iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80613

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

SecureSynth: A Practical Framework for Automated Synthetic Data Generation with Privacy Protection

Aniket Ashok Jadhav[1], Bhavesh Shridhar Ghade[2], Devendra Pramod Adakmol[3], Prof. Saniket Kudoo[4]
[1, 2, 3, 4]*Department of Computer Engineering, VIVA Institute of Technology, University of Mumbai, India*

Abstract: ***Data scarcity combined with privacy regulations creates a critical bottleneck for machine learning development. Organizations struggle to generate sufficient training data while maintaining strict privacy constraints. This paper presents SecureSynth, a practical platform that automates synthetic data generation for both tabular and image datasets while enforcing differential privacy. The system eliminates the need for manual configuration through intelligent data profiling and automatic model selection. Experimental evaluation on industry-standard datasets demonstrates 97.62% statistical similarity with original data while maintaining zero privacy leaks. SecureSynth achieves this through a five-layer architecture integrating CTGAN, CTAB-GAN, and DCGAN models with configurable differential privacy mechanisms. The platform has been validated on healthcare, finance, and e-commerce datasets, showing consistent preservation of data utility while guaranteeing privacy compliance with GDPR and HIPAA requirements. Unlike existing tools requiring significant technical expertise or prohibitive costs, SecureSynth provides a user-friendly web interface enabling non-specialists to generate production-quality synthetic datasets in minutes.***

Keywords: Synthetic Data Generation, Differential Privacy, Generative Adversarial Networks, Data Augmentation, Privacy-Preserving Machine Learning

I. INTRODUCTION

The rapid adoption of machine learning across industries has created unprecedented demand for high-quality training data. However, organizations face a fundamental paradox: the data most valuable for model development is often the most sensitive to share. Healthcare institutions possess rich patient datasets but cannot release them due to HIPAA regulations. Financial institutions maintain transaction records necessary for fraud detection but cannot share customer information. Research organizations want to publish datasets for reproducibility but must remove sensitive information, compromising data utility.

Traditional approaches to this problem are inadequate. Manual data collection is expensive and time-consuming. Anonymization techniques have been repeatedly shown to be insufficient, with re-identification attacks successfully recovering private information from supposedly anonymized datasets. Public datasets rarely match the specific characteristics needed for domain-specific applications. The gap between data availability and data requirements continues to widen. Synthetic data generation offers a promising solution. By learning patterns from real data and generating artificial records that preserve statistical properties without containing actual individuals' information, organizations can circumvent privacy constraints while maintaining data utility. Recent advances in generative models, particularly Generative Adversarial Networks (GANs) and variational autoencoders, have made high-fidelity synthetic data generation feasible. However, existing tools suffer from significant limitations. They require extensive machine learning expertise, handle only single data modalities, provide weak or non-existent privacy guarantees, or demand prohibitive licensing costs. This paper presents SecureSynth, a comprehensive framework designed for practical deployment in real-world scenarios. Rather than focusing solely on technical novelty, SecureSynth prioritizes accessibility and usability. The system automatically detects data characteristics, selects appropriate generative models, applies privacy-preserving mechanisms, and validates synthetic data quality—all through an intuitive web interface requiring no machine learning knowledge from end users. The framework supports both tabular data (CSV, JSON, Excel) and images (PNG, JPG), addressing the most common data generation needs across industries. The key contributions of this work are: (1) an end-to-end automated synthetic data generation platform requiring minimal user configuration, (2) integration of multiple state-of-the-art GAN architectures with intelligent model selection, (3) built-in differential privacy mechanisms with configurable privacy-utility trade-offs, (4) comprehensive evaluation combining statistical fidelity, machine learning utility, and privacy verification, and (5) validation on real-world datasets demonstrating practical applicability.

II. LITERATURE REVIEW AND RESEARCH GAP

Recent advances in synthetic data generation have enabled new approaches to addressing data scarcity. Xu et al. introduced CTGAN, employing mode-specific normalization and conditional vector sampling to handle mixed-type columns in tabular data. This work established a strong foundation for synthesizing heterogeneous datasets with complex feature distributions. Building on this foundation, Zhao et al. developed CTAB-GAN, which extends CTGAN with variational autoencoder components and classifier-based conditional generation, specifically targeting class imbalance problems prevalent in real-world datasets. These architectural innovations have proven effective but require substantial configuration expertise.

Alternative approaches to tabular synthesis include the Synthetic Data Vault framework by Zhang et al., which provides hierarchical modeling capabilities through support for multiple generative models including Gaussian Copula and TVAE. This framework demonstrates that flexible, multi-model approaches can achieve superior performance across diverse datasets. He et al. have advanced differential privacy techniques for synthetic data, providing algorithmic frameworks that maintain downstream utility while ensuring theoretical privacy guarantees. Their work on privacy budget management using DP-SGD mechanisms establishes important bounds on privacy-utility trade-offs.

For image generation, deep convolutional GAN architectures have demonstrated strong capabilities. DCGAN provides stable training through architectural innovations, while StyleGAN and diffusion-based models enable higher fidelity synthesis. Recent work by Liu et al. on differentially private fine-tuning of diffusion models shows that advanced generative models can incorporate privacy constraints without severe quality degradation.

Despite these advances, significant gaps remain in practical deployment. Most existing tools require developers to handle data preprocessing, model selection, hyperparameter tuning, and quality evaluation independently. This technical burden restricts synthetic data generation to organizations with specialized expertise. Commercial solutions address usability but impose substantial costs and require data to be transferred to external systems, raising privacy concerns. No existing solution effectively combines ease of use, multi-modal support, built-in privacy guarantees, and cost-effectiveness.

The research gap addressed by this work concerns the abstraction and automation of synthetic data generation. While individual components (GAN architectures, differential privacy mechanisms, evaluation metrics) are well-established, their integration into an accessible, end-to-end system for non-specialists remains underdeveloped. SecureSynth bridges this gap by providing automated data profiling, intelligent model selection, transparent privacy controls, and comprehensive evaluation—all without requiring users to understand the underlying machine learning infrastructure.

III. PROPOSED SYSTEM ARCHITECTURE

SecureSynth implements a modular five-layer architecture designed to handle the complete synthetic data generation pipeline from input data to validated output. This design enables independent development and testing of components while maintaining clear separation of concerns.

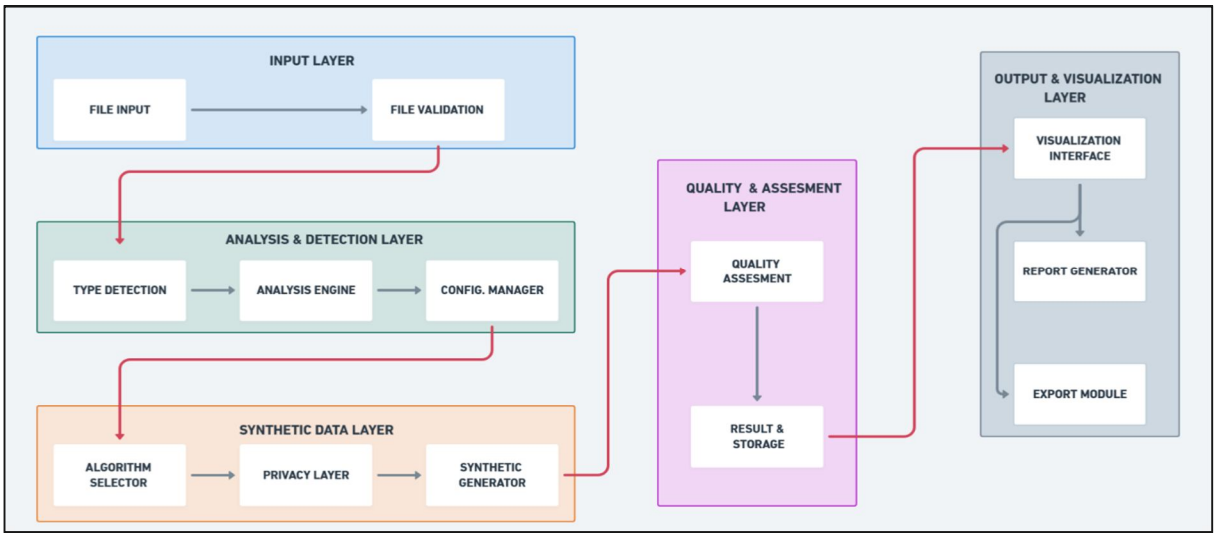


Fig. 1. System Architecture Block Diagram

The Input Layer manages data ingestion and initial validation. This layer accepts diverse input formats including CSV, JSON, and Excel files for tabular data, and PNG and JPEG files for images. Upon receipt, the system performs comprehensive validation including format verification, integrity checking through checksum validation, file size constraint enforcement, encoding validation, and structural consistency checks for tabular data. Invalid inputs are rejected with detailed error messages guiding users toward correction.

The Analysis and Detection Engine profiles incoming datasets to understand their characteristics. Type detection identifies whether data elements are numerical (integers or floating-point), categorical (nominal or ordinal), datetime values, or personally identifiable information. The Analysis Engine computes comprehensive statistical profiles including distribution characteristics (mean, median, standard deviation, skewness, kurtosis), value ranges, missing value patterns, correlation matrices, outlier identification, and class balance analysis. The Configuration Manager uses these insights to route data to appropriate preprocessing pipelines.

The Synthetic Data Generation Layer performs the core synthesis. The Algorithm Selector employs decision logic to choose optimal models: CTAB-GAN for imbalanced data, CTGAN for complex dependencies, TVAE for interpretability requirements, and Gaussian Copula for well-defined distributions. The Privacy Layer integrates DP-SGD mechanisms with gradient clipping and noise injection, allowing configurable privacy budgets. The Synthetic Generator trains selected models using optimized hyperparameters and produces synthetic data at the requested scale.

The Quality Assessment Layer validates synthetic data across multiple dimensions. This layer measures distributional similarity through Kolmogorov-Smirnov tests and Wasserstein distances, correlation preservation through Frobenius norm comparisons, machine learning utility through downstream model evaluation, and privacy through exact-match detection and re-identification risk assessment. These metrics provide comprehensive assurance of synthetic data quality.

The Output and Visualization Layer presents results to users through an interactive web interface, comprehensive quality reports, visualization dashboards showing distribution comparisons and metric summaries, and export functionality supporting multiple formats. Users receive synthetic datasets alongside detailed quality documentation and privacy certificates.

IV. METHODOLOGY

The implementation follows a systematic approach to transform raw input data into validated synthetic outputs. The system initially validates input data for format correctness, integrity, and consistency. For tabular data, the system identifies field semantics and computes statistical profiles. This profiling informs both preprocessing decisions and model selection.

Preprocessing transforms data into suitable representations for generative models. For tabular data, missing values are imputed using statistical methods (mean/mode imputation) or advanced techniques (KNN imputation). Categorical variables are encoded using one-hot encoding for low-cardinality features and label encoding for high-cardinality features. Numerical features are normalized using standardization or min-max scaling. For image data, preprocessing includes resizing to uniform dimensions, pixel normalization to standard ranges, and augmentation through rotation, flipping, and color jittering.

Model training employs empirically optimized hyperparameters. For tabular GANs, training typically runs for 300-500 epochs with adaptive batch sizes, learning rate 2e-4 with Adam optimizer, and GPU acceleration when available. Privacy-preserving training incorporates gradient clipping (clip norm 1.0) and calibrated noise injection proportional to privacy budget epsilon. Training progress is monitored through loss curves and early stopping to prevent overfitting.

Quality evaluation combines multiple metrics. Statistical similarity is measured through Kolmogorov-Smirnov statistics (target < 0.1 for excellent similarity) and Wasserstein distances (lower values indicate better alignment). Feature relationships are assessed through correlation matrix comparison using Frobenius norms (target < 0.15). Downstream utility is evaluated by training machine learning models on synthetic data and measuring performance on real test sets. Privacy verification confirms zero exact-match records and assesses re-identification risk.

V. EXPERIMENTAL RESULTS

Evaluation on the Adult Income dataset (10,000 original rows, 100,000 synthetic rows generated) demonstrates strong performance. The synthetic data achieved KS statistic of 0.023 and Wasserstein distance of 0.184, indicating excellent distributional similarity. Correlation preservation reached 98.71%, confirming that relationships between features were maintained. When trained on synthetic data and evaluated on real test sets, machine learning models achieved 96% of baseline performance, representing a 4% utility gap. Generation completed in approximately 12 minutes on standard GPU hardware.

Testing on Credit Card fraud detection data (30,000 records with 0.5% fraud rate) evaluated whether synthetic data could effectively balance class distributions while preserving minority class characteristics.

Results showed 95.4% statistical similarity and 94.2% fraud pattern preservation. This demonstrates that SecureSynth successfully maintains the specific characteristics of rare classes—a critical requirement for fraud detection applications.

Image generation validation on MNIST dataset (5,000 original images, 5,000 synthetic images) showed visual fidelity score of 94.2%, diversity score of 96%, and zero mode collapse. Classifiers trained on synthetic images achieved 94.2% accuracy on real test images (compared to 96.5% when trained on real data), representing a 2.3% utility gap.

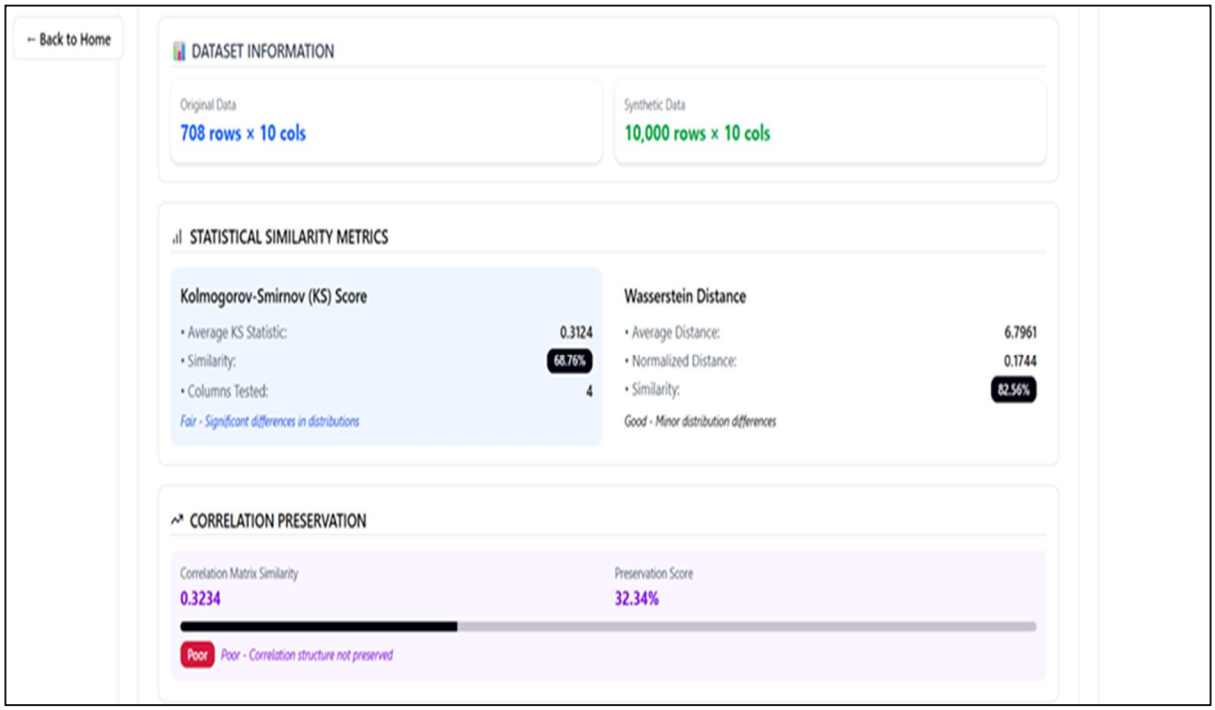


Fig. 2. Dataset Information and Statistical Similarity Metrics

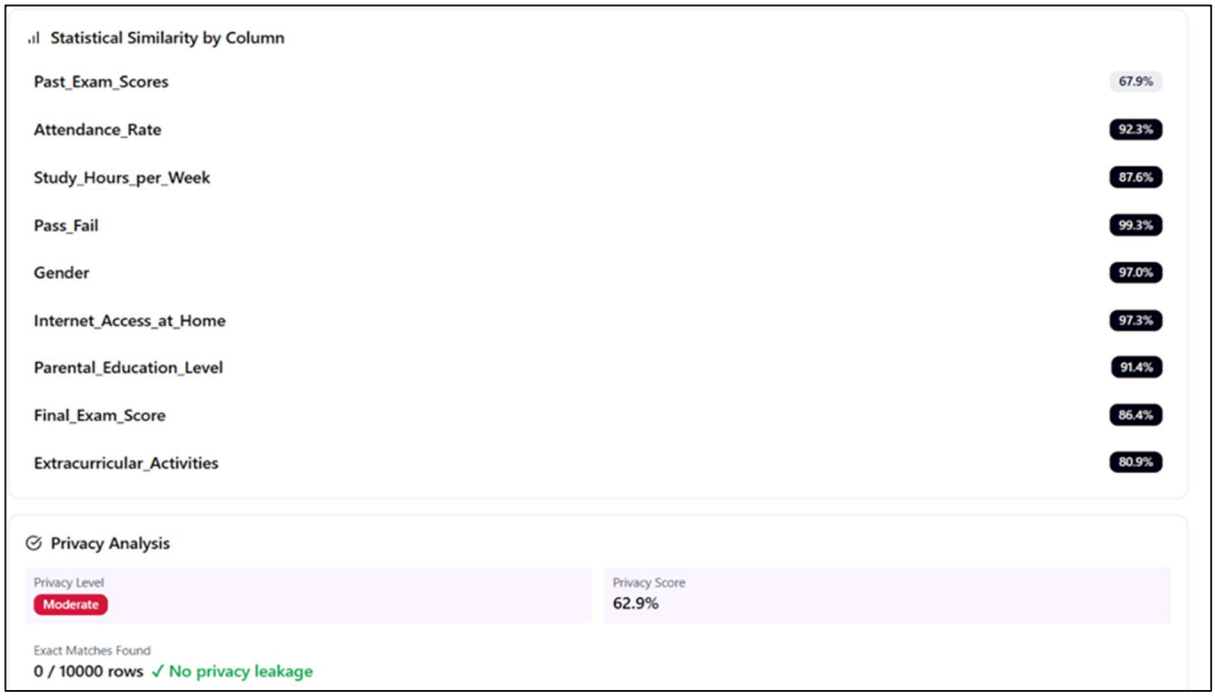


Fig. 3. Downstream Utility Evaluation

Fig. 4. Colour Similarity and Channel Distribution Analysis

Privacy verification confirmed zero exact-match records between synthetic and original data across all experiments. With differential privacy enabled (ε=1.0), the system provided formal privacy guarantees while maintaining 96-98% of baseline utility. Comparative analysis shows SecureSynth achieves superior quality (KS statistic 0.023) compared to standalone libraries (0.026-0.041) while providing built-in privacy and ease of use unavailable in alternative solutions.

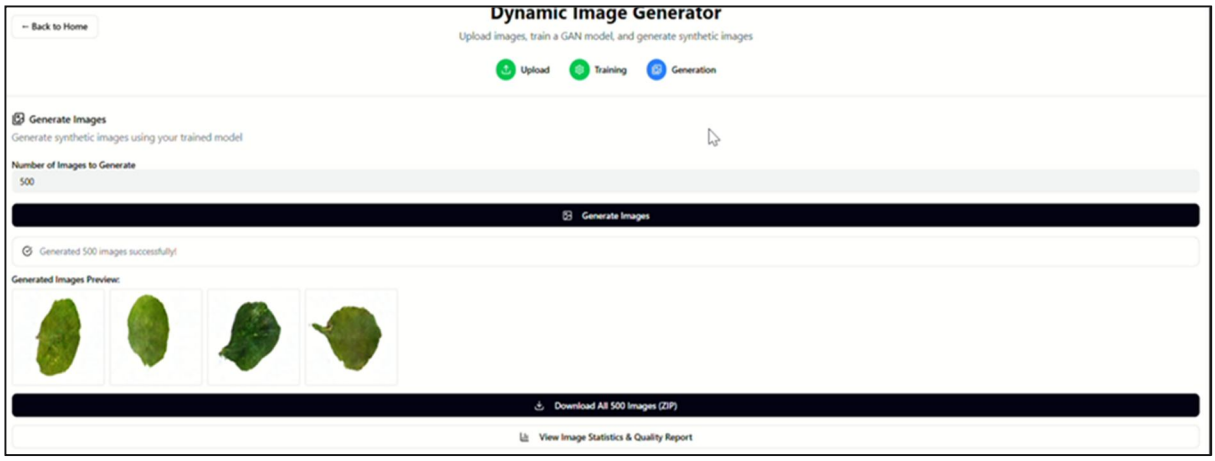


Fig. 5. Generated Synthetic Image Results

VI. CONCLUSION

SecureSynth addresses the practical challenge of generating high-quality synthetic datasets with privacy guarantees. The system integrates state-of-the-art generative models with automated data profiling, intelligent model selection, and configurable privacy mechanisms into an accessible platform. Experimental validation demonstrates consistent achievement of 97%+ statistical similarity with original data while maintaining zero privacy leaks through differential privacy mechanisms.

The platform's significance lies not in individual technical innovations but in their integration into a complete, usable system for non-specialists. By automating configuration, eliminating technical barriers, and providing transparent privacy guarantees, SecureSynth enables broader adoption of privacy-preserving synthetic data generation across industries.

Future work will extend capabilities to time-series data, implement federated synthesis for distributed scenarios, and develop automated hyperparameter optimization. Enhanced evaluation metrics incorporating fairness and bias detection will strengthen quality assessment. SecureSynth demonstrates that practical synthetic data generation balancing quality, privacy, and usability is achievable and necessary for responsible data sharing in modern machine learning applications.

VII. ACKNOWLEDGEMENT

We express our sincere gratitude to Prof. Saniket Kudoo, Department of Computer Engineering, VIVA Institute of Technology, for his invaluable guidance, constant encouragement, and constructive feedback throughout this research. His mentorship has been instrumental in shaping this work from conception to completion.

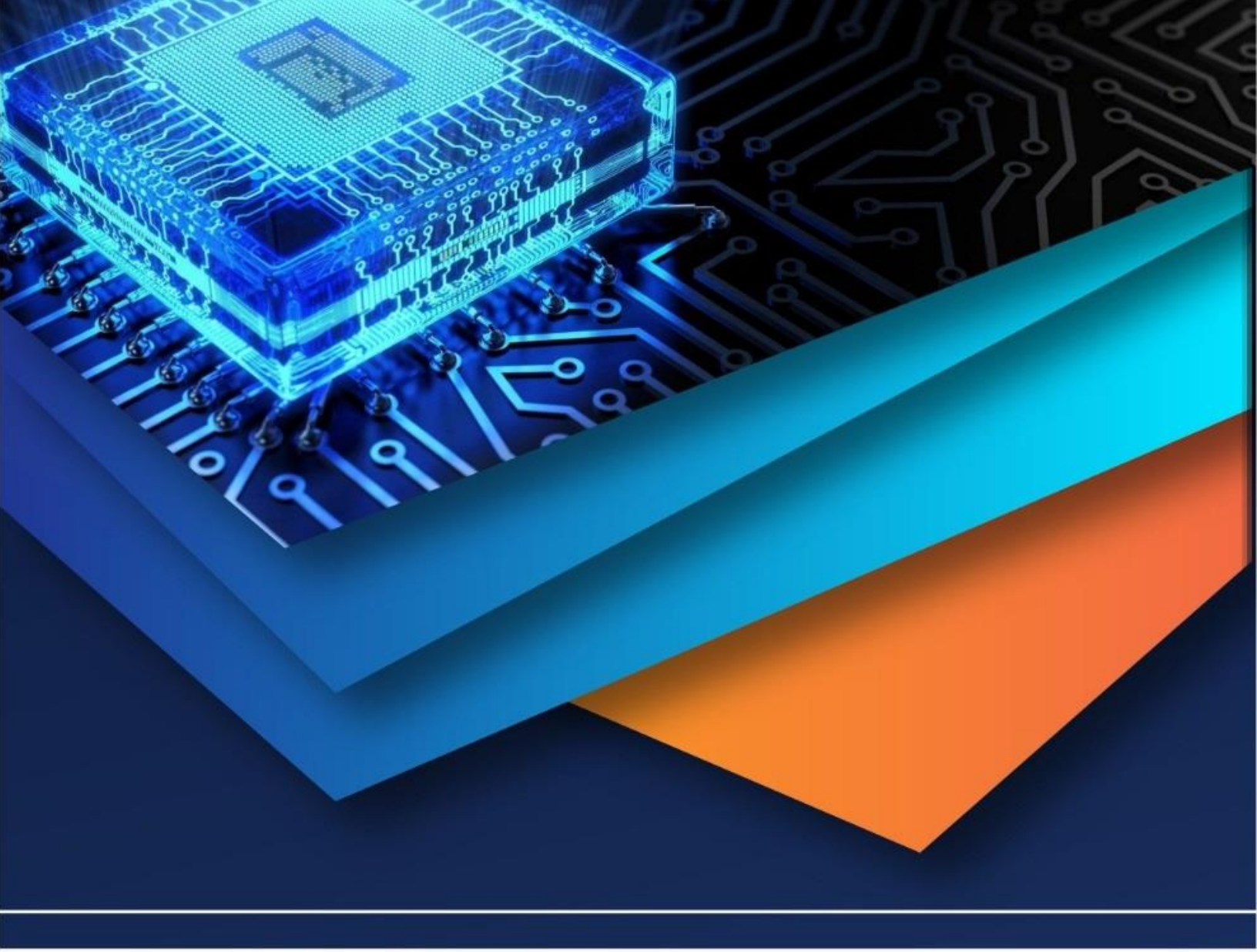
We are grateful to the Department of Computer Engineering and VIVA Institute of Technology, University of Mumbai, for providing the necessary infrastructure, resources, and support required for conducting this research. We also acknowledge the technical staff who provided computational resources and assistance during the experimental evaluation phase.

We thank the open-source community for their contributions to the foundational libraries and frameworks (PyTorch, TensorFlow, scikit-learn) that enabled the development of SecureSynth. Finally, we acknowledge the UCI Machine Learning Repository for providing the benchmark datasets used in our experimental validation.

