SECURE TIME SERIES ENERGY DATA WITH GENERATIVE AI

Student
Thi Van Dai Dong

Supervisors
Sam West, Dr Mahathir Almashor

Introduction

Data sharing is constrained by concerns about privacy and legal matters. Data synthesis emerges as a prominent solution to address the challenges associated with data sharing currently. Our project aims to:

- Generate time-series energy data using a Generative Adversarial Network (GAN) model.
- Implement privacy protection filters within the generative model to prevent potential attacks.

Data Source

Internal Data: The dataset contains energy consumption data recorded at 30-minute intervals.
Public Data: The dataset includes 30-minute interval meter readings of electricity consumption and generation for each participating household in the Smart Grid Smart City project.

Methods

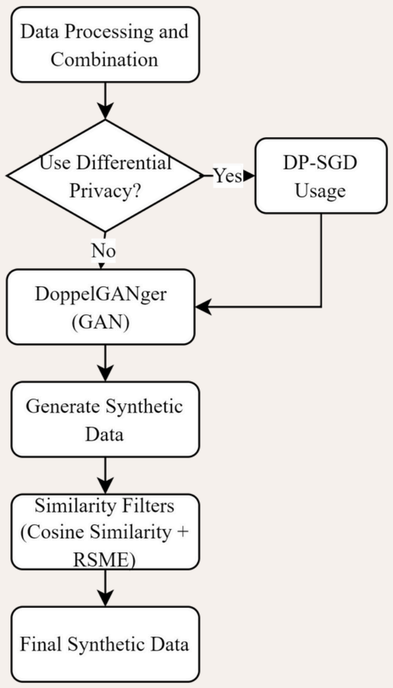

Fig 1. Experiment Flow

Differentially Private Stochastic Gradient Descent (DP-SGD)

Differentially Private Stochastic Gradient Descent (DP-SGD) is used to adjust the gradients utilised in stochastic gradient descent (SGD) in order to obscure the distribution of certain parts of the data.

DoppelGANger

- DoppelGANger was first introduced in 2020 to address issues:
  - The inability to capture long-term effects.
  - Mode collapse.
- The generator employs **LSTM** to sequence data and output multiple time points to enhance **temporal correlations**.
- The **Wasserstein loss** is implemented to mitigate vanishing gradients and address **mode collapse.**

Evaluation

Visualizations

- TSNE: Assess the similarity and diversity in data distributions.
- Time-series plots

Train Synthetic, Test on Real

- Two models are trained: one on the real dataset and another on the synthetic dataset.
- Evaluation is conducted using the test set from the real dataset.
- Performance assessed using **RMSE** and **MAE**.
  - Ideally, the RMSE and MAE scores of a model trained on synthetic data **closely match** those of a model trained on real data, with scores **approaching 0**.

References

Results

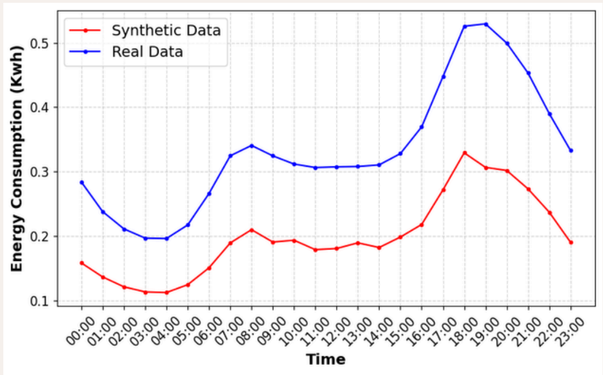

Fig 2. Mean Hourly Energy Consumption Across All Households

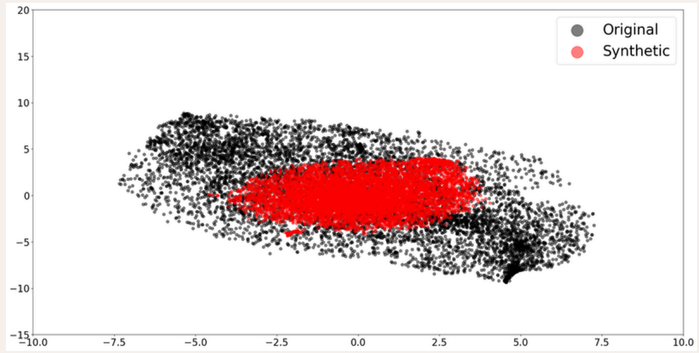

Fig 3. T-SNE Plot: Data Diversity and Distribution

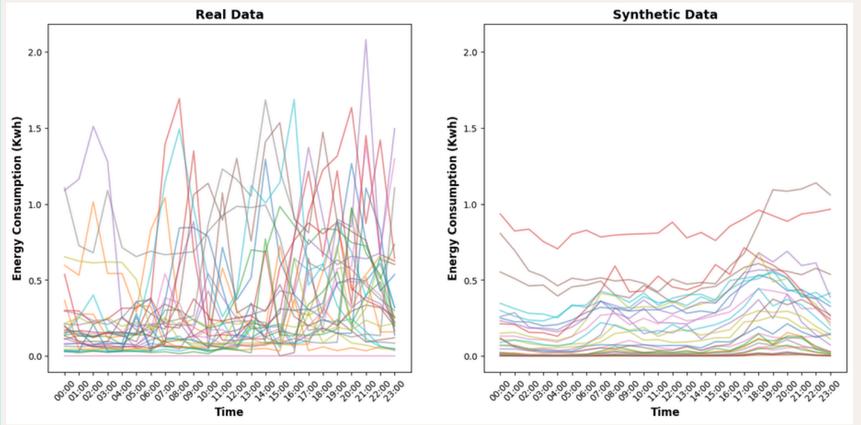

Fig 4. Hourly Energy Consumption of 30 Random Households

	RMSE	MAE
Real	0.11	0.07
Synthetic	0.15	0.15

Tbl 1. RMSE and MAE Scores on Test Set (Train Synthetic Test Real method)

- The time-series plots suggest that the synthetic dataset can **capture the short-term patterns** observed in the real dataset.
- The TSTR method indicates that the synthetic data performs comparably well to the real data in the downstream ML task.
- However, the t-SNE visualisation indicates that there is potential for enhancing the model to increase the diversity of the synthetic series.

Conclusions

- DoppelGANger can effectively capture the short-term patterns observed in the real dataset but still struggles with long-term patterns.
- The model requires intensive training time.

Future Works

- Exploring unrolled GANs to assess if the generated series exhibit greater diversity.
- Experimenting with GAN models training using input data split by seasons, weekdays, and weekends to examine variations in diversity within the generated series.
- Conducting literature reviews on GAN solutions that facilitate learning long-term patterns in time series.