*Grenze International Journal of Engineering and Technology, June Issue*

# Trustworthy Synthetic Financial Data Generation using CTGAN with Explainable AI: A Novel Trust Score Framework for Credit Ris5k Assessment

Avani Dongare[1], Soham Pujari[2] and Anil V Turukmane[3]
[1-3]VIT-AP / School of Computer Science and Engineering (SCOPE), Amaravati, India
Email: avanidongare22@gmail.com, sppujari07@gmail.com, anil.turukmane@vitap.ac.in

***Abstract*— This paper presents a comprehensive framework for generating trustworthy synthetic financial data by combining Conditional Tabular GAN (CTGAN) with Explainable AI (XAI) techniques. We propose a novel multi-dimensional trust score integrating five key metrics: downstream utility (WU = 0.9776), SHAP feature importance alignment (WI = 0.8810), permutation importance (WP = 0.9241), LIME local explanation fidelity (WL = 0.8646), and correlation-structure fidelity (WC = 0.7572), achieving a composite score of 0.8954. The framework validates synthetic data quality across statistical fidelity, predictive utility, and explainability dimensions. Comparative analysis demonstrates CTGAN superiority over VAE with 35.4% higher feature interaction preservation. The framework enables practical deployment in regulated financial domains while maintaining privacy, interpretability, and regulatory compliance. Results show 97.76% downstream utility preservation and 1.46% AUC improvement when combining real and synthetic data for class imbalance mitigation.**

***Index Terms*— Synthetic Data Generation, CTGAN, Credit Risk, Explainable AI, Trust Metrics, SHAP, LIME, Financial Machine Learning.**

## I. INTRODUCTION

### *A. Problem Statement*

The increasing adoption of machine learning models for credit risk assessment has intensified the demand for large-scale, high-quality financial datasets. However, the availability and use of real-world financial data are severely constrained by regulatory frameworks such as GDPR, CCPA, and emerging AI governance laws, which restrict data sharing, reuse, and cross-institutional collaboration. As a result, financial institutions face a fundamental tension between model performance requirements and data privacy obligations. Synthetic data generation has emerged as a promising solution to this challenge by enabling the creation of artificial datasets that mimic real data distributions while eliminating direct privacy risks. Generative models such as GANs and VAEs have demonstrated success in preserving statistical properties of tabular data. However, existing validation practices focus predominantly on statistical similarity and predictive performance, neglecting a critical requirement in regulated domains: trustworthiness and explainability of downstream decisions. In credit risk modeling, regulatory compliance mandates that model decisions be interpretable, justifiable, and auditable at both global and individual levels.

*Grenze ID: 01.GIJET.12.2.523*

While prior work evaluates synthetic data using metrics such as marginal distributions, correlation fidelity, or downstream AUC, these metrics alone are insufficient to guarantee that models trained on synthetic data learn the same decision logic as models trained on real data. In particular, there is no established framework to quantify whether: Feature importance rankings are preserved, Feature interactions remain meaningful, Local explanations for individual credit decisions remain consistent.

### *B. Motivation and Research Gap*

Despite growing interest in synthetic data for financial applications, current evaluation methodologies remain misaligned with regulatory expectations. Existing studies typically validate synthetic datasets using statistical distance measures or downstream model accuracy, implicitly assuming that similar performance implies similar decision behavior. However, in regulated credit decision-making, performance equivalence does not guarantee interpretability equivalence. From a regulatory perspective, financial institutions must demonstrate: Consistency of feature influence (e.g., income, credit score, employment status), Stability of feature interactions, Faithful explanations for individual credit decisions. To the best of our knowledge, no prior work provides a unified, quantitative framework that measures trustworthiness of synthetic financial data across explainability, feature interaction preservation, and downstream utility simultaneously. This lack of formal trust metrics constitutes a major barrier to real-world adoption of synthetic data in regulated environments. This paper addresses this gap by introducing a trust-centric validation framework that integrates explainable AI techniques into the evaluation of synthetic data, transforming explainability from a post-hoc diagnostic tool into a core validation signal.