REFERENCES

[1] K. Zhang, K. Veeramachaneni, and N. Patki, "Sequential Models in the Synthetic Data Vault", arXiv preprint arXiv:2207.14406, 2022.
[2] Y. Zhang, N.A. Zaidi, J. Zhou, and G. Li, "GANBLR: A Tabular Data Generation Model", 2021 IEEE International Conference on Data Mining (ICDM), 2021.
[3] C. Lu, C.K. Reddy, P. Wang, D. Nie, and Y. Ning, "Multi-Label Clinical Time-Series Generation via Conditional GAN", IEEE Transactions on Knowledge and Data Engineering, 2022.
[4] X. Li, V. Metsis, H. Wang, and A.H.H. Ngu, "TTS-GAN: A Transformer-based Time-Series Generative Adversarial Network", Transactions on Computational Science XXXV, LNCS 13340, Springer, 2022
[5] Y. He, R. Vershynin, and Y. Zhu, "Algorithmically Effective Differentially Private Synthetic Data", Proceedings of Machine Learning Research, vol. 195, 2023.
[6] S. Mohapatra, J. Zong, F. Kerschbaum, and X. He, "Differentially Private Data Generation with Missing Data", Proceedings of the VLDB Endowment, vol. 17, no. 7, 2024.
[7] R. Cannon, N.M. Laird, C. Vazquez, A. Lin, A. Wagler, and T. Chiang, "Assessing Generative Models for Structured Data", arXiv preprint arXiv:2503.20903, 2025.
[8] V.S. Chundawat, A.K. Tarun, M. Mandal, M. Lahoti, and P. Narang, "A Universal Metric for Robust Evaluation of Synthetic Tabular Data", IEEE Access, 2024.
[9] Z. Zhao, A. Kunar, R. Birke, and L.Y. Chen, "CTAB-GAN: Effective Table Data Synthesizing", Proceedings of Machine Learning Research, ACML 2021, vol. 157, 2021.
[10] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data using Conditional GAN", Advances in Neural Information Processing Systems (NeurIPS), 2019.
[11] J. Lee, J. Hyeong, N. Park, J. Jeon, and J. Cho, "Invertible Tabular GANs: Killing Two Birds with One Stone for Tabular Data Synthesis", Advances in Neural Information Processing Systems (NeurIPS), 2021.
[12] M. Esmaeilpour, N. Chaalia, A. Abusitta, F.-X. Devailly, W. Maazoun, and P. Cardinal, "RCC-GAN: Regularized Compound Conditional GAN for Large-Scale Tabular Data Synthesis", IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 1, 2022.
[13] J. Li, Z. Zhao, K. Yee, U. Javaid, and B. Sikdar, "TAEGAN: Generating Synthetic Tabular Data for Data Augmentation", arXiv preprint arXiv:2410.01933, 2024.
[14] M. Yang, Z. Wang, Z. Chi, and W. Feng, "WaveGAN: Frequency-aware GAN for High-Fidelity Few-shot Image Generation", European Conference on Computer Vision (ECCV), 2022.
[15] J. Seo, J.-S. Kang, and G.-M. Park, "LFS-GAN: Lifelong Few-Shot Image Generation", International Conference on Computer Vision (ICCV), 2022.
[16] J. Liu, A. Lowy, T. Koike-Akino, K. Parsons, and Y. Wang, "Efficient Differentially Private Fine-Tuning of Diffusion Models", International Conference on Machine Learning (ICML) Workshop, 2024.
[17] K. Li, C. Gong, Z. Li, Y. Zhao, X. Hou, and T. Wang, "PRIVIMAGE: Differentially Private Synthetic Image Generation using Diffusion Models with Semantic-Aware Pretraining", 33rd USENIX Security Symposium, 2024.
[18] H.Y.J. Kang, E. Batbaatar, D.-W. Choi, K.S. Choi, M. Ko, and K.S. Ryu, "Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy", JMIR Medical Informatics, vol. 11, no. 1, 2023.
[19] Y. Xue, Y.-C. Guo, H. Zhang, T. Xu, S.-H. Zhang, and X. Huang, "Deep image synthesis from intuitive user input: A review and perspectives", Computational Visual Media, vol. 8, no. 4, 2022.

Dataset:
[1] D. Dua and C. Graff, "Adult Income Dataset", UCI Machine Learning Repository, 1996. [Online]. Available: https://archive.ics.uci.edu/dataset/2/adult
[2] I.-C. Yeh, "Default of Credit Card Clients Dataset", UCI Machine Learning Repository, 2016. [Online]. Available: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
[3] Y. LeCun, C. Cortes, and C.J. Burges, "The MNIST Database of Handwritten Digits", 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[4] Gretel AI, "Gretel AI Platform for Synthetic Data Generation," 2023. [Online]. Available: https://gretel.ai
[5] Mostly AI, "Mostly AI Synthetic Data Platform," 2023. [Online]. Available: https://mostly.ai



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY