2023 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia) | 979-8-3503-2774-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISGTAsia54891.2023.10372697

A Framework for Generating Attack Samples in Power Energy System using VAEs

Tran L.T Le, David K.Y Yau, Justin Albrethsen

tranlytu_le@mymail.sutd.edu.sg, david_yau@sutd.edu.sg, justin_albrethsen@sutd.edu.sg

***Abstract*—Real-world datasets from the cyber-physical systems (CPS) are indeed valuable for training and evaluating machine learning models in the context of cybersecurity analysis and defense against attacks. Generating realistic and practical attack samples is essential for understanding the vulnerabilities of power energy systems and developing more effective defenses. However, existing methods for generating attack samples must often be more balanced and realistic. This can lead to defense techniques that could be more effective against real-world attacks. Our work proposes a temporal convolutional network (TCN) based variational autoencoder (VAE) to learn disentangled attack characteristics from the real dataset and generate diverse time delay attack (TDA) sample sets. The TCN-VAE model's ability to capture complex temporal patterns and dependencies make it effective in representing attack features. In addition, integrating XGBoost as a classifier model, we compute correlation information between features and latent dimensions, gaining insights into feature importance for generating automatic generate control (AGC) frequency with desired characteristic features. Our framework addresses the imbalance in quantity and quality of the original dataset, leading to a more comprehensive and diverse dataset for improved model performance. Through two use cases, our work demonstrates that training on a large dataset with good quality results in better performance and higher correct classification accuracy and correct estimation accuracy of the cycle at which AGC frequency steps into unsafe levels in the power energy system.**

***Index Terms*—Data Generation, TCN-VAE(Time-Convolutional Variational Autoencoder), Power Energy System, Representative Learning, Machine Learning, Smart Grid, Time Series Data, Latent Space.**

I. Introduction

Real-world datasets from cyber-physical systems (CPS) are valuable for training and evaluating machine learning models for cyber defense. However, such datasets are typically limited in size and diversity. These limitations can impact a model's ability to generalize to different scenarios, which is crucial for robust cybersecurity applications. Additionally, the time-consuming and expensive nature of data collection in real CPS can lead to datasets that do not fully capture all possible variations and edge cases. Imbalance and bias in the real-world datasets are common issues that can further degrade the performance of machine learning tasks. Underrepresented events or scenarios and significant imbalances in sample size among scenarios make it challenging to develop accurate and reliable models. Two main aspects of a good data set: quantity and quality, are critical for successful machine learning applications. However, real CPS data might contain missing values or noisy measurements due to sensor errors or communication issues, leading to data size leakage after preprocessing. This further complicates the dataset's quality and makes it difficult to achieve the desired performance. Moreover, the dynamic and real-time nature of data collection in CPS environments make it challenging to reproduce the same data for experimentation and model validation. This leads to an even more imbalanced dataset, posing difficulties in identifying anomalies or attack behaviors in machine learning applications, especially in the energy domain. To tackle these challenges, our proposed work is to utilize the TCN-VAE model to learn disentangled attack characteristics from the real dataset and generating diverse attack sample sets. The model can understand complex temporal patterns and dependencies, allowing it to learn and represent attack features effectively. The integration of XGBoost as a classifier model is to compute correlation information between features and latent, providing insights into feature importance and guiding the generation of AGC frequency with desired characteristic features. By generating attack sample sets based on attack characteristics, our framework overcomes the imbalance in quantity and quality of the original dataset. This approach enables us to create a more comprehensive and diverse dataset for training and evaluation, leading to improved model performance.

Machine learning has been increasingly used in attack classification and mitigation for power grid cyberattacks [4-6]. [1] leveraged latent space representation to improve anomaly detection, while [2] explored interpretability of learned representations and attention mechanisms for identifying anomalies in energy time series data. However, the limited availability of real-world attack data hinders models from being trained on all possible scenarios, leading to performance degradation as the models may struggle to accurately identify new or emerging attacks.

Our paper makes valuable contributions to the generation of diverse time delay attack scenarios using collected attack patterns from simulated datasets. A time delay attack manipulates control signals to introduce delays in the system's response. This can cause instability, disrupt normal operation, and lead to cascading failures by delaying critical control commands such as power generation adjustments or load shedding.The proposed framework offers a systematic and effective approach to enhance the diversity of attack scenarios. We introduce the novel concept of generating time-series data based on attack data characteristics from simulated datasets, which has the potential to impact power system security research.

Emphasizing the importance of dataset size and quality on model performance, we demonstrate that a model trained with a high-quality dataset achieves higher accuracy in estimation and classification tasks. In summary, our research significantly contributes to the field of power system security, providing new insights for further exploration in this area.

The paper is categorized into five sections: section I includes Introduction. Simulation design and explanation of our simulation datasets in section II. Section III introduces and explains our proposed frameworks. Results and discussion part is in section IV, before our discussion is concluded in section V.

II. Materials

A. Dataset Simulation Design

The TCN-VAE model is trained on sequential time series data simulated from Powerworld simulator, inspired by previous work [4]. Using a 37-bus system with three areas as a case study is suitable, representing a realistic small to mid-scale grid found in about a third of national grids [4]. The simulation in PowerWorld enhances dataset authenticity, aligning it with real-world data. The dataset focuses on time-delayed attacks against ACE signal transmission, which can significantly impact power grid stability and reliability [5]. Time delay magnitude is varying between 1 to 10, which indicates the delay timing of events in the power system, can lead to incorrect decision-making by protective relays, control devices, and other components of the grid. By adjusting load change and capacity parameters in the simulator, different utilization scenarios affecting AGC frequency are introduced, inducing stronger and more apparent attack disturbances. This approach provides a comprehensive dataset with varying utilization conditions, reflecting real-world power system scenarios and dynamics.

B. Data Preprocessing

Simulations were conducted for 300 cycles, excluding the initial 13 cycles to allow the system to stabilize. Attacks randomly start between 60 to 80 cycles. Demand changes are categorized into three levels (high, medium, and low) to capture different load scenarios, significantly impacting AGC frequency and leading to more substantial disturbances during the attack. Delays less than three cycles are ignored to ensure meaningful and significant effects on AGC frequency [4]. A window size of 64, after striding, effectively captures all attack characteristics, including the maximum and minimum attack values of AGC frequency and the cycles where the frequency reaches max or min peak or starts to become unsafe. Approximately 18,000 samples were split into 80% for training and 20% for evaluation in our work.

III. Proposed Framework

Inspired by the outstanding performance of TCNs compared to LSTMs in forecasting energy-related time-series data [3], our proposed framework leverages the strengths of TCNs and VAEs to handle energy-related time-series data. The framework consists of two main parts that is illustrated in figure 1.

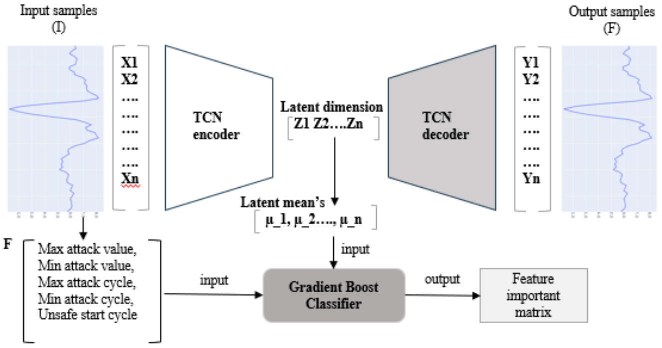


Fig. 1. Framework architecture

1. Integrating TCNs as hidden layers in VAE architecture allows for effective learning and disentanglement of data features. TCNs are well-suited for handling time series data due to their ability to capture complex temporal patterns and dependencies present in the energy-related time-series data. When combined with VAEs, the model can extract interpretable and disentangled latent representations of the time-series data, facilitating a deeper understanding of the underlying patterns and dynamics.

2. The XGBoost classifier (Extreme Gradient Boosting) is utilized to classify attack features importance level to the latent space representation. This information is crucial for generating AGC frequency with specific attack characteristic features within the preferred value range.