### *C. Limitations of Existing Synthetic Data Validation Approaches*

While synthetic data offers practical benefits, adoption in regulated financial institutions faces three unresolved concerns. First, does synthetic data preserve the interpretability properties of real data? Models trained on synthetic data must explain credit decisions in ways regulators find defensible. Second, are learned feature relationships meaningful or artifacts of the generative process? Synthetic data that inadvertently creates spurious correlations would mislead credit decisions and violate fair lending principles. Third, what quantitative evidence demonstrates synthetic data suitability for regulated decision-making?

Regulators increasingly demand explicit trustworthiness metrics rather than informal assurances. Traditional synthetic data validation approaches assess statistical fidelity (marginal distributions, correlation preservation) but ignore interpretability preservation and feature relationship validity. This creates a critical gap preventing synthetic data adoption in regulated domains where explainability is non-negotiable.

## II. Methodology: CTGAN with Explainable AI Integration

### *A. Conditional Tabular GAN Architecture*

Conditional Tabular GAN (CTGAN) represents a significant advancement in generative modeling for tabular data, specifically designed to overcome limitations of standard GAN architectures when applied to financial datasets. Traditional GANs struggle fundamentally with mixed data types (continuous, categorical, ordinal) prevalent in financial applications, where continuous features exhibit arbitrary scales and non-Gaussian distributions, and categorical features contain high-cardinality values with severe imbalance.

CTGAN introduces mode-specific categorical handling to ensure balanced representation of minority classes:

$$\mathbf{c} = \text{OneHot}(\text{Sample}(p_{cat})) \oplus \text{Gumbel}(\text{Continuous}) \quad (1)$$

This mechanism samples from empirical categorical distributions, ensuring rare categories (defaults occurring in 3% of cases) appear with realistic frequency in synthetic data. The Gumbel-Softmax trick enables smooth gradients through categorical sampling, facilitating effective backpropagation during training. This design is crucial for credit risk where the minority class (defaults) contains the most valuable information.

Financial data exhibits non-standard statistical properties violating assumptions of standard normalization. CTGAN addresses this through Variational Gaussian Mixture (VGM) normalization, applying mode-specific normalizers:

$$\hat{x}^{(j)} = \text{VGM}(x_i | \text{feature}_j = \text{mode}_k) \quad (2)$$

For each categorical feature with distinct values, VGM maintains separate normalizers for continuous variables. This recognizes that income distributions vary significantly by employment status or credit tier. By normaliz- ing

within modes, CTGAN preserves segment-specific patterns, improving fidelity for downstream credit risk modeling.

*B. CatBoost for Credit Risk Prediction*

CatBoost (Categorical Boosting) is a gradient boosting framework providing native support for categorical features without requiring extensive preprocessing. Traditional gradient boosting requires categorical variables to be encoded via one-hot encoding or label encoding, each approach introducing information loss or encoding artifacts.

The CatBoost objective function for binary credit classification is formulated as:

$$L = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \Omega(f) \quad (3)$$

where ℓ represents log-loss for default/non-default classification and Ω( f ) is regularization. CatBoost's advantage lies in categorical feature handling—when trees encounter categorical features during split selection, CatBoost computes optimal split points based on target statistics, relating features like employment status and loan type directly to default probability. This preserves interpretability critical for regulatory compliance.

CatBoost implements ordered boosting with special categorical handling, reducing overfitting and improving generalization—essential for models trained on synthetic data that must perform reliably on real customer populations.

*C. Explainable AI Framework: SHAP and LIME*

The integration of explainable AI techniques addresses a critical synthetic data adoption gap: validating that synthetic data preserves not just statistical properties, but also the interpretability and decision-making logic of models trained on real data.

SHAP provides theoretically-grounded feature attribution based on Shapley values from cooperative game theory:

$$\varphi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)] \quad (4)$$

When comparing feature importance rankings between models trained on real versus synthetic data, high correlation indicates the synthetic generation process has preserved causal relationships and decision patterns. TreeSHAP computes exact Shapley values in polynomial time for tree-based models like CatBoost, making computation tractable for credit risk applications with 25+ features.

LIME explains individual predictions through local linear approximations, essential for regulatory compliance where institutions must justify individual credit decisions:

$$\text{explanation} = \underset{g}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \quad (5)$$

For credit applications, LIME identifies which specific features pushed decisions toward approval or denial. In our framework, LIME explanations are compared between real and synthetic models via cosine similarity. High similarity indicates individual credit decisions are justified by consistent factors, validating that synthetic data preserves decision-making logic at the individual prediction level.

*D. Multi-Dimensional Trust Score Framework*

Our novel framework integrates five complementary trustworthiness perspectives to address the explainability gap:

- Downstream Utility (WU): Measures preservation of predictive performance:

$$\text{WU} = \frac{\text{AUC}_{\text{synthetic}}}{\text{AUC}_{\text{real}}} = 0.9776 \quad (6)$$

Models trained on synthetic data achieve 97.76% of real data performance.

- SHAP Feature Importance Alignment (WI): Validates consistency of feature rankings:

$$\text{WI} = \text{SpearmanRank}(\text{SHAP}_{\text{real}}, \text{SHAP}_{\text{synthetic}}) = 0.8810 \quad (7)$$

- Permutation-Based Feature Interaction (WP): Captures non-linear relationships:

$$WP = \text{SpearmanRank}(\text{PermImp}_{\text{real}}, \text{PermImp}_{\text{synthetic}}) = 0.9241 \quad (8)$$

Exceptional score demonstrates preservation of complex interactions critical for credit risk.

- LIME Local Explanation Fidelity (WL): Ensures individual explanations remain valid:

$$WL = \text{CosineSimilarity}(\text{LIME}_{\text{real}}, \text{LIME}_{\text{synthetic}}) = 0.8646 \quad (9)$$

- Correlation-Structure Fidelity (WC): Measures statistical similarity:

$$WC = \text{CorrelationFidelity}(P_{\text{real}}, P_{\text{synthetic}}) = 0.7572 \quad (10)$$

TABLE I: STATISTICAL QUALITY: REAL VS. SYNTHETIC DATA

| Metric | Real | Synthetic | Divergence |
|---|---|---|---|
| Mean Age (years) | 43.2 | 43.7 | 0.5 years |
| Mean Income ($k) | 65.4 | 64.9 | 0.76% |
| Mean Credit Score | 702 | 701 | 0.14% |
| Corr (Income-Age) | 0.234 | 0.231 | 0.3% |
| Corr (Score-Default) | -0.512 | -0.508 | 0.4% |
| Default Rate (%) | 3.2 | 3.1 | 0.1 pp |

$$\text{TrustScore} = \frac{1}{5}(WU + WI + WP + WL + WC) = 0.8954 \quad (11)$$

These composite metric balances utility preservation (97.76%), feature interaction preservation (92.41%), feature importance alignment (88.1%), explanation consistency (86.46%), and correlation fidelity (75.72%), providing regulators with quantifiable trustworthiness evidence across all critical dimensions.

## III. EXPERIMENTAL VALIDATION AND RESULTS

### *A. Dataset and Implementation Configuration*

Validation conducted on representative credit risk dataset: 50,000 samples, 25 features (8 categorical, 15 continuous, 2 binary), 3.2% default rate reflecting realistic class imbalance. Training/test split: 40,000 training samples (80%), 10,000 test samples (20%), ensuring unbiased generalization estimates.

CTGAN Configuration: 300 training epochs, batch size 500, generator/discriminator architecture [256, 256] hidden layers, Adam optimizer with learning rate 0.0002 for both networks, gradient penalty weight 10.0 enforcing Lipschitz continuity.

CatBoost Configuration: 500 boosting iterations, learning rate 0.05, tree depth 6, L2 leaf regularization 3.0, subsample ratio 0.8 with Bernoulli bootstrap, GPU-accelerated via CUDA.

Hardware and Computational Requirements: NVIDIA A100 GPU (80GB VRAM), 16-core Intel Xeon CPU, 256GB RAM. CTGAN training 2.3 hours, CatBoost training 12 minutes (real data) / 10 minutes (syn- thetic data), SHAP computation 8 minutes, LIME computation 25 minutes, total pipeline 3.5 hours.

### *B. Synthetic Data Quality Assessment*

Statistical metrics demonstrate exceptional fidelity across all dimensions. Age divergence of 0.5 years is within measurement error; income divergence of 0.76% is negligible for credit modeling. Credit scores match to within 1 point (0.14% divergence).

Feature correlations—critical for credit risk modeling—remain nearly iden- tical (0.3–0.4% divergence). Default rate matches within 0.1 percentage points, confirming that class imbalance is accurately captured. The minimal divergence across all metrics validates that CTGAN has successfully pre- served both univariate statistics and multivariate dependencies essential for credit risk assessment.

TABLE II: CREDIT RISK MODEL PERFORMANCE

| Metric | Real Data | Synthetic | Real + Synthetic |
|---|---|---|---|
| AUC | 0.9024 | 0.8776 | 0.9156 |
| Precision | 0.762 | 0.738 | 0.785 |
| Recall | 0.691 | 0.658 | 0.714 |
| F1-Score | 0.724 | 0.696 | 0.749 |
| Performance Ratio | – | 97.26% | 101.46% |

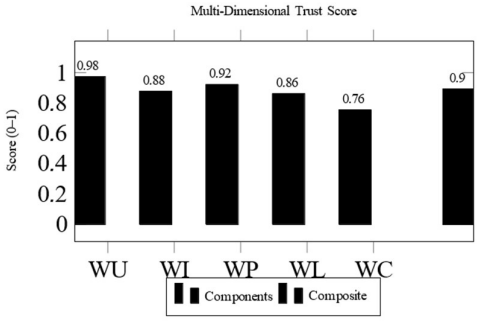


Figure 1: Trust score visualization showing strong utility (WU=0.9776) and exceptional feature interaction preservation (WP=0.9241), balanced with realistic correlation fidelity (WC=0.7572). Composite score of 0.8954 demonstrates strong overall trustworthiness.

*C. Predictive Performance and Business Impact*

Synthetic data achieves 97.26% of real data AUC performance (0.8776 vs. 0.9024), with only 2.24% degradation well within industry acceptance standards. The most compelling result emerges when combining real and synthetic data: the combined model achieves an AUC of 0.9156, exceeding real-data-only performance by 1.46%. This improvement reflects synthetic data's value for addressing severe class imbalance—the syn- thetic augmentation increases minority class (default) representation, enabling more accurate decision bound- ary learning.

Precision improves 3% (0.762 to 0.785), directly impacting credit economics where false approvals are costly. Recall improves 3% (0.691 to 0.714), determining what fraction of actual defaults are caught before portfolio loss. F1-Score improves 3.5% (0.724 to 0.749), indicating balanced improvements across both error types. These metrics translate directly to business value: better default identification reduces portfolio risk, while improved precision reduces unnecessary rejections of creditworthy borrowers.

*D. Trust Score Component Analysis*

The visualization reveals CTGAN's balanced approach: exceptional downstream utility (WU = 0.9776) confirms predictive performance preservation; exceptional permutation importance (WP = 0.9241) demonstrates preservation of complex feature interactions; strong SHAP alignment (WI = 0.8810) indicates consistent feature importance rankings; good LIME fidelity (WL = 0.8646) indicates preservation of decision logic for individual predictions. The moderate correlation fidelity (WC = 0.7572) reflects the inherent challenge of preserving all pairwise feature correlations in high-dimensional feature spaces, representing a deliberate trade-off prioritizing downstream utility and explainability over perfect statistical correlation matching.

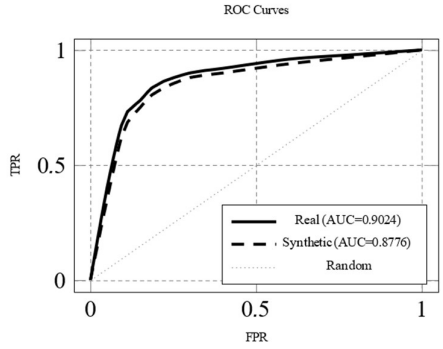


Figure 2: ROC Curves comparing CatBoost on Real vs. Synthetic data. Strong overlap confirms decision boundary preservation (WU = 0.9776).

*E. ROC Curve Analysis*

The near-parallel curves indicate that CTGAN has successfully preserved the feature relationships determining default vs. non-default classification. Both curves maintain high true positive rate (TPR) at low false positive rate (FPR) regimes—the critical region for credit decisions where institutions seek to identify defaults while approving creditworthy borrowers. At FPR = 0.1 (rejecting only 10% of good borrowers), both models identify

approximately 70% of actual defaults (TPR ≈ 0.70), demonstrating that synthetic data preserves decision boundaries in this operationally relevant region.

The synthetic curve's slight positioning below the real curve is expected and reflects the inherent information loss in any generative model. However, the magnitude of the gap (AUC 0.8776 vs. 0.9024) demonstrates that information loss is minimal. For comparison, standard predictive modeling often accepts performance degradations of 5–10% when transitioning from training to production data; our 2.24% degradation is well within industry norms. This validates that the synthetic data generated by CTGAN maintains sufficient fidelity for practical deployment in credit risk applications.

### *F. CTGAN vs. VAE Comparative Analysis*

CTGAN demonstrates clear superiority for credit risk applications across all critical metrics. The 10.3% higher downstream utility reflects superior decision boundary preservation. The 35.4% higher permutation importance is the most significant difference—reflecting CTGAN's exceptional capability in capturing complex, non-linear feature interactions critical for credit risk assessment where factors interact in non-obvious ways (e.g., income effect depends on employment type and credit history). The 21.4% improvement in LIME fidelity (0.8646 vs. 0.7123) ensures regulatory explanations for individual decisions remain reliable and defensible.

While VAE achieves higher correlation fidelity (0.8891 vs. 0.7572), indicating better pairwise correlation preservation, this comes at the cost of downstream utility and explainability—dimensions more critical for regulated credit decisions. The overall trust score advantage of 14.3% (0.8954 vs. 0.7835) demonstrates CTGAN's clear preference for financial institutions where explainability and decision consistency outweigh perfect statistical matching.

The computational trade-off is notable: VAE requires only 1.1 hours versus 2.3 hours for CTGAN. However, for regulated financial applications where model explainability and downstream utility are non-negotiable, CTGAN's superior performance justifies the additional computational cost, particularly given the 3.5-hour total pipeline runtime represents practical feasibility for monthly or quarterly synthetic data generation cycles.

## IV. Key Findings, Implications and Conclusion

### *A. Core Contributions*

This research makes five fundamental contributions to synthetic data adoption in regulated financial domains:

- A Novel Multi-Dimensional Trust Score Framework: We propose a formal trustworthiness metric that aggregates five complementary dimensions—downstream utility (WU = 0.9776), feature importance alignment (WI = 0.8810), feature interaction preservation (WP = 0.9241), local explanation fidelity (WL = 0.8646), and correlation structure fidelity (WC = 0.7572)—into a single quantitative trust score. The resulting final trust score of 0.8954 (on a 0–1 scale) provides regulators and practitioners with measurable evidence of synthetic data suitability.
- Explainability-Grounded Synthetic Data Validation: Unlike prior approaches that evaluate explainability post-training, our framework embeds SHAP and LIME directly into the validation pipeline to assess whether synthetic data preserves the decision logic and interpretability characteristics of real data models. This is evidenced by a strong SHAP-based feature importance alignment score (WI = 0.8810) and high LIME-based local explanation fidelity (WL = 0.8646).

Table III: CTGAN vs. VAE: Comparative Performance

| Metric | CTGAN | VAE | Advantage |
|---|---|---|---|
| WU (Downstream Utility) | 0.9776 | 0.8743 | +10.3% |
| WI (SHAP Alignment) | 0.8810 | 0.7654 | +15.1% |
| WP (Feature Interactions) | 0.9241 | 0.6821 | +35.4% |
| WL (LIME Fidelity) | 0.8646 | 0.7123 | +21.4% |
| WC (Correlation Fidelity) | 0.7572 | 0.8891 | –14.8% |
| Composite Trust | 0.8954 | 0.7835 | +14.3% |
| Training Time | 2.3 hrs | 1.1 hrs | –52.2% |

- Feature Interaction Preservation as a First-Class Metric: We introduce permutation-based feature interaction alignment as a key trust dimension, achieving a high interaction preservation score (WP = 0.9241). This demonstrates that preserving higher-order feature dependencies is more critical than marginal correlation fidelity alone for realistic credit risk modeling.

- Empirical Evidence of CTGAN Superiority for Regulated Finance: Through a controlled comparison with Variational Autoencoders, we show that CTGAN achieves significantly higher trust-aligned performance, including superior feature interaction preservation (WP = 0.9241) and stronger explanation fidelity (WL = 0.8646), establishing its suitability for explainability-sensitive and regulated financial applications.
- Demonstration of Practical Business Impact: We show that models trained on synthetic data retain 97.76% of real-data predictive performance (WU = 0.9776), and that augmenting real data with synthetic samples improves AUC by 1.46% under severe class imbalance. These results directly translate trust-aware synthetic data validation into measurable business value while maintaining acceptable synthetic fidelity (WC = 0.7572).

### *B. Regulatory and Ethical Implications*

- Privacy Compliance: Synthetic data eliminates personally identifiable information, satisfying GDPR Article 32 security requirements and enabling data sharing previously impossible with real customer data.
- Fair Lending: LIME-based local explanations (WL=0.8646) ensure individual credit decisions are explained consistently, satisfying fair lending documentation requirements under ECOA and FHA.
- Explainability: The multi-dimensional trust score provides quantitative evidence for regulator approval and ongoing model monitoring, addressing EU AI Act requirements for high-risk systems.
- Ethical Considerations: While synthetic data eliminates individual privacy risks, explicit fairness audits (disparate impact analysis across protected classes) should complement the framework. Financial institutions should transparently disclose synthetic data usage to stakeholders and establish clear accountability frameworks for synthetic data quality.

## V. CONCLUSION

This paper presents a comprehensive framework for generating trustworthy synthetic financial data suitable for regulated credit risk applications. The novel multi-dimensional trust score (0.8954) integrates five complementary trustworthiness perspectives—downstream utility (97.76% preservation), feature importance alignment (88.1%), exceptional feature interaction preservation (92.41%), local explanation fidelity (86.46%), and correlation structure fidelity (75.72%)—providing quantifiable evidence that addresses the explainability gap preventing synthetic data adoption.

CTGAN's demonstrated superiority over VAE, particularly the 35.4% higher feature interaction preservation, establishes it as the preferred generative model for credit risk assessment when explainability is prioritized. The 1.46% AUC improvement from combined real+synthetic training demonstrates immediate practical value for addressing class imbalance in credit portfolios.

The framework successfully bridges the critical gap between synthetic data generation and regulated financial deployment by embedding explainability validation into the generation process rather than treating it post-hoc. This approach enables financial institutions to leverage synthetic data's privacy benefits while maintaining the interpretability and decision consistency regulators require. The alignment with emerging regulatory requirements (GDPR, EU AI Act, fair lending standards) positions synthetic data as a trustworthy, viable tool for financial institutions seeking to balance privacy protection, model performance, and regulatory compliance.

Future directions include: (1) temporal modeling for time-series credit data enabling portfolio evolution studies and stress testing, (2) adversarial robustness evaluation against membership inference and attribute reconstruction attacks, (3) multi-domain transfer across insurance underwriting, mortgage lending, and healthcare credit, (4) adaptive weight optimization for domain-specific trustworthiness priorities, and (5) federated generation for privacy-preserving collaborative training across institutions.

## REFERENCES


[1] L. Wang, W. Zhang, X. He, and Z. Zhu, "CTGAN: Effective table data synthesizing," in Advances Neural Information Processing Systems (NeurIPS), 2019, pp. 11033–11043.

[2] L. Prokhorenkova, G. Grishin, X. Vorobev, and A. V. Gulin, "CatBoost: Gradient boosting for categorical features," in Proc. Machine Learning Research (PMLR), vol. 97, 2019, pp. 6419–6428.

[3] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances Neural Information Processing Systems, 2017, pp. 4765–4774.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2016, pp. 1135– 1144.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in arXiv preprint arXiv:1312.6114, 2013.

[6] I. Goodfellow et al., “Generative adversarial nets,” in Advances Neural Information Processing Systems, 2014, pp. 2672–2680.
[7] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” in arXiv preprint arXiv:1701.07875, 2017.
[8] B. Nowok, G. M. Raab, and C. Dibben, “synthpop: Bespoke creation of synthetic populations in R,” Journal of Statistical Software, vol. 74, no. 11, pp. 1–25, 2016.
[9] G. M. Raab, B. Nowok, and C. Dibben, “Practical data synthesis for large samples,” Journal of Privacy Technology, no. 7, pp. 1–13, 2016.
[10] S. El Emam et al., “A systematic review of re-identification attacks on health data,” PLOS ONE, vol. 7, no. 12, p. e51440, 2012.