Evaluating Privacy-Utility Tradeoffs in Synthetic Smart Grid Data

Andre Catarino[1]*, Rui Melo[1,2], Luis Cruz[3], Rui Abreu[1],

[1]Faculty of Engineering, University of Porto
[2]Carnegie Mellon University
[3]TU Delft

Abstract

The widespread adoption of dynamic Time-of-Use (dToU) electricity tariffs requires accurately identifying households that would benefit from such pricing structures. However, the use of real consumption data poses serious privacy concerns, motivating the adoption of synthetic alternatives. In this study, we conduct a comparative evaluation of four synthetic data generation methods, Wasserstein–GP Generative Adversarial Networks (WGAN), Conditional Tabular GAN (CTGAN), Diffusion Models, and Gaussian noise augmentation, under different synthetic regimes. We assess classification utility, distribution fidelity, and privacy leakage. Our results show that architectural design plays a key role: diffusion models achieve the highest utility (macro-F1 up to 88.2%), while CTGAN provide the strongest resistance to reconstruction attacks. These findings highlight the potential of structured generative models for developing privacy-preserving, data-driven energy systems.

Introduction

The transition toward sustainable energy systems needs effective demand-side management strategies. One such approach is the adoption of dynamic electricity tariffs, including dToU tariffs, which incentivise consumers to shift electricity usage from peak to off-peak periods. This not only improves grid efficiency and facilitates renewable energy integration but also helps reduce carbon emissions. However, deploying such demand-side strategies relies heavily on detailed household consumption data, typically collected via smart meters, raising significant privacy concerns.

Synthetic data generation has emerged as a promising solution to balance the need for detailed data analytics with privacy preservation. In this context, generative models offer the ability to create realistic but artificial energy consumption data, enabling utility providers and researchers to conduct meaningful analyses without exposing sensitive user information. WGANs (Goodfellow et al. 2020), CTGAN (Mirza and Osindero 2014), Diffusion Models (Sohl-Dickstein et al. 2015), and Gaussian Noise Augmentation (Maalej and Rebai 2021a) represent a spectrum of approaches for synthetic data generation, each with distinct advantages for modelling temporal and structured data.

While synthetic data techniques aim to protect user privacy, they are not immune to privacy attacks. Membership Inference Attacks (MIA) (Shokri et al. 2017a) attempt to identify whether specific data points were part of a model's training set, while reconstruction attacks aim to recover sensitive information from model outputs. Therefore, rigorous evaluation of these vulnerabilities is essential for assessing the trade-off between data utility and privacy.

Our work extends the application of synthetic data augmentation from load forecasting to the prediction of household suitability for dynamic tariffs, an area less explored in current literature (Moon et al. 2020). We systematically benchmark multiple generative models with respect to their impact on classification accuracy and resilience to privacy attacks. To our knowledge, this is the first study to holistically evaluate the privacy-utility trade-offs of synthetic data in the context of tariff suitability rather than conventional load prediction. Our contributions aim to inform data-driven energy policy and promote privacy-aware research in smart grid environments.

To guide our investigation, we formulate the following:

- **RQ1:** Can machine learning models accurately classify households based on their suitability for dToU tariffs using behavioural features?
- **RQ2:** How do different synthetic data generation methods affect downstream classification utility under semi-synthetic and full-synthetic regimes?
- **RQ3:** To what extent do synthetic data generation methods reduce privacy leakage (via MIA and reconstruction) compared to real data, and how do privacy guarantees trade off with utility?

Related Work

Victor von Loessl (von Loessl 2023) analysed survey data from German households, revealing that only 23.6% of respondents exhibited low privacy concerns, primarily due to transparent data-handling practices. These practices help mitigate aversion to smart meter-enabled tariffs, particularly among consumers with heightened privacy concerns.

The effectiveness of dynamic tariffs in influencing household energy consumption depends heavily on tariff de-

*Corresponding author. Email: up202408593@up.pt

sign. Freier et al. (Freier and von Loessl 2022) proposed a methodology for structuring time-varying electricity prices to balance supply and demand while promoting renewable energy integration. Their findings indicated that short-term price variations play a crucial role in financial savings, yet households with limited flexibility struggle to achieve significant benefits. Similarly, researchers in (Guo and Weeks 2022) developed a two-stage game-theoretic model demonstrating that dynamic tariffs enhance market efficiency by aligning retail and wholesale prices. However, they highlight the need for regulatory interventions to ensure fair redistribution of economic gains among consumers.

Regarding privacy, concerns over smart meter data have driven the development of synthetic data generation techniques. Sheng Chai et al. (Chai and Chadney 2024) introduced Faraday, a Variational Autoencoder (VAE)-based model trained on 300 million UK smart meter readings. This model generates synthetic load profiles that closely resemble real energy consumption patterns while ensuring user privacy. Additionally, recent advances in generative models have further improved synthetic data quality for smart grid applications. Bilgi Yilmaz et al. (Yilmaz and Korn 2022) explored RCGAN, TimeGAN, CWGAN, and RCWGAN

In contrast, work by Asma Maalejand and Chiheb Rebai (Maalej and Rebai 2021a) introduced a Gaussian noise-based augmentation strategy for Support Vector Regression (SVR) models, demonstrating the utility of noise-based techniques to improve the performance of machine learning models in energy forecasting tasks.

