NORTH CAROLINA AGRICULTURAL AND TECHNICAL STATE UNIVERSITY

Enhancing CPS Trustworthiness and Reliability through Synthetic Data and Ensemble learning

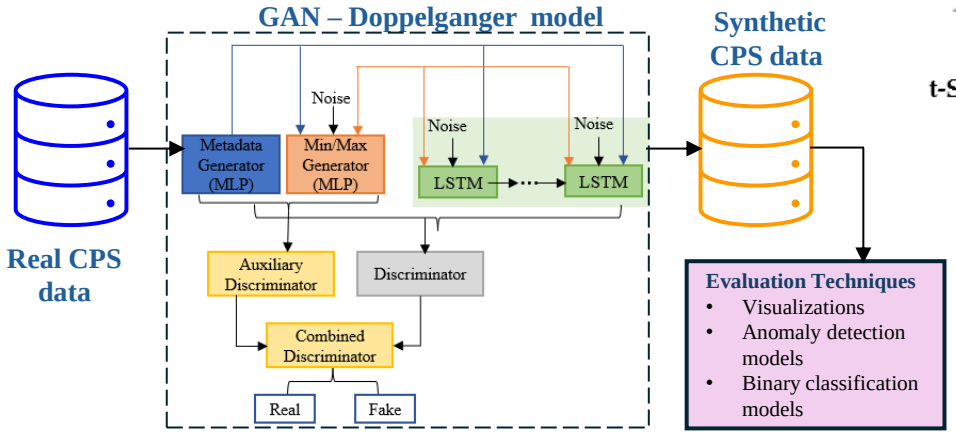
Why use synthetic data in CPS?

- Development and validation of models, algorithms, and control systems for Cyber-physical Systems (CPS) heavily rely on the availability of datasets.
- Public datasets are limited due to:
  - time constraints,
  - Privacy concerns,
  - high costs, and
  - the need for specialized technical expertise in data collection.
- Promising approach – creation of synthetic CPS datasets using Generative Adversarial Networks (GANs).

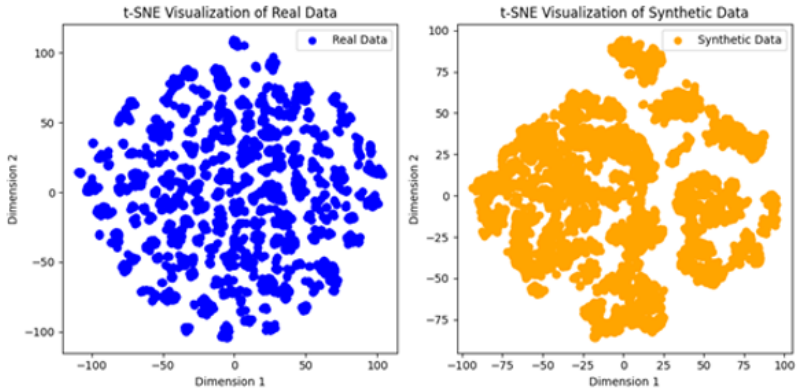
Objectives

- Generate quality synthetic data to improve training and validation.
- Evaluate the effectiveness of the synthetic dataset.
- Enhance system reliability and safety.

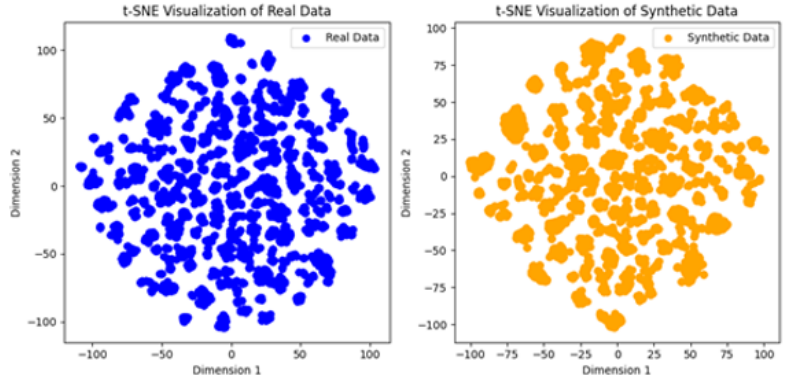
Methodology: Water Distribution System CPS Application



Results Comparison based on t-SNE Visualization



t-SNE Visualization of Synthetic and Real CPS Data for 250 Epochs.



t-SNE Visualization of Synthetic and Real CPS Data for 1000 Epochs.

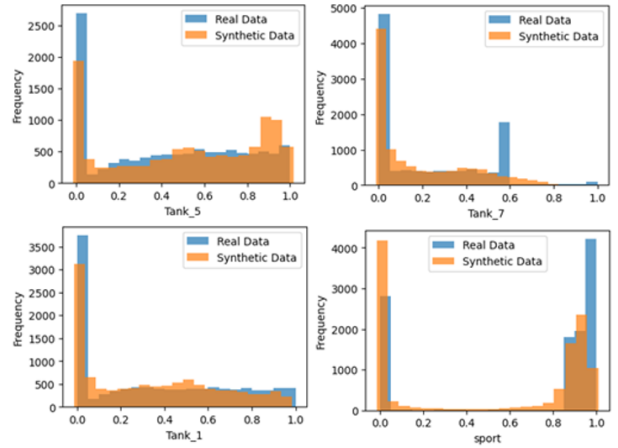
Comparative Assessment of PCA-Isolation Forest: Real vs. Synthetic CPS Data

Dataset	Precision (weighted)	Recall (weighted)	F1-score (weighted)	Accuracy
Synthetic Data	0.87	0.84	0.78	0.84
Real Data	0.68	0.80	0.73	0.80

Comparative Analysis by Binary Classification Models

	Accuracy		F1 score	
Models	Real	Synthetic	Real	Synthetic
RF	0.980	0.990	0.930	0.992
DT	0.738	0.880	0.719	0.837
AdaBoost	0.900	0.959	0.776	0.963
XGBoost	0.748	0.887	0.819	0.899
LR	0.678	0.919	0.702	0.912

Comparison of RF Top Four Important features: Real vs Synthetic CPS Data



Conclusion

- The results suggest a reasonable alignment between the real and synthetic datasets.
- The synthetic data generated successfully captures essential patterns.