A. TCN-VAE Model

The default parameter settings for the TCN model may not be optimal for all datasets and tasks. Through experimentation on our power energy dataset, we chose 2 stacks, a kernel size of 3, and 32 filters, enabling TCN layers to capture long-term dependencies in time-series data. Exponentially increasing dilation for subsequent layers (e.g., 2, 4, 8, ...) enhances the receptive field, effectively capturing broader context and dependencies. Combining TCN and VAE ensures our model achieves good reconstruction quality and disentanglement of attack features.

The choice of the latent dimension significantly affects the performance and behavior of the TCN-VAE model, impacting the mean square error (MSE) and Kullback-Leibler (KL) divergence, essential components in VAE's loss function. A model with 6 latent dimensions strikes a reasonable balance between MSE and KL scores, allowing us to generate well-matched attack samples closely resembling real attack samples in the distribution of attack features.

B. Gradient Boost Classifier

Ensemble Gradient Boosting is employed to assess the contributions of each latent dimension in representing attack features, utilizing latent values from the VAE encoder. The calculated importance score measures the impact of each latent

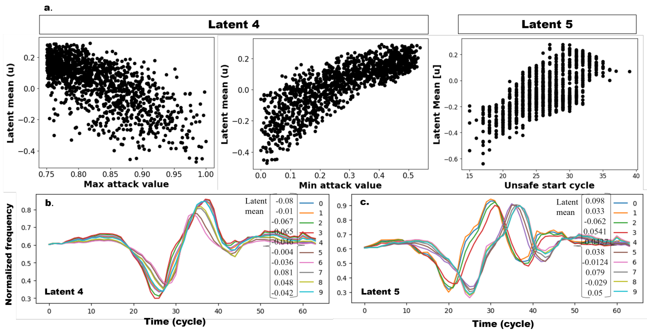


Fig. 2. a. Correlation plot, b. Fake samples generated from latent 4th controls feature max and min attack values, c. Fake samples generated from latent 5th controls feature unsafe start cycle.

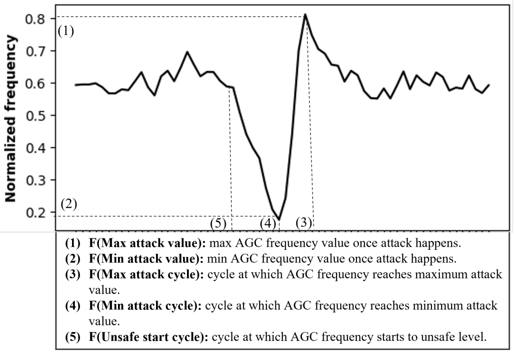
dimension on attack characteristics, guiding the generation of AGC frequencies within the preferred range. The model achieved an accuracy score of 82%, demonstrating strong performance in the classification task. A latent impact matrix evaluates the effect of each latent dimension on the attack characteristic prediction, with higher scores indicating a greater impact. Figure 3 explains the feature definitions that support the important matrix table. Table 1 shows high scores for the 4th and 5th latent dimensions, confirming the TCN-VAE's proficiency in learning disentangled representations, where a change in one latent dimension corresponds to a change in attack characteristics.

C. Latent control to generate attack samples.

The correlation plot shows how strongly each attack feature is correlated with each latent dimension and the disentanglement score measures how well each latent dimension captures a single attack feature. As shown in figure 2a, the 4th latent dimension exhibits a strong correlation with both the maximum and minimum attack peak values. This suggests that this particular latent dimension can be utilized to generate attack samples with specific ranges of values for these features. In Figure 2b and figure 2c, fake attack samples are generated by controlling the 4th and 5th latent dimensions independently. Figure 2b showcases fake attack samples with varying maximum and minimum peak values, while figure 2c presents fake attack sample traces with altered unsafe start cycles.

TABLE I
IMPORTANT MATRIX

latent_i	F(1)	F(2)	F(3)	F(4)	F(5)
1	0.09	0.03	0.02	0.14	0.06
2	0.12	0.03	0.07	0.09	0.07
3	0.07	0.01	0.02	0.11	0.07
4	**0.45**	**0.75**	0.02	0.18	0.12
5	0.15	0.13	**0.47**	**0.26**	**0.5**
6	0.12	0.05	0.4	0.22	0.18



(1) F(Max attack value): max AGC frequency value once attack happens.
(2) F(Min attack value): min AGC frequency value once attack happens.
(3) F(Max attack cycle): cycle at which AGC frequency reaches maximum attack value.
(4) F(Min attack cycle): cycle at which AGC frequency reaches minimum attack value.
(5) F(Unsafe start cycle): cycle at which AGC frequency starts to unsafe level.

Fig. 3. Features definition

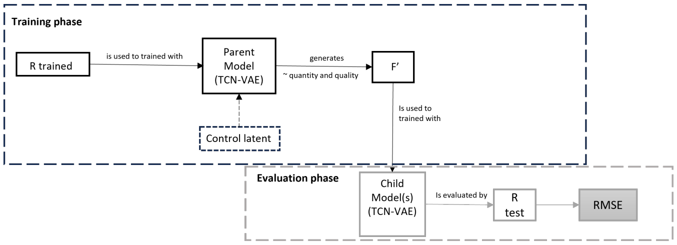


Fig. 4. Evaluation flow

IV. Results and Discussion

Our proposed work focuses on the crucial aspect of attack dataset generation, aiming to deliver not only massive dataset volume but also its quality. Having a large and high-quality dataset is indeed crucial for training models effectively. In the following subsections, we provide evidence of how the attack dataset generated from our proposed work affects general model performance, as well as classification and unsafe attack start cycle estimation results. This evidence will help establish the relationship between dataset characteristics and model performances, validating the effectiveness of our proposed method. The parent model learns from the actual data and captures the underlying patterns present in real-world scenarios. It is then leveraged to generate diverse quantities and qualities of attack data, effectively expanding the dataset for further evaluation. The child models train on these fake attack-sample sets. This allows the child models to learn and adapt to the characteristics and patterns of the generated data. This step assesses how well the child models can utilize the information encoded in the fake attack samples. Evaluating the child models' performance on the real dataset demonstrates how well the child models can predict or reconstruct real attack samples. This indicates the effectiveness of the attack dataset generation process.

A. Model performance

1) Dataset size impact

It takes years to generate a dataset from a real CPS system, and an insufficient attack dataset is a drawback that downgrades the efficiency of machine learning applications in the power energy domain. In machine learning, models trained with a high

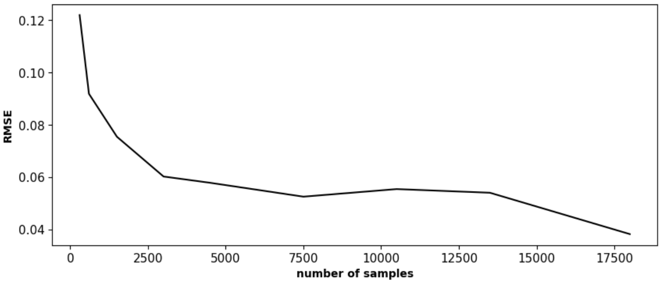


Fig. 5. Dataset size impact to model performance

quantity of samples tend to exhibit much better performance than models trained with a small quantity. Figure 5 shows that our model's performance increases gradually, indicated by lower RMSE metric scores, as the quantity of samples ascends. This demonstrates that the more attack samples are input to the model, the more accurately the attack features can be mimicked.

1) Dataset quality impact

Dataset quality effects model performance, with good quality defined as covering all possible attack feature values and having an even number of samples among those value ranges. The hypothesis is that training models with good quality datasets can enhance performance even more than large dataset volume. However, real CPS often lacks sufficient good quality datasets, presenting a natural challenge. Our proposed work addresses this issue by generating more attack samples to fill the gaps in the original dataset. Models learn attack features from the real dataset collected from CPS and then control latent dimensions to generate additional fake samples with desired value ranges, volumes, and quality.

To test our hypothesis, we train three child models with even, uneven, and special uneven generated datasets collected from the parent model and collect RMSE metric scores for comparison. Even datasets have balanced sample numbers, while uneven and special uneven datasets exhibit imbalances, with the latter lacking specific value ranges. To ensure fairness, all three datasets have an equal sample size of around 5,000 and cover the exact same physical-feature values range. Table II illustrates that our model performs best when trained with an even dataset compared to other sets, confirming the validity of our hypothesis. Our proposed work provides a large and high-quality attack dataset, facilitating better attack feature learning in models.

TABLE II
QUALITY OF DATASET IMPACT ON MODEL PERFORMANCE

TCN-VAE	**RMSE**
Even	**2.19E-05**
Uneven	3.80E-02
Special Uneven	5.24E-02

B. *Use cases*

This section includes two use cases in power energy domain that illustrates the evidence proving that how impact and supportive a large and good quality generate dataset to estimation or classification accuracy.

Use case 1: Unsafe-attack start cycle estimation
After the attack executes, the system maintains stability and safety for a few cycles as the grid has an automatic restore mechanism until it exceeds its ability and becomes unstable. According to [6], once the system is detected as falling to an unsafe situation, mitigation is necessary to restore safety. Our experiment aims to assess the impact of dataset quality on the model's ability to accurately estimate the cycle when the system transitions to an unsafe zone, ensuring grid stability and safety in the energy domain. The XGBoost regression model is trained using 10,000 pairs of latent features ”unsafe start cycle” and mean vectors from the parent TCN-VAE model. The test set consists of latent features and mean vectors from the child models' latent dimensions, trained with diverse quality generated attack datasets. The primary objective is to evaluate how accurately the XGBoost regression model can estimate the unsafe start cycle based on provided latent features and mean vectors. Accuracy is measured using residuals, which are the differences between the predicted “unsafe start cycle” values and the ground truth values. Estimation is accurate when the predicted unsafe cycle values fall within the residual range. The number of correct estimation cases over the total number of samples is estimated accuracy percentage. Table III demonstrates that the model trained with an even fake sample set consistently outperforms the child models trained with uneven sets. The findings confirm the importance of a good quality and evenly distributed dataset for accurate estimation. The superior performance of the model trained with fake sample sets validates the effectiveness of our approach in ensuring grid stability and safety.

TABLE III
UNSAFE START CYCLE ESTIMATION ACCURACY (%)

Residuals	**Even**	**Uneven**	**Special uneven**
[-5,5]	67	50	35
[-7,7]	83	79	55.4

Use case 2: Unsupervised classification
Both TDA and false injection attacks (FDI) are cyber-attacks that disrupt the power energy system's normal operation, differing in their approach techniques. TDA manipulates control signals to introduce time delays in the system's response, while FDI manipulates data sent to the power system's control devices or sensors. Our work focuses on unsupervised classification between TDA and FDI attack cases, using the TCN-VAE model trained with TDA attack samples and RMSE scores. We create three different quality fake datasets, each containing around 5,000 samples, to train three child models. A test set of approximately 3,000 real attack samples, containing both attack types, is used for evaluation. Our primary objective is

to assess how accurately the TCN-VAE child models, trained with different quality datasets, can classify TDA and FDI attack types. The classification accuracy, calculated as the total number of correct classifications divided by the total number of test samples, measures the model's ability to correctly identify FDI and TDA attack cases. The results in Table IV demonstrate that dataset quality significantly impacts the classification outcome. Notably, using a high-quality training dataset improves the accuracy (91.72%) and true positive rate (80.02%), leading to more accurate classification of FDI attack samples. These findings highlight the importance of having a good quality training dataset for unsupervised classification tasks. Our proposed approach shows promise in effectively distinguishing between different types of attack samples, enhancing security measures in power energy systems.

TABLE IV
IMPACT OF DATASET QUALITY ON CLASSIFICATION ACCURACY (PERCENTAGE)

	Area under ROC	TPR
Even	**91.72**	**80.02**
Uneven	84.71	73.30
Special uneven	85.10	73.81

C. Model comparison with baseline models

In our work, TCN-VAE serves as the primary method, but comparing it with baseline models is crucial to understand its reconstruction and prediction capabilities comprehensively. To ensure a fair comparison, all three models are trained and evaluated using the exact same dataset samples, with equal size and scenarios. The training progress is not constrained by a limited number of epochs, allowing each model to converge to its optimal state and maximize learning potential. The training automatically terminates after 10 epochs if no significant progress is observed. TCN-VAE demonstrates the lowest MSE score compared to baseline models, while vanilla VAE performs relatively weaker with the highest MSE score. Consistent results across other evaluation metrics like MAPE and MAE also support the conclusion that LSTM-VAE performs well but falls behind TCN-VAE. Considering the parameter trained epochs for performance assessment, vanilla VAE achieves optimal performance with the least number of epochs but falls short compared to TCN-VAE. Although LSTM-VAE demands the highest number of epochs, approximately six times more than TCN-VAE, it still fails to outperform TCN-VAE. Table V presents the comparison results, strongly suggesting that TCN-VAE delivers the best overall performance, combining superior performance and reasonable training time.

TABLE V
TCN-VAE PERFORMANCE COMPARISON TO BASELINE MODELS

Model	MSE	MAPE	MAE	Trained epoch(s)
1D-CNN VAE	1.41E-04	1.37E-00	7.54E-03	200
LSTM-VAE	8.05E-05	1.10E-00	6.12E-03	2900
TCN-VAE	**2.19E-05**	**5.92E-01**	**3.24E-03**	**500**

V. CONCLUSION

This paper presents a novel TCN-VAE-based technique to generate additional time-series data, specifically for time delay attack patterns in AGC frequency. Our proposed framework provides an effective solution that harnesses the strengths of machine learning models to address challenges in real-world CPS datasets, advancing cybersecurity research. It emphasizes the importance of generating high-quality attack sample sets to enhance the security and reliability of energy systems. The results underscore the significance of dataset quality in the success of unsupervised classification and cycle estimation tasks. A model trained with a high-quality dataset demonstrates higher accuracy in both tasks, validating the practical relevance of our approach in enhancing energy system security. Additionally, the work compares the reconstruction performance of other baseline models, VAE, and LSTM-VAE, using evaluation metrics like MSE, MAE, and MAPE. The results conclude that TCN-VAE exhibits outstanding performance, while VAE performs the weakest overall on our dataset.

REFERENCES

[1] D. Kaur, S. N. Islam and M. A. Mahmud, ”A Variational Autoencoder-Based Dimensionality Reduction Technique for Generation Forecasting in Cyber-Physical Smart Grids,” 2021 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, QC, Canada, 2021, pp. 1-6, doi: 10.1109/ICCWorkshops50388.2021.9473748.

[2] J. Pereira and M. Silveira, ”Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention,” 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 1275-1282, doi: 10.1109/ICMLA.2018.00207.

[3] Lara-Benítez P, Carranza-García M, Luna-Romera JM, Riquelme JC. Temporal Convolutional Networks Applied to Energy-Related Time Series Forecasting. Applied Sciences. 2020; 10(7):2322. https://doi.org/10.3390/app10072322

[4] S. Ghahremani, R. Sidhu, D. K Yau, N.-M. Cheung, and J. Albrethsen, “Defense against Power System Time Delay Attacks via Attention-based Multivariate Deep Learning,” in IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), to be published. IEEE, 2021, pp. 1–6.

[5] S. Ghahremani, D. K. Y. Yau, J. Albrethsen, R. Sidhu and N. -M. Cheung, ”Time Delay Attack Detection Using Recurrent Variational Autoencoder and K-means Clustering,” 2021 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia), Brisbane, Australia, 2021, pp. 1-5, doi: 10.1109/ISGTAsia49270.2021.9715557.

[6] X. Lou, C. Tran, R. Tan, D. K. Yau, and Z. T. Kalbarczyk, “Assessing and mitigating impact of time delay attack: a case study for power grid frequency control,” in Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems, 2019, pp. 207–216.