Methodology

Our approach follows a structured pipeline designed to ensure accurate and privacy-preserving predictions of household suitability for dToU tariffs. The methodology consists of multiple stages, including data preprocessing, feature engineering, the experimental details of the synthetic data generation and model training, and the evaluation framework, as highlighted in Figure 1.

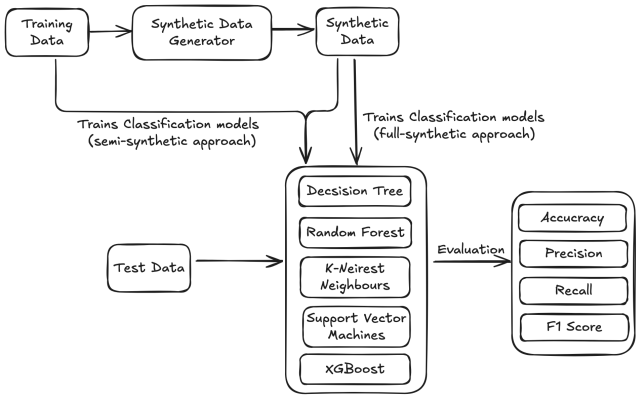

Figure 1: Proposed methodology overview.

Data Preprocessing

The dataset originates from the UK Power Networks' Low Carbon London project (2011–2014), comprising 167 million half-hourly electricity consumption records from 5,567 households in Greater London. It includes two groups: approximately 1,100 households enrolled in a dToU tariff trial in 2013, receiving real-time price signals, and around 4,500 households on a fixed-rate tariff serving as a control group. The dToU tariff featured variable pricing tiers, influencing household consumption patterns based on peak and off-peak rates. To enable meaningful comparisons across households with varying energy consumption levels, we standardised each consumption value.

Problem Formulation and Label Construction

Let each household $h_i \in \mathcal{H}$ be described by a vector of behavioural features $\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}] \in \mathbb{R}^d$, derived from smart meter readings under dynamic pricing conditions. These include metrics such as the ratio of energy consumed during high-tariff periods, load entropy, and changes in weekday/weekend usage. The goal is to learn a binary function $f : \mathbb{R}^d \to \{0, 1\}$, where $f(\mathbf{x}_i) = 1$ denotes that the household is *responsive* (i.e., likely to shift consumption in response to price signals), and $f(\mathbf{x}_i) = 0$ denotes otherwise.

However, the real data $\mathcal{D}_{\text{real}} = \{(\mathbf{x}_i, y_i)\}$ poses privacy risks due to its granularity. To mitigate this, we introduce a synthetic dataset $\mathcal{D}_{\text{syn}} = \{(\tilde{\mathbf{x}}_j, \tilde{y}_j)\}$, generated via models such as WGAN, Diffusion, and noise-based augmentations.

Since ground truth labels for responsiveness are unavailable, we adopt an unsupervised scoring method. First, all behavioural features are standardised using z-score normalisation. We then apply Principal Component Analysis (PCA) and extract the first principal component (PC1), which captures the dominant axis of variance across households. This yields a *responsiveness score* for each household:

$$s_i = \sum_{j=1}^{d} \mathbf{w_j} \cdot z_i^{(j)} \tag{1}$$

where $z_i^{(j)}$ is the standardized value of feature j for household i, and $\mathbf{w_j}$ is the loading of that feature in PC1. Households with higher s_i values exhibit behaviours more aligned with the dominant pattern of tariff responsiveness.

To obtain binary labels, we set a threshold q corresponding to the 75th percentile of s_i values across all households:

$$\text{Responsive}_i = \mathbb{I}[s_i > \text{Quantile}_q(s)] \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Households with scores above this threshold are labelled as responsive (1); others are labelled as non-responsive (0). In our experiments, we set $q = 0.75$, capturing the top 25% of households with the most favourable behavioural indicators.

While ground truth responsiveness labels are not available, our use of PCA over engineered behavioural features is consistent with prior work in behavioural segmentation and responsiveness inference in smart grid contexts. Similar dimensionality-reduction and clustering techniques have

been used to derive latent consumption patterns or segment user behaviour in demand response studies (Beckel et al. 2014; Albert and Rajagopal 2013). The resulting scores align well with expert intuition and tariff policy objectives, offering a practical proxy for real-world suitability.

Features included in the construction of the composite responsiveness score are the following:

- `high_usage_ratio` – fraction of total energy used during high-tariff periods; penalized in the score.
- `low_usage_ratio` – proportion of consumption during low-tariff periods; rewarded in the score.
- `peak_hour_ratio` – share of usage during peak daily hours (16:00–20:00); lowers the responsiveness score.
- `weekend_shift` – difference in average consumption between weekends and weekdays; large shifts reduce score due to behavioural inconsistency.
- `load_entropy` – entropy of usage distribution; moderate entropy indicates more regular, responsive behaviour.
- `load_factor_low` – efficiency of consumption during low-tariff periods; higher values are rewarded.

Households that exhibit reduced electricity usage during high-tariff periods and demonstrate more regular load patterns, such as a higher load factor during low-tariff periods, receive higher responsiveness weights. Positive PC1 loadings (e.g., `low_usage_ratio`, `load_factor_low`) contribute positively to the responsiveness score, indicating alignment with dToU incentives. In contrast, negative loadings (e.g., `high_usage_ratio`, `peak_hour_ratio`) reflect behaviours that are less responsive to such incentives.

Baseline Classifiers and Evaluation Protocol We benchmark five standard tabular classifiers: Decision Tree (DT), Random Forest (RF), k-Nearest Neighbours (KNN), Support-Vector Machine with an *RBF* kernel (SVM), and Extreme Gradient Boosting (XGBOOST), using a fixed nested-cross-validation pipeline. Each estimator is wrapped in a STANDARDSCALER and tuned by **randomised search** (`n_iter`$=10$) on an inner *3-fold* stratified CV. Performance is assessed on an outer **5-fold** stratified CV. The hyperparameter search space is listed in Table 1.

Table 1: Randomised search spaces for the baseline classifiers. "a:b:c" follows `range(a, c+1, b)`.

Model	Parameter grid
RF	`n_estimators` $\in\{100, 300, 500\}$; `max_depth` $\in\{$None, 10, 20$\}$; `min_samples_split` $\in\{2, 5\}$
KNN	`n_neighbors` = 3:2:15; `weights` $\in\{$uniform, distance$\}$
SVM	`C` $\in\{1, 10\}$; `kernel` = `rbf`; `gamma` = `scale`
XGB	`n_estimators` $\in\{100, 300, 500\}$; `max_depth` $\in\{3, 6, 10\}$; `learning_rate` $\in\{0.01, 0.1, 0.3\}$; `subsample` $\in\{0.8, 1.0\}$
DT	`max_depth` $\in\{$None, 10, 20, 30$\}$; `criterion` $\in\{$gini, entropy$\}$

Our primary utility measure is the **macro-F1 score**:

$$\mathrm{F1}_{\text{macro}} \;=\; \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c+\mathrm{Recall}_c} \tag{3}$$

with $C = 2$ classes. Macro-F1 weights all classes equally, making it the appropriate choice for our moderately imbalanced dataset (class ratio 3:1). For each classifier, we report $\mu_{\mathrm{F1}} \pm \sigma$ and the 95% confidence interval across the outer folds. The outer splits are *identical* for real, semi-synthetic, and full-synthetic datasets, enabling paired significance testing via the Wilcoxon signed-rank and paired t-test.

Generative Data Synthesis Approaches

Notation. We denote the numerical feature matrix by $X \in \mathbb{R}^{n\times d}$ and the binary target by $y \in \{0,1\}^n$. A generator produces $(\tilde{X}, \tilde{y})$ of the same dimensionality.

(1) Wasserstein–GP GAN Our WGAN comprises a generator G and a critic C (*three-layer* MLPs, hidden 128, RELU). Spectral norm imposes the 1-Lipschitz constraint on C, and we train with the Wasserstein loss plus gradient penalty ($\lambda_{\mathrm{gp}} = 10$) (Gulrajani et al. 2017). To avoid label-mode collapse, we add three terms: (i) entropy maximisation that regularises G to encourage sample diversity (Khoramshahi et al. 2020); (ii) a label-balance term $\big(\mathrm{E}_{\tilde{y}}[\tilde{y}]-p_r\big)^2$; and (iii) an MSE penalty to the empirical class ratio p_r.

We run five critical updates per generator update with Adam optimiser (Kingma and Ba 2017) ($\alpha = 10^{-4},\ \beta_1 = 0,\ \beta_2 = 0.9$), batch size 32 for 100 epochs. Preliminary ablation experiments revealed that removing the entropy and class-balance regularizers led to severe mode collapse, particularly in the label distribution. In early trials, the generator defaulted to producing only the majority class. These regularizers were therefore retained in all WGAN experiments to ensure balanced class synthesis and stable training.

(2) DDPM-Based Diffusion Model Our diffusion-based generator is inspired by the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) framework and adapted for the tabular classification setting. The model employs a three-layer MLP backbone (hidden size 128) with ReLU activations, and incorporates time-step conditioning via a learned embedding layer ($t \in \{0,\dots,T-1\}$, with T=100 steps). The architecture jointly predicts two outputs: (i) the additive Gaussian noise $\varepsilon_\theta(x_t, t)$ and (ii) the binary class logit $\hat{y}$, using a dual-head design.

We train the model using a composite objective:

$$\mathcal{L} = \mathrm{MSE}(\varepsilon, \hat{\varepsilon}) + \mathrm{BCE}(y, \hat{y}) \tag{4}$$

balancing both the denoising accuracy and classification capability. The optimiser is Adam (α=10^{-3}, batch size 32), with a training horizon of 50 epochs.

To improve generation stability, we maintain an Exponential Moving Average (EMA) of model weights (decay 0.999), applied during inference. Sampling follows a deterministic DDIM-style reverse process. Starting from Gaussian noise $x_T \sim \mathcal{N}(0, I)$, the model refines the sample through time-stepped denoising conditioned on the learned time embeddings. Generated features are decoded through

the learned denoising path, and class labels are derived from the class logit head using a thresholded sigmoid output.

This formulation allows for joint synthetic feature-label generation in a time-consistent, noise-aware manner, while incorporating practical design elements like EMA smoothing and dual-task training to ensure stable and realistic synthetic data production.

(3) CTGAN We adopt the SDV `CTGANSynthesizer` implementation of CTGAN (Xu et al. 2019a), designed to handle the unique challenges of mixed-type tabular data. CTGAN models both continuous and categorical features by conditioning the generator and discriminator on sampled values of discrete columns, allowing for realistic class-conditional generation.

We specify metadata using the `SingleTableMetadata` interface, registering all continuous behavioural features with "numerical" type and the binary `TargetClass` as "categorical". This ensures schema compliance and enables CTGAN to embed and conditionally sample from discrete feature spaces.

The model is trained with the following hyperparameters: batch size of 500, learning rate of 2×10^{-4}, and 300 training epochs. The training objective follows the WGAN-GP formulation to stabilise convergence. During each iteration, CTGAN samples a conditional vector from the discrete columns, embeds categorical features, and concatenates them with Gaussian noise to form the generator input. The discriminator receives both real and synthetic data with their associated conditions, ensuring the generator learns contextually valid feature-label combinations.

Gaussian Noise Injection for Data Augmentation In this approach, Gaussian noise $\mathcal{N}(\mu, \sigma^2)$ is added to the numerical features during training, following the methodology outlined in (Maalej and Rebai 2021b). This noise injection serves to regularise training, preventing the model from memorising training patterns and improving its generalisation ability.

Dynamic Noise Adjustment To ensure effective augmentation, noise parameters are dynamically tuned based on feature distributions. For each feature x_i, the noise follows:

$$x_i' = x_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_{x_i}^2) \tag{5}$$

where σ_{x_i} is adaptively set as a fraction of the empirical standard deviation of x_i.

Evaluation and Validation

Experimental Protocol

The effectiveness of synthetic data augmentation, generated using WGAN, CTGAN, a diffusion model, and Gaussian noise augmentation, is assessed through two approaches: **semi-synthetic** and **full-synthetic**. The *semi-synthetic* approach augments the original dataset with synthetic observations, while the *full-synthetic* approach replaces real data entirely with synthetic samples. The impact of these approaches is evaluated based on the following criteria:

Distribution Fidelity: The similarity between synthetic and real data distributions is measured using Kullback-Leibler (KL) (Kullback and Leibler 1951) and Jensen-Shannon (JS) (Menéndez et al. 1997) to assess how well synthetic data preserves the statistical properties of the original dataset.

Utility: Classification models trained on semi-synthetic and full-synthetic datasets are compared against those trained on real data alone. Performance metrics are analysed to determine whether data augmentation enhances generalisation.

Privacy Robustness: The degree to which synthetic data protects privacy is assessed using attack simulations, including MIA and Reconstruction Attacks.

Fidelity

Visual inspection. Figure 2 overlays the t-SNE embeddings of the original records and eight synthetic variants. Large overlaps (e.g., Gaussian noise augmentation vs Original) indicate that synthetic points populate the same manifold regions as the real data, whereas visible cluster gaps reveal distributional drift (e.g., WGAN synthetic samples).

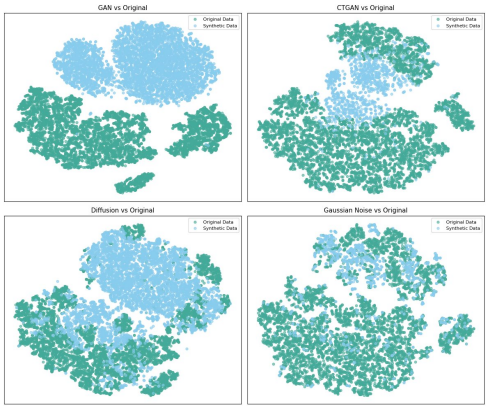

Figure 2: t-SNE (perplexity 30) of real vs. synthetic samples. Colours denote generator + synthesis regime.

Table 2: Fidelity and utility metrics across synthetic data types. Best KL/JS in bold, worst underlined.

Synthetic data	KL	JS	Best Classifier	Macro-F1 (%)
Real	—	—	XGB	67.5 ± 3.8
WGAN Semi-synthetic	2.68	0.083	XGB	63.4 ± 5.3
WGAN Full-synthetic	6.12	0.221	KNN	50.7 ± 4.2
CTGAN Semi-synth.	**0.16**	**0.014**	XGB	68.3 ± 5.4
CTGAN Full-synth.	0.44	0.045	RF	82.5 ± 4.3
Diffusion Semi-synth.	0.74	0.021	RF	73.0 ± 5.1
Diffusion Full-synth.	1.44	0.054	SVM	88.2 ± 3.0
Noise Baseline Semi	0.27	0.007	KNN	75.7 ± 5.1
Noise Baseline Full	0.52	0.016	KNN	81.6 ± 2.9

Divergence metrics. To move beyond visual impression, we compute the Kullback–Leibler (KL) and Jensen–Shannon (JS) divergences between the multivariate Gaussian

Table 3: Macro-F1 (%) on each dataset. ↑ / ↓: significant improvement / drop vs. real data ($p < 0.05$).

Dataset	RF	KNN	SVM	DT	XGB
Real (baseline)	66.5 ± 4.1	62.1 ± 3.2	65.0 ± 4.0	62.0 ± 3.0	**67.5** ± 3.8
WGAN Semi-Synth.	$61.0 \pm 5.3\downarrow$	$55.3 \pm 5.\downarrow$	61.6 ± 4.1	60.0 ± 5.5	**63.4** ± 5.3
WGAN Full-Synth.	$42.6 \pm 2.6\downarrow$	**50.7** $\pm 4.2\downarrow$	$50.5 \pm 3.5\downarrow$	$48.6 \pm 3.4\downarrow$	$45.7 \pm 2.9\downarrow$
CTGAN Semi-Synth.	67.6 ± 4.2	61.1 ± 4.9	67.5 ± 4.4	62.3 ± 4.8	**68.3** ± 5.4
CTGAN Full-Synth.	**82.5** $\pm 4.3\uparrow$	$75.1 \pm 3.4\uparrow$	$80.1 \pm 4.2\uparrow$	$70.2 \pm 3.6\uparrow$	$80.0 \pm 3.1\uparrow$
Diffusion Semi-Synth.	73.0 ± 5.1	62.5 ± 3.4	$71.5 \pm 3.6\downarrow$	$69.9 \pm 3.8\downarrow$	$72.0 \pm 4.$
Diffusion Full-Synth.	$82.9 \pm 3.1\uparrow$	$68.8 \pm 3.0\uparrow$	**88.2** $\pm 3.0\uparrow$	$79.2 \pm 4.0\uparrow$	$87.4 \pm 4.0\uparrow$
Noise Semi-Synth.	$74.8 \pm 5.9\uparrow$	$75.7 \pm 5.1\uparrow$	$73.6 \pm 4.6\uparrow$	$69.3 \pm 2.7\uparrow$	$75.5 \pm 5.8\uparrow$
Noise Full-Synth.	$76.0 \pm 3.7\uparrow$	**81.6** $\pm 2.9\uparrow$	$77.2 \pm 4.4\uparrow$	$72.6 \pm 3.0\uparrow$	$79.2 \pm 3.0\uparrow$

kernel-density estimates of the real data and each synthetic distribution (Table 2).[1]

Both metrics reward exact overlap, but JS is symmetric and bounded, making it easier to compare across methods.

Observations. CTGAN Semi-synthetic achieves the lowest divergence on both metrics, indicating that a modest proportion of real samples is sufficient for that generator to learn high-order structure. Diffusion models come second, followed closely by the calibrated Gaussian-noise injection. WGAN Full-synthetic is an outlier with a KL of 6.1, evidence of model collapse and excessive variance inflation.

Summary-statistics parity. Beyond divergence metrics, we assess fidelity by comparing means, skewness, and kurtosis of 24 numerical features. **CTGAN Semi-Synthetic** best replicates original distributions, preserving asymmetry and tail behavior in complex features like `Peak_to_Mean_Ratio`. In contrast, **WGAN Full-Synthetic** flattens key statistics. Diffusion and noise-based models preserve central moments but tend to exaggerate tails.

Overall fidelity score. Combining (i) low KL/JS, (ii) moment parity, and (iii) the qualitative t-SNE overlap, we conclude that **CTGAN Semi-synthetic delivers the highest distributional fidelity**, closely followed by Diffusion Semi-synthetic. WGAN full-synthetic approach, while visually plausible, deviates substantially in tail behaviour and is therefore *not recommended* for downstream tasks that rely on accurate peak or burst modelling.

Utility

Performance Comparison Table 3 summarises the classification performance across the different augmentation strategies, highlighting that diffusion-based full-synthetic data yields the highest Macro-F1 score.

Significance Testing To compare synthetic against real data, we apply two paired tests across the outer-fold macro-F1 vectors: (i) two-sided Wilcoxon signed-rank (Wilcoxon 1945), and (ii) paired Student t-test (Student 1908). The null hypothesis states that the mean performance difference is zero. $p < 0.05$ (after Holm–Bonferroni correction across five classifiers) flags a *significant utility drop*. (See Tables 2 and 3). To obtain an unbiased estimate of downstream utility, we adopt a fixed *nested* cross-validation pipeline.

Privacy Robustness

Protecting individual records from disclosure is a prerequisite when synthetic data are released or used to train downstream models. We study the canonical *membership inference attack* (MIA) under the strong **posterior-only black-box** threat model: The adversary observes the prediction vector of the target model for an arbitrary query but has no access to gradients, weights, or training loss.

MIA

- **Shadow ensemble.** For each dataset we fit $n_{\text{shadow}} = 5$ Shadow models with heterogeneous architectures (Random Forest and 2-layer MLP). Hyper-parameters and seeds follow the public YAML file in our code release.
- **Attack features.** From every shadow inference we extract the maximum posterior probability, a standard and practically obtainable signal (Shokri et al. 2017b).
- **Attack classifiers.** We train both a Random Forest (200 trees) and an MLP (32–32 units) on the aggregated shadow-generated attack dataset, using a 10% hold-out portion for early stopping. Only the *stronger* attacker per setting is reported.
- **Repeats and confidence intervals.** The entire pipeline is repeated for five independent seeds. We report the mean attack AUC together with the *two-sided 95% t-interval*.

Discussion. Across all eight synthetic configurations illustrated in Table 4, the attack AUC stays in the narrow range 0.61–0.64, only slightly above random guessing. Differences between *semi-* and *full-synthetic* variants of the same generator never exceed $\Delta\text{AUC} = 0.03$. The Gaussian-noise approach offers the strongest privacy (0.61)). In contrast, diffusion-based full synthesis yields the best trade-off: a modest AUC of 0.64 while increasing the target's accuracy.

In our evaluation, MIAs are largely ineffective against these synthetic data pipelines. The reported attack AUCs suggest that adversaries cannot meaningfully distinguish between training and non-training samples. Reconstruction analysis offers a more informative lens under these conditions.

