DEEP LEARNING IN NETWORK TRAFFIC ANALYSIS USING SYNTHETIC DATA FOR PRIVACY PROTECTION AND CYBER ATTACK MITIGATION

Swikriti Neupane , Dr Pratap Sahu

Department of Computer Science, School of Natural Sciences and Mathematics
Claflin University, Orangeburg, SC

UR 2 PhD
Undergraduate Research to PhD Mentoring

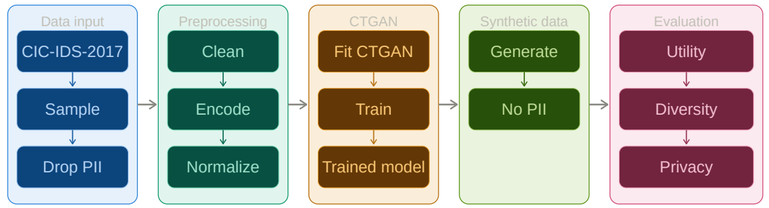
Background

- **Intrusion Detection Systems (IDS)** are used to monitor and detect abnormal or malicious network activity. They rely on large, labeled datasets for training machine learning models effectively.
- **Challenge with Real Data:** Real-world network datasets contain Personally Identifiable Information (PII) such as IP addresses, hostnames, and user credentials. Sharing these datasets directly creates serious privacy and security risks, which limits collaboration and open research.
- **PII Anonymization:** Before any processing, sensitive fields are identified using keyword matching and removed from the dataset. Techniques include tokenization (e.g., IP → IP_001) and generalization (e.g., exact timestamp → date only), ensuring sensitive details are hidden prior to model training.
- **Synthetic Data Generation:** CTGAN (Conditional Tabular GAN) is used to generate synthetic network traffic that statistically mirrors real data distributions. The synthetic data is validated on three key factors:

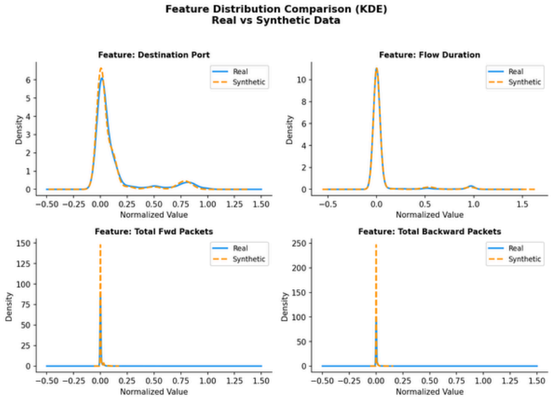
Objective

- Remove personally identifiable information (PII) from network datasets while maintaining their utility for research.
- Generate high-quality synthetic datasets using GAN-based models (CTGAN).
- Validate synthetic data on utility, diversity, and privacy metrics.
- Enable secure data sharing for IDS research without compromising sensitive information.

Pipeline



Methodology



Dataset Used the CIC-IDS-2017 dataset (Friday PortScan file) because it reflects both real network activity and multiple types of cyberattacks. The full dataset contains 286,467 records and 79 features. A random sample of 5,000 records was used for efficient model training.

PII Detection Applied keyword-based pattern matching to locate personal data fields. In this dataset, Source IP, Destination IP, and Timestamp were identified as PII. Other common PII types (MAC addresses, hostnames, URLs) were checked but were not present, as this CSV is a pre-processed feature extraction output from CICFlowMeter.

Anonymization:

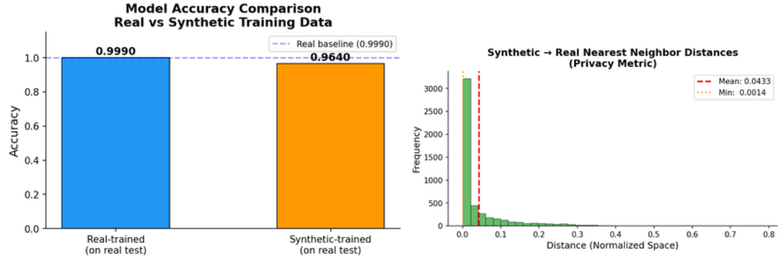
- Tokenization — replaced sensitive values with coded labels (e.g., IP_001).
- Generalization — converted specific timestamps to date-level granularity.
- Drop strategy — PII columns were removed entirely before any model training.

Synthetic Data Generation CTGAN was trained on the preprocessed dataset for 300 epochs to learn the statistical distributions of real network traffic. The Label column was treated as a discrete variable. After training, 4,994 synthetic samples were generated matching the size of the real dataset.

Evaluation

- Utility : trained two Random Forest classifiers (one on real data, one on synthetic data) and compared accuracy on a shared held-out real test set to detect any performance drop.
- Diversity : compared label distributions and feature distributions using KDE plots and bar charts.
- Privacy : used Nearest Neighbors (k=1) to measure how far each synthetic sample is from its closest real record in normalized feature space. Larger distances indicate better privacy protection.

Results & Analysis



To evaluate the synthetic data pipeline, real and synthetic datasets were compared across three dimensions: Utility, Diversity, and Privacy.

- Utility: A Random Forest model trained on synthetic data achieved comparable accuracy to one trained on real data when tested on the same held-out real test set, confirming minimal performance drop ($\Delta \approx 0$).
- Diversity: Label distributions were consistent across real and synthetic datasets. KDE plots confirmed that CTGAN successfully learned the underlying data patterns without simply repeating original samples.
- Privacy: Nearest neighbor distances computed in normalized feature space confirmed that synthetic records are not copies of real data, with no exact duplicates found.
- PCA Projection: Overlapping clusters of real and synthetic data in 2D feature space confirm that synthetic data is statistically representative of the original dataset.

Conclusion & Future Plans

- Synthetic data generation successfully preserved model performance with minimal accuracy drop ($\Delta \approx 0$), confirming it as a viable privacy-safe substitute for real network traffic data in IDS research.
- Privacy was validated through nearest neighbor distances in normalized feature space, confirming no exact copies of real records exist in the synthetic dataset.
- Future work will explore advanced generative models such as Graph Neural Networks and diffusion models to improve synthetic data quality across all 79 features.
- Automated end-to-end pipelines and better handling of rare attack class imbalance will be prioritized to make the framework more robust and deployable in real-world IDS systems.

References

1. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of ICISSP, 108–116.
2. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems (NeurIPS), 32.
3. Goodfellow, I., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 27.