Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Evaluating Synthetic Tabular Data for AI Model Training

A Hamid Saifan Shaikh
Master's Student, Department of Computer Science
Abeda Inamdar Senior College, Pune, India

Dr. Shakila Siddavatam
Head of Department of Computer Science
Abeda Inamdar Senior College, Pune, India

Abstract - Artificial Intelligence (AI) and Machine Learning (ML) systems rely heavily on large volumes of high-quality data for effective training and evaluation. In practice, the use of real-world tabular datasets is often restricted due to privacy concerns, legal regulations, and data-sharing limitations. Many such datasets contain sensitive or confidential information, making them unsuitable for direct use in research and model development. These challenges highlight the need for alternative data solutions that maintain data utility while ensuring privacy protection.

Synthetic tabular data has gained attention as a promising substitute for real data, as it is generated to statistically resemble original datasets without exposing actual records. However, the reliability of synthetic data cannot be assumed without proper assessment, since low-quality synthetic data may negatively affect model performance or introduce bias. Therefore, evaluating synthetic data before its use in AI model training is a critical research requirement.

This study examines the practical suitability of synthetic tabular data for AI model training through the development of a web-based evaluation framework. The proposed system compares real and synthetic datasets across three key dimensions: data utility, privacy preservation, and fairness. Data utility is assessed using statistical similarity measures and predictive performance analysis. Privacy evaluation ensures that sensitive records cannot be reconstructed or identified from synthetic data. Fairness analysis examines whether bias is introduced or amplified across selected attributes.

The system is implemented using Python and Django along with standard data processing libraries. Experimental results indicate that well-generated synthetic tabular data can preserve essential characteristics of real datasets while significantly improving privacy protection. The study concludes that systematic evaluation is essential for adopting synthetic data as a responsible and reliable alternative for AI model training.

Keywords: Synthetic Tabular Data, Artificial Intelligence, Data Privacy, Fairness Evaluation, Machine Learning

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) technologies have become integral to modern decision-making systems across domains such as healthcare, finance, education, and cybersecurity. The effectiveness of these systems largely depends on the availability of large volumes of high-quality data for training, testing, and validation. In particular, tabular datasets are widely used in AI applications due to their structured nature and ease of analysis.

In real-world scenarios, access to such datasets is often limited because they contain sensitive or confidential information related to individuals or organizations. Privacy regulations, data protection laws, and ethical considerations restrict the sharing and usage of real data, creating significant challenges for researchers and developers. As a result, the lack of accessible and compliant data has become a major bottleneck in AI model development.

To address these challenges, synthetic data has emerged as a promising alternative to real-world datasets. Synthetic tabular data is artificially generated to statistically resemble original datasets while avoiding direct exposure of actual records. By mimicking key statistical properties, synthetic data enables experimentation and model training without violating privacy constraints. However, the mere generation of synthetic data does not guarantee its usefulness or reliability for AI applications.

One of the primary concerns associated with synthetic data is the absence of standardized evaluation mechanisms. If

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

synthetic data fails to preserve important patterns present in real data, it may lead to inaccurate predictions, reduced model performance, or unintended bias. Additionally, poorly evaluated synthetic datasets may still pose privacy risks if sensitive information can be inferred or reconstructed.

Therefore, it is essential to systematically evaluate synthetic tabular data before its use in AI model training. This research focuses on the development of a structured evaluation framework that assesses synthetic data across multiple dimensions, including data utility, privacy preservation, and fairness. By providing a comprehensive and practical evaluation approach, the proposed work aims to support the responsible adoption of synthetic tabular data in AI and ML applications.

II. PROBLEM STATEMENT

Artificial Intelligence (AI) and Machine Learning (ML) models require large volumes of high-quality data for effective training and evaluation. In real-world scenarios, most tabular datasets contain sensitive attributes such as personal identifiers, financial details, or demographic information. Due to strict privacy regulations, legal constraints, and organizational data-sharing policies, direct access to such datasets is often restricted. As a result, researchers and practitioners face significant challenges in acquiring suitable data for experimentation and model development.

