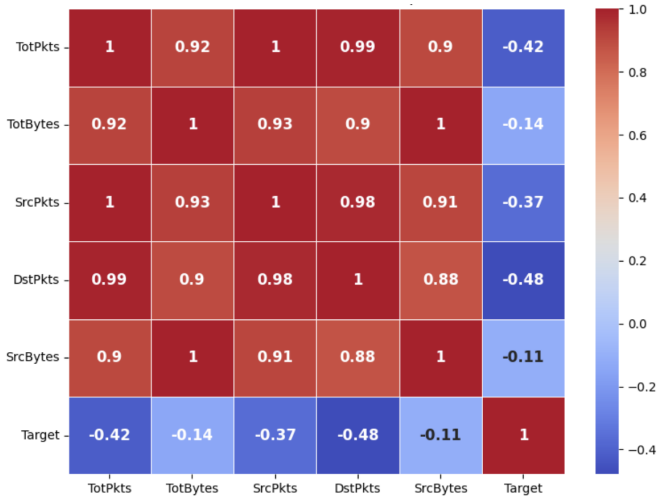
In Figure 3 and Figure 4, we can see the comparison of "Sport" and "TotPkts" data between the original message and generated intelligent attacks via GAN and Autoencoder. In particular, when plotted, the data points from the Autoencoder-

Fig. 5: The penetration performance by Autoencoder-based native GenAI.

reproduced attacks closely resemble those of the original control message, demonstrating a high degree of replication accuracy. On the other hand, the GAN-generated intelligent attack displays noticeable differences from the original control message. The Autoencoder outperformed the GAN in reproducing attack vectors, as demonstrated in Figures 3 and Figure 4. The Autoencoder's reconstructed data shows a high level of fidelity to the original data, particularly when compared through correlation matrices. This strong performance is further supported by significantly lower mean squared error (MSE) values, highlighting the Autoencoder's superior reconstruction accuracy over the GAN. Further, we analyze the capability of intelligent attack generation by the developed Autoencoder-based native GenAI in Figure 5.

We compare the newly generated intelligent attack vector in Figure 6. In particular, Figure 6a represents the original control and communication message of a smart grid framework, while Figure 6b describes the GenAI driven attack vector space. Further, Figure 6 shows that the GenAI-driven intelligent attack vector contains a new type of attack signature and has the capability to mimic the original control message to xerox the smart grid operation.

We then examine our proposed BBN-based protection framework by considering the newly generated intelligent attack vector to justify the efficacy of the proposed model. In Figure 7 and Figure 8 show how well the Autoencoder model preserves important message relationships by comparing key variables in both the original and Autoencoder-generated intelligent attack of control messages. Figures 7 and 8 illustrate that the mutual information between SrcBytes and TotBytes was 1.236 in the original message and 1.250 in the Autoencoder-generated message. This small difference demonstrates the model's ability to closely replicate the original messages. Similarly, the dependency between TotPkts and DstPkts remained largely consistent, with values of 0.963 and 0.964, respectively. These results indicate that the Autoencoder can create synthetic messages that closely resemble the original messages.



(a) Original message vector.

(b) GenAI attack vector.

Fig. 6: Correlation analysis between the input feature vector and the GenAI-induced adversarial attack vector within Smart Grid communication networks.

In summary, the developed Bayesian Belief Network (BBN)-based risk explanation and protection framework enables the understanding capabilities of new intelligently generated cyber attacks such as replay attacks by native GenAI. The calculated mutual information among the BBN node (i.e., smart grid operational parameters) of the BBN to determine the prominent attack features of a particular native GenAI-driven attack for protecting the smart grid operation from attacks.

V. Conclusion

In this research, we have developed a new framework for investigating the potential of intelligent attack generation capability on smart grid control massage by native generative AI models such as GAN and Autoencoder while developing a Bayesian Belief Network (BBN)-based risk explanation and protection mechanism to enable the understanding capabilities of such attack. We demonstrate the effectiveness of these models in simulating various attack scenarios, with Autoencoder exhibiting superior performance in preserving data accuracy. By constructing a BBN network using Pearson correlation

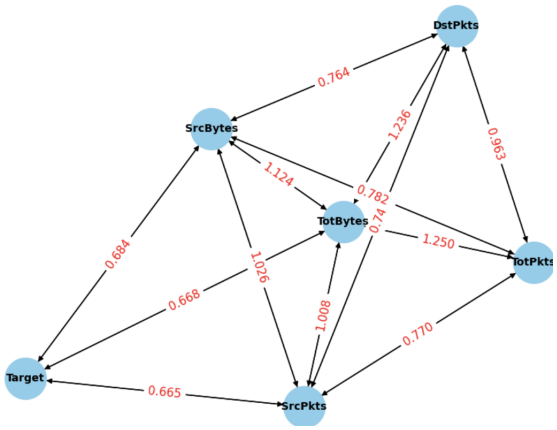


Fig. 7: Mutual information-based trust dependencies for the original smart grid communication data, representing authentic network behavior.

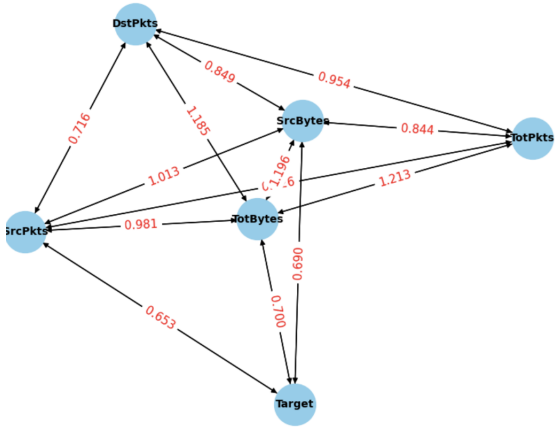


Fig. 8: Mutual information-based trust dependencies of GenAI (i.e., autoencoder) generated intelligent attack vector.

coefficients and calculating mutual information among the control message parameters, we can effectively detect and analyze new attack vectors generated by the AI models. This approach allows for a deeper understanding of the attack's characteristics. Overall, our findings highlight the critical need for advanced AI-based security solutions to protect smart grids from emerging cyber threats. As a result, the proposed framework can differentiate between original and generated data, which is crucial for detecting and mitigating adversarial attacks and ensuring the safety and reliability of smart grid operations.

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," *Computers and Electrical Engineering*, vol. 107, p. 108638, 2023.

[2] 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.

[3] M. K. Hasan, A. A. Habib, Z. Shukur, F. Ibrahim, S. Islam, and M. A. Razzaque, "Review on cyber-physical and cyber-security system in smart grid: Standards, protocols, constraints, and recommendations," *Journal of Network and Computer Applications*, vol. 209, p. 103540, 2023.

[4] 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, 2023.

[5] K. D. Kumar, M. A. Jawale, M. Sujith, and D. Pardeshi, "Cybersecurity threats, detection methods, and prevention strategies in smart grid: Review," in *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 2023, pp. 1609–1614.

[6] 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.

[7] M. S. Munir, S. Proddatoori, M. Muralidhara, W. Saad, Z. Han, and S. Shetty, "A zero trust framework for realization and defense against generative ai attacks in power grid," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 2482–2488.

[8] M. Muralidhara, M. S. Munir, S. Proddatoori, S. Shetty, and K. Gold, "Detecting attacks and optimizing routes in radio-frequency networks using machine learning and graph theory," in *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*, 2024, pp. 157–165.

[9] M. Razghandi, H. Zhou, M. Erol-Kantarci, and D. Turgut, "Variational autoencoder generative adversarial network for synthetic data generation in smart home," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 4781–4786.

[10] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[11] M. Zolanvari, "Wustl-iiot-2021, ieee dataport," https://dx.doi.org/10.21227/yftq-n22, accessed: September 1, 2023.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[13] S. Mishra, P. Mehta, N. Chouhan, N. Pethani, and I. Saha, "Edit any face – image synthesis using gan's," in *2022 International Conference on Futuristic Technologies (INCOFT)*, 2022, pp. 1–5.

[14] U. Michelucci, "An introduction to autoencoders," 2022.

[15] Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization," *IEEE Access*, vol. 8, pp. 42 169–42 184, 2020.

[16] M. S. Munir, S. H. Dipro, K. Hasan, T. Islam, and S. Shetty, "Artificial intelligence-enabled exploratory cyber-physical safety analyzer framework for civilian urban air mobility," *Applied Sciences*, vol. 13, no. 2, p. 755, 2023.

[17] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. suppl_2, pp. S231–S240, 2002.