[1]Following Xu *et al.* (Xu et al. 2019b), we estimate the probability density of each dataset with a Gaussian KDE whose bandwidth is chosen by Scott's rule (Scott 1992), i.e. $h_j = \sigma_j n^{-1/(d+4)}$. For our real data ($n = 1117$, $d = 24$) this yields $h_j \approx 0.78\,\sigma_j$ for every feature.

Table 4: MIA results (higher AUC = more vulnerable). Parentheses show 95% confidence interval.

Synthetic data	Shadow MIA AUC
WGAN Semi-synthetic	0.62 (0.60, 0.63)
WGAN Full-synthetic	0.62 (0.61, 0.64)
CTGAN Semi-synthetic	0.63 (0.61, 0.64)
CTGAN Full-synthetic	0.64 (0.62, 0.65)
Diffusion Semi-synthetic	0.62 (0.61, 0.63)
Diffusion Full-synthetic	0.64 (0.62, 0.65)
Noise Semi-synthetic	0.63 (0.61, 0.64)
Noise Full-synthetic	0.61 (0.60, 0.63)
Original data	**0.65 (0.64, 0.66)**

Reconstruction attack

Threat model. An adversary receives the released *synthetic table*. Using only the published attributes, it trains a regression model that tries to *reconstruct* a hidden target feature of interest, *average consumption*. Successful reconstruction reveals fine-grained private behaviour.

Attack procedure.

(1) **Model sweep.** Train five regressors {Random Forest, Gradient-Boosting, Ridge, Lasso, MLP} on the synthetic data. Pick the model with the lowest mean-squared error (MSE) on a held-out *real* test set. This "best of sweep" yields an upper bound on what a resourceful adversary could achieve.

(2) **Real-data upper bound.** Train the same Random Forest directly on the real table; its error is the maximum leak possible if the raw data were published.

(3) **Privacy metrics.** *Privacy gap* $\Delta = \mathrm{MSE}_{\mathrm{noise}} - \mathrm{MSE}_{\mathrm{syn}}$ and *privacy-risk score* $\mathrm{PRS} = \Delta/(\mathrm{MSE}_{\mathrm{noise}} - \mathrm{MSE}_{\mathrm{real}})$ (ratio in $[0, 1]$; higher $\Rightarrow$ bigger leak).

From table 5 we extract the following findings:

Table 5: Reconstruction results (lower MSE / PRS = safer). Best value per column in **bold**, worst underlined. "SS" stands for Semi-synthetic and "FS" for Full-synthetic.

Synthetic data	Best model	MSE ($\times 10^{-3}$)	ρ	R^2	PRS
WGAN SS	RF	0.18	0.996	0.990	0.99
WGAN FS	GB	8.30	0.760	0.561	0.62
CTGAN SS	RF	0.27	0.994	0.986	0.99
CTGAN FS	**Lasso**	**19.1**	**- 0.022**	**–0.010**	**0.16**
Diffusion SS	GB	0.07	0.998	0.996	0.99
Diffusion FS	Ridge	0.35	0.994	0.982	0.98
Noise SS	GB	0.00	1.000	0 999	1. 0
Noise FS	GB	0.05	0.999	0 997	1. 0

1. **Semi-synthetic sets leak almost as much as the real data.** Their errors are three orders of magnitude smaller than the noise baseline ($\mathrm{MSE} \leq 0.27 \times 10^{-3}$) and correlations exceed 0.99, giving $\mathrm{PRS} \approx 1$.
2. **CTGAN Full-synthetic offers the strongest protection.** Error jumps two orders of magnitude ($\mathrm{MSE} = 19.1 \times 10^{-3}$), correlation vanishes, and PRS drops to 0.16—*84 % less leak* than raw data. Diffusion Full-synthetic keeps good fidelity ($\mathrm{MSE} = 0.35$) but still leaks almost as much as noise-shuffled data (PRS 0.98).
3. **WGAN Full-synthetic** partially mitigates risk (PRS 0.62) but still leaves a strong linear correlation between synthetic and real data ($\rho = 0.76$).

Observations. When the secret attribute is a fine-grained load profile, **CTGAN full-synthetic is the only generator that meaningfully degrades reconstruction attacks** without resorting to external noise addition. Combining these results with the MIA study, we recommend CTGAN full synthesis for any public release that prioritises privacy; diffusion full synthesis remains acceptable for internal analytics where some residual leakage is tolerable.

Privacy–Utility Pareto Figure 3 presents the Privacy–Utility Pareto frontier, mapping each strategy along two axes: privacy, quantified by the Privacy Risk Score, and utility, measured by macro-F1 classification performance.

The figure highlights clear trade-offs, where structured generative models, especially CTGAN and Diffusion, define the Pareto frontier, confirming CTGAN full synthesis offers the strongest defence, while diffusion provides the best compromise between privacy and predictive utility.

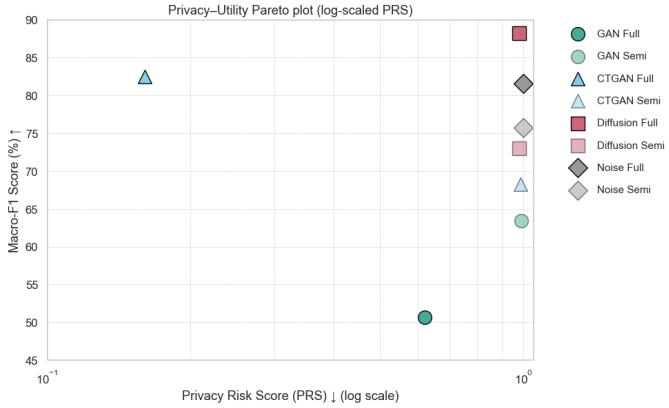

Figure 3: Privacy–Utility Pareto plot. Marker shape encodes model type (GAN: circle, CTGAN: triangle, Diffusion: square, Noise: diamond). Color encodes data type (Full: darker; Semi: lighter).

Discussion

This section synthesizes results to answer our three research questions (RQ1–RQ3) concerning model effectiveness, synthetic data utility, and privacy protection.

Model Effectiveness

Classifier performance varies considerably across synthetic data strategies. **SVM and XGBoost achieve the highest macro-F1 scores overall (up to 88.2%) on Diffusion Full-Synthetic data**, confirming that structured synthetic data can support accurate tariff responsiveness classification. However, no single model is optimal across all

datasets: **Random Forest** performs best on CTGAN Full-Synthetic (82.5%), while **KNN** performs best on the Noise Baseline (81.6%). Decision Trees consistently underperform, showing lower stability and greater overfitting sensitivity. These findings align with prior smart-grid classification studies (Petrlik et al. 2022; Mhaske et al. 2022) but extend them to the tariff suitability domain.

Impact of Synthetic Data on Performance

The utility of classifiers trained on synthetic data reveals clear distinctions between generation methods. **Diffusion Full-Synthetic data consistently achieves the highest utility**, outperforming the real-data baseline by over 20 macro-F1 points in some cases. **CTGAN Full-Synthetic** also performs strongly (82.5%, RF), surpassing its semi-synthetic counterpart. This supports prior findings that hybrid or structured generative models enhance performance in energy analytics (Liang, Wang, and Wang 2024), but our results demonstrate this effect in classification rather than forecasting. Moreover, while Gaussian noise augmentation can improve forecasting (Maalej and Rebai 2021a), diffusion models deliver greater performance gains for classification, suggesting that architecture plays a decisive role in utility.

Privacy Considerations

Full-synthetic datasets provide stronger defences against privacy attacks than real or semi-synthetic data. Across all generators, MIA's AUCs remain near random (0.61–0.64), replicating the limited leakage observed for diffusion models (Wu et al. 2025). In reconstruction attacks, however, privacy diverges sharply: **CTGAN Full-Synthetic** achieves the lowest privacy-risk score (PRS = 0.16), confirming robustness against attribute recovery (Alshantti, Rasheed, and Westad 2024). In contrast, WGAN-GP variants leak more information, consistent with prior empirical rankings (Hyeong et al. 2022).

Conclusion

This study evaluated how different synthetic data generation architectures influence the trade-off between utility and privacy in household classification for dToU tariffs. We benchmarked four generation strategies—WGAN, CTGAN, Diffusion Models, and Gaussian noise augmentation—under both full- and semi-synthetic regimes. Utility was measured using macro-F1 score, fidelity through KL/JS divergence, and privacy via MIA and feature reconstruction attacks.

Our findings, illustrated in Figure 3, empirically confirm the hypothesis that **architectural design of generative models strongly shapes both privacy and utility**. Full-synthetic data based on diffusion, provided the highest predictive utility. CTGAN full-synthetic data achieved the strongest privacy preservation, reaching the lowest privacy-risk score (PRS = 0.16) while maintaining competitive accuracy. In contrast, WGAN full-synthetic data performed weakest on both fronts, reflecting instability and mode collapse. Gaussian noise augmentation yielded deceptively high utility but closely mirrored the real distribution (low KL/JS), resulting in greater reconstruction leakage. Overall, the findings highlight that structured architectures (e.g., Diffusion and CTGAN) are better suited for privacy-preserving synthetic data generation in smart grid applications.

Threats to Validity

Label construction bias. The ground-truth labels for responsiveness are derived from unsupervised dimensionality reduction and quantile-based thresholding. While this procedure is grounded in behavioural features, it may not align perfectly with economic responsiveness or actual behavioural change. The absence of external validation data, such as expert annotations or outcomes from real-world dToU interventions, limits our ability to confirm the semantic validity of these labels. **Contextual scope limitations.** Our analysis is grounded in a single empirical dataset from the UK Power Networks trial. While this dataset is representative of real smart meter deployments and tariff experiments, its demographic, regional, and temporal scope may not generalize to other electricity markets or consumption behaviours. **Limitations of empirical privacy evaluation.** Our privacy assessment relies on empirical adversaries, membership inference and reconstruction attacks, without incorporating formal guarantees. These attacks are practical and widely used but may underestimate leakage against more adaptive or theoretically grounded threats.

Future Work

This research demonstrates the feasibility of synthetic data for predicting household responsiveness to dToU tariffs, yet several avenues remain to advance generative modeling. A first avenue is integrating **time-aware generative architectures** that capture temporal dependencies and consumption dynamics. For example, TimeGAN (Yoon, Jarrett, and Schaar 2019) combines recurrent models with adversarial training to preserve temporal order and feature dependencies. More recent diffusion-based time-series models, such as TiDE (Gupta 2023) and CSDI (Tashiro, Song, and Ermon 2021), leverage transformers and conditional score-based imputation to improve temporal fidelity and controllability. In parallel, embedding explicit **causal structure** into the generative process could enable principled simulations of consumption responses under varying tariff schemes.

A second direction concerns the **construction of responsiveness labels**. The current approach relies on unsupervised PCA-based thresholds, which are consistent but indirect proxies for responsiveness. Future work should analyse alternative thresholding strategies and triangulate results with expert annotation, user studies, or empirical evidence from tariff trials (Faruqui and Sergici 2010; von Loessl 2023).

Finally, achieving **formal privacy guarantees** remains an open challenge. Beyond empirical evaluation via membership inference and reconstruction attacks, future work should integrate certified mechanisms such as Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al. 2016) or the Private Aggregation of Teacher Ensembles (PATE) (Papernot et al. 2017) to establish provable bounds on privacy leakage and move toward verifiable privacy–utility trade-offs.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, 308–318. New York, NY, USA: Association for Computing Machinery. ISBN 9781450341394.

Albert, A.; and Rajagopal, R. 2013. Smart Meter Driven Segmentation: What Your Consumption Says About You. *IEEE Transactions on Power Systems*, 28(4): 4019–4030.

Alshantti, A.; Rasheed, A.; and Westad, F. 2024. Privacy Reidentification Attacks on Tabular GANs. arXiv:2404.00696.

Beckel, C.; Sadamori, L.; Staake, T.; and Santini, S. 2014. Revealing household characteristics from smart meter data. *Energy*, 78: 397–410.

Chai, S.; and Chadney, G. 2024. Faraday: Synthetic Smart Meter Generator for the smart grid. arXiv:2404.04314.

Faruqui, A.; and Sergici, S. 2010. Household response to dynamic pricing of electricity: a survey of 15 experiments. *Journal of Regulatory Economics*, 38(2): 193–225.

Freier, J.; and von Loessl, V. 2022. Dynamic electricity tariffs: Designing reasonable pricing schemes for private households. *Energy Economics*, 112: 106146.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved Training of Wasserstein GANs. arXiv:1704.00028.

Guo, B.; and Weeks, M. 2022. Dynamic tariffs, demand response, and regulation in retail electricity markets. *Energy Economics*, 106: 105774.

Gupta, R. e. a. 2023. TiDE: Time-series Diffusion for Forecasting and Imputation. *arXiv preprint arXiv:2305.13395*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Hyeong, J.; Kim, J.; Park, N.; and Jajodia, S. 2022. An Empirical Study on the Membership Inference Attack against Tabular Data Synthesis Models. arXiv:2208.08114.

Khorramshahi, P.; Souri, H.; Chellappa, R.; and Feizi, S. 2020. GANs with Variational Entropy Regularizers: Applications in Mitigating the Mode-Collapse Issue. arXiv:2009.11921.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kullback, S.; and Leibler, R. A. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1): 79–86.

Liang, X.; Wang, Z.; and Wang, H. 2024. Synthetic Data Generation for Residential Load Patterns via Recurrent GAN and Ensemble Method. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–12.

Maalej, A.; and Rebai, C. 2021a. Sensor Data Augmentation Strategy for Load Forecasting in Smart Grid Context. In *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*, 979–983.

Maalej, A.; and Rebai, C. 2021b. Sensor Data Augmentation Strategy for Load Forecasting in Smart Grid Context. 979–983.

Menéndez, M. L.; Pardo, J.; Pardo, L.; and Pardo, M. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2): 307–318.

Mhaske, D.; Satam, R.; Londhe, S.; and et al. 2022. An Efficient Electricity Theft Detection Using XGBoost. *International Journal of Engineering Applied Sciences and Technology*, 6(10): 282–287.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784.

Moon, J.; Jung, S.; Park, S.; and Hwang, E. 2020. Conditional Tabular GAN-Based Two-Stage Data Generation Scheme for Short-Term Load Forecasting. *IEEE Access*, 8: 205327–205339.

Papernot, N.; Abadi, M.; Úlfar Erlingsson; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. arXiv:1610.05755.

Petrlik, I.; Lezama, P.; Rodriguez, C.; and et al. 2022. Electricity Theft Detection using Machine Learning. *International Journal of Advanced Computer Science and Applications*, 13(12): 420–428.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017a. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017b. Membership Inference Attacks against Machine Learning Models. arXiv:1610.05820.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1): 1–25.

Tashiro, Y.; Song, J.; and Ermon, S. 2021. CSDI: Conditional Score-based Diffusion Models for Imputation. *NeurIPS*.

von Loessl, V. 2023. Smart meter-related data privacy concerns and dynamic electricity tariffs: Evidence from a stated choice experiment. *Energy Policy*, 180: 113645.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80–83.

Wu, X.; Pang, Y.; Liu, T.; and Wu, S. 2025. Winning the MIDST Challenge: New Membership Inference Attacks on Diffusion Models for Tabular Data Synthesis. *arXiv preprint arXiv:2503.12008*.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019a. Modeling Tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems*.

Xu, L.; Skoularidou, M.; Cueto, A.; and González, J. 2019b. Modeling Tabular Data Using Conditional GAN. In *Advances in Neural Information Processing Systems 32*, 7335–7345.

Yilmaz, B.; and Korn, R. 2022. Synthetic demand data generation for individual electricity consumers : Generative Adversarial Networks (GANs). *Energy and AI*, 9: 100161.

Yoon, J.; Jarrett, D.; and Schaar, M. 2019. Time-series Generative Adversarial Networks.