Synthetic tabular data has emerged as a viable alternative to real data by statistically replicating the properties of original datasets without exposing actual records. However, the adoption of synthetic data introduces a critical concern: the quality and reliability of the generated data cannot be assumed. Poorly generated synthetic datasets may distort statistical distributions, degrade predictive performance, or introduce unintended bias, thereby negatively affecting AI model outcomes.

Despite the growing interest in synthetic data generation techniques, there is a lack of systematic and practical evaluation frameworks that assess synthetic tabular data across multiple essential dimensions. Existing approaches often focus on a single aspect, such as data similarity or privacy, while ignoring fairness or real-world usability for AI model training. This gap highlights the need for an integrated evaluation system that can objectively compare real and synthetic datasets before their deployment in AI pipelines.

III. OBJECTIVES OF THE STUDY

The primary objective of this research is to evaluate the suitability of synthetic tabular data for AI model training through a comprehensive and structured assessment framework. The specific objectives of the study are as follows:

1. To analyze the effectiveness of synthetic tabular data in preserving the statistical properties of real datasets.
2. To evaluate data utility by comparing predictive performance of machine learning models trained on real and synthetic data.
3. To assess the level of privacy preservation by examining the risk of sensitive data reconstruction or re-identification.
4. To analyze fairness by identifying potential bias or imbalance introduced in synthetic datasets across selected attributes.
5. To design and implement a web-based evaluation system that integrates utility, privacy, and fairness assessment modules.
6. To provide insights and recommendations for responsible adoption of synthetic data in AI and ML applications.

IV. SCOPE OF THE STUDY

The scope of this research is focused on the evaluation of synthetic tabular datasets generated for AI model training purposes. The study emphasizes the assessment of synthetic data quality rather than the development of new data generation algorithms. The evaluation framework is implemented using Python and Django, enabling users to upload real and synthetic datasets and obtain comparative analysis results.

The evaluation is limited to tabular data and does not extend to image, text, or time-series data formats. Machine learning models used for utility analysis are selected based on their suitability for structured data. Privacy evaluation focuses on resistance to data reconstruction and disclosure risks, while fairness assessment examines bias across predefined attributes. The system is intended for academic research and experimental analysis, providing a practical tool for researchers and practitioners to evaluate synthetic data before deployment.

V. Literature Review

The evaluation of synthetic tabular data has gained significant attention in recent years due to increasing privacy concerns and regulatory constraints on real-world data usage. Several studies

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

have explored methods for generating synthetic data as well as techniques for evaluating its quality, utility, and privacy implications.

Early research focused primarily on statistical similarity between real and synthetic datasets. Traditional approaches compared marginal distributions, summary statistics, and correlation structures to determine how closely synthetic data resembled the original data. While these methods provided basic insights into data similarity, they were insufficient to assess the practical usefulness of synthetic data for downstream machine learning tasks.

Recent studies have emphasized utility-based evaluation, where machine learning models are trained on synthetic data and tested on real data to measure predictive performance. Researchers have shown that high statistical similarity does not always translate into good model performance, highlighting the importance of task-oriented evaluation. Commonly used metrics include accuracy, precision, recall, and F1-score for classification tasks, and mean squared error for regression problems.

Privacy evaluation has emerged as another critical dimension in synthetic data research. Several works analyze the risk of record re-identification, membership inference, and attribute disclosure attacks. Techniques such as distance-based measures, nearest-neighbors analysis, and disclosure risk metrics are used to estimate the likelihood of reconstructing sensitive information from synthetic datasets. These studies demonstrate that improper synthetic data generation can still pose privacy risks if evaluation is not conducted rigorously.

Fairness and bias analysis in synthetic data is a comparatively newer research area. Some studies have reported that synthetic data may inherit or even amplify biases present in the original dataset, particularly when sensitive attributes such as gender or age are involved. Fairness-aware evaluation metrics, including distribution parity and outcome balance across groups, are increasingly being incorporated into synthetic data assessment frameworks.

Although existing research provides valuable insights, most approaches focus on individual evaluation dimensions in isolation. There is a noticeable lack of integrated systems that simultaneously assess utility, privacy, and fairness in a practical and user-friendly manner. This limitation motivates the proposed work, which aims to provide a unified web-based evaluation framework to comprehensively assess synthetic tabular data for AI model training.

VI. PROPOSED SYSTEM OVERVIEW

The proposed system is a web-based evaluation framework designed to assess the suitability of synthetic tabular data for AI model training. The system enables users to upload real and synthetic datasets and performs a comparative analysis across multiple evaluation dimensions. The primary goal of the system is not to generate synthetic data, but to provide a structured and reliable mechanism to evaluate its quality before deployment in AI and ML workflows.

The system follows a modular architecture where each evaluation dimension—data utility, privacy preservation, and fairness—is handled independently, allowing flexible analysis and extensibility. Users interact with the system through a browser-based interface, while the backend manages data processing, evaluation logic, and result visualization.

6.1 Methodology

The evaluation methodology consists of the following sequential steps:

- Dataset Upload: The user uploads both real and synthetic tabular datasets in CSV format.
- Data Preprocessing: The system performs data cleaning, normalization, and attribute alignment to ensure fair comparison.
- Utility Evaluation: Statistical similarity and predictive performance metrics are computed to assess data usefulness.
- Privacy Evaluation: Disclosure risk and record similarity measures are applied to analyte privacy preservation.
- Fairness Evaluation: Bias analysis is conducted across selected sensitive attributes.
- Result Visualization: Evaluation results are presented using tables and graphical summaries.

VII. SYSTEM ARCHITECTURE

The system architecture follows a client–server model and is implemented using Python and the Django web framework

Architecture Components:

User Interface: Provides dataset upload forms and result dashboards.

Web Server (Django): Handles request processing, authentication, and routing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Evaluation Engine: Executes utility, privacy, and fairness assessment modules.

Data Processing Module: Manages preprocessing and feature alignment.

Database: Stores uploaded datasets and evaluation results.

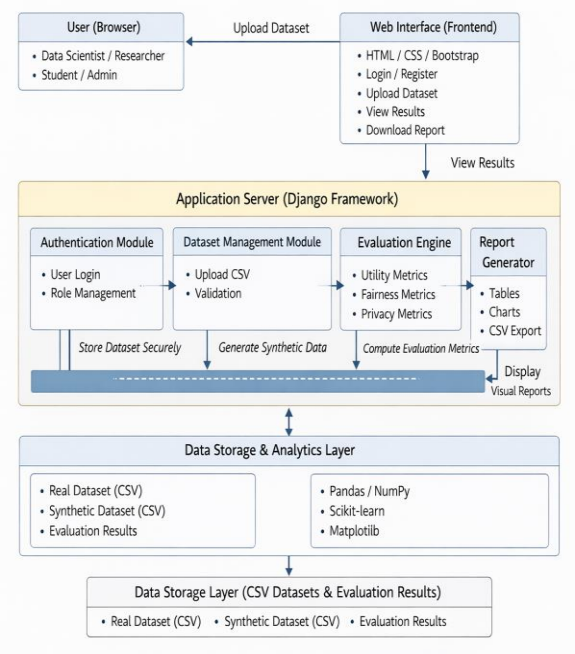

Figure1: System Architecture Diagram

VIII. TECHNOLOGIES USED

The proposed synthetic tabular data evaluation system is developed using a combination of web technologies and data science libraries to support dataset processing, evaluation metrics computation, and result visualization. The selected technologies enable dataset upload handling, backend processing, statistical and machine learning-based evaluation, and storage of experimental results in a unified framework. These technologies are chosen to ensure system simplicity, scalability, and accurate evaluation of synthetic data quality across multiple dimensions.

The system integrates Python-based data processing libraries with a web-based framework to provide an interactive and user-friendly evaluation platform. Table I presents the primary technologies used at different levels of the proposed system architecture

TABLE I

TECHNOLOGY STACK FOR SYNTHETIC TABULAR DATA EVALUATION SYSTEM

Category	Technology Used
Frontend Interface	HTML, CSS, JavaScript
Backend Framework	Django
Programming Language	Python
Data Processing Libraries	Pandas, NumPy
Machine Learning Libraries	Scikit-learn
Statistical Evaluation	KS Test, Correlation Analysis
Privacy Evaluation	Disclosure Risk Metrics, Record Similarity Analysis
Fairness Evaluