ADScAI
Applied Data Science & Artificial Intelligence Symposium

Generative AI for Cybersecurity: Crafting Network Intrusion Datasets

Mathushaharan Rathakrishnan
Dept. of Electronic and Telecommunication Engineering
University of Moratuwa
Moratuwa, Sri Lanka
mathushaharan28@uom.lk

Samiru Gayan
Dept. of Electronic and Telecommunication Engineering
University of Moratuwa
Moratuwa, Sri Lanka
samirug@uom.lk

***Keywords*—Intrusion detection systems, generative adversarial networks, multi-model framework, explainable AI, data privacy.**

I. Introduction

Rapid advances in networking technology and the increasing complexity of cyberattacks have created the need for Artificial Intelligence (AI)-powered Intrusion Detection Systems (IDSs). However, the performance of AI-based IDS is limited due to the lack of labeled intrusion data, class imbalances, and restrictions to share intrusion data due to the General Data Protection Regulation (GDPR). In addition, it is expensive and risky to simulate attacks and collect intrusion data.

Recently, Generative Adversarial Networks (GANs) have proven to be promising solutions for synthesizing data. However, most current GAN architectures are designed with Gaussian-like data distributions in mind (such as images) and are difficult to adapt to tabular data characterized by non-Gaussian, mixed-type, and multi-modal features.

To address the above-mentioned limitations, we propose a multi-model framework based on Conditional Tabular GAN (CTGAN) [1] that can be used to synthesize network intrusion data. The framework incorporates rigorous testing and validations to ensure the accuracy and real-world applicability of synthesized intrusion data.

II. Literature Review

Generative AI methods, particularly GANs, have demonstrated strong capabilities for synthetic data generation. Esteban et al. [2] introduced Recurrent GAN (RGAN) for medical time-series data generation. Similarly, [3] presented an approach for generating discrete (binary) data using adversarial training (CorrGAN). Moving forward, in [4], the authors presented table-GAN, which is a model that can synthesize tabular data with categorical and continuous variables. However, this model primarily handles Gaussian-distributed data and struggles with multi-modal continuous features common in real-world scenarios.

Recently, CTGAN has been successfully applied to network traffic synthesis [5] and shown improved performance than Topological Auto Encoder (TVAE) [6]. Nevertheless, existing literature lacks dedicated frameworks for synthesizing and rigorously validating network intrusion data, especially with integrated explainability.

III. Methodology

A. Dataset and Preprocessing

This research uses the NSL-KDD dataset [7], which contains network intrusion records categorized into basic, content-based, and time-based features. We preprocess the data by converting attack labels to binary labels (attack, normal), scaling numerical values to [0,1], and encoding categorical features using one-hot encoding.

B. Proposed Multi-Model Framework

Fig. 1 illustrates our proposed multi-model framework comprising three main components:

- Conditional Tabular GAN (CTGAN) for synthetic intrusion data generation.
- Explainable AI (XAI)-enabled Artificial Neural Network (ANN)-based binary classifier for Train-on-Real, Test-on-Synthetic (TRTS) approach.
- ANN-based binary classifier for Train-on-Synthetic, Test-on-Real (TSTR) approach.

The CTGAN architecture consists of two neural networks: the generator and the critic. The generator is responsible for data synthesis, while the critic improves its performance through continuous feedback by estimating the distance between learned conditional distribution and conditional distribution on real data. Additionally, two binary classifiers assist in evaluating the synthesized data using the TRTS and TSTR methods as depicted in Fig. 1. CTGAN uses Training-by-sampling and mode-specific normalization to perform better on data synthesis compared to other GAN architectures when it comes to multi-modal tabular data.

https://doi.org/10.31705/ADScAI.2025.22

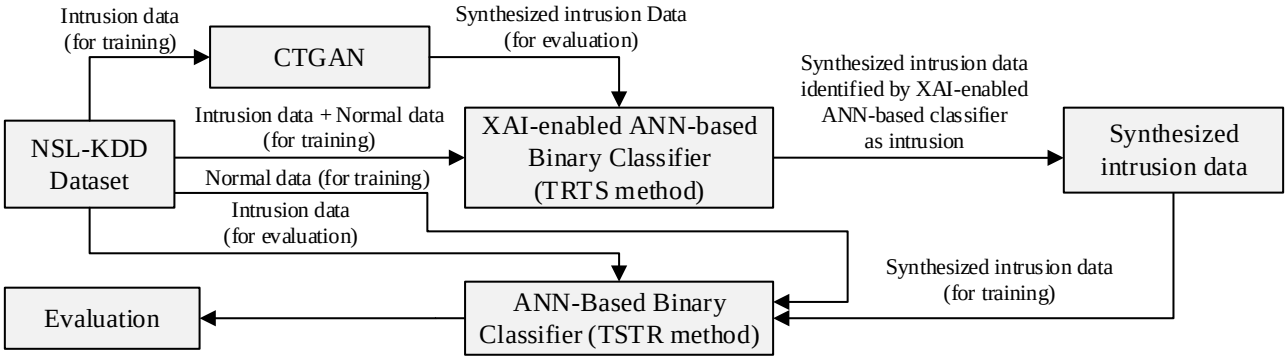

Fig. 1. Proposed multi-model framework.

IV. Results and Discussion

The proposed multi-model framework was evaluated through qualitative and quantitative assessments. Qualitative assessment using overlaid density plots confirmed that CTGAN-synthesized data matched real-world distributions accurately. As an example, Fig. 2 shows the distribution of the dst_host_srv_count feature, which represents the number of connections in the past 100 connections to the same destination host that used the same service (e.g., HTTP, FTP). The high overlap demonstrates CTGAN's ability to accurately capture intrusion-related feature patterns. For quantitative evaluation, we employed the TRTS method, the TSTR method, and the Kolmogorov–Smirnov (KS) test. In the TRTS method, a binary classifier trained on real intrusion data and normal data correctly classified 96.19% of the synthesized intrusions with 98.36% classification accuracy. Additionally, Local Interpretable Model-Agnostic Explanation (LIME) was integrated on top of TRTS to provide interpretability. In the TSTR method, a classifier trained on synthesized intrusions and real-world normal data identified 98.59% of real-world intrusions as intrusions with 97.95% classification accuracy, demonstrating real-world applicability of the CTGAN-synthesized intrusion data. The KS test further validated the similarity between real and synthesized data with KS statistic values near zero and an average similarity score of 0.92. As a limitation, while LIME was applied in the TRTS method, no explainability was integrated into the CTGAN model itself. Future work could explore embedding XAI into the CTGAN to better interpret the intrusive nature of the synthesized data.

V. Conclusion

This paper proposed a multi-model framework for synthesizing network intrusion data and validated it through qualitative and quantitative approaches. The synthesized data achieved a similarity score of 0.92.

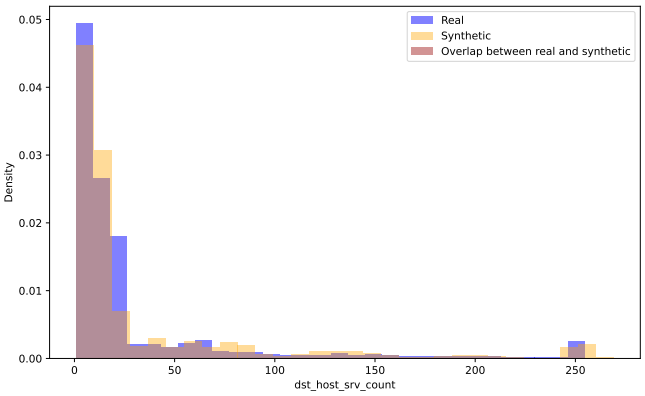

Fig. 2. Overlaid density histogram comparing real and synthetic distributions of the feature dst_host_srv_count. This feature represents the number of connections in the past 100 connections to the same destination host that used the same service (e.g., HTTP, FTP). The high overlap indicates that CTGAN accurately captures the underlying distribution of intrusion-related features.

References

[1] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 659.

[2] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," *arXiv preprint arXiv:1706.02633*, 2017. [Online]. Available: https://arxiv.org/abs/1706.02633

[3] S. Patel, A. Kakadiya, M. Mehta, R. Derasari, R. Patel, and R. Gandhi, "Correlated discrete data generation using adversarial training," *arXiv preprint arXiv:1804.00925*, 2018. [Online]. Available: https://arxiv.org/abs/1804.00925

[4] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proc. VLDB Endow.*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018, doi: 10.14778/3231751.3231757.

[5] M. N. A. Khatiman, A. Abu-Samah, M. A. Azman, R. Nordin, and N. F. Abdullah, "Generation of synthetic 5G network dataset using generative adversarial network (GAN)," in *Proc. IEEE 16th Malaysia Int. Conf. Commun. (MICC)*, 2023, pp. 141–145.

[6] M. Moor, M. Horn, B. Rieck, and K. Borgwardt, "Topological autoencoders," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, p. 653.

[7] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "NSL-KDD dataset," 2009. [Dataset]. An improved version of the KDD'99 dataset for network intrusion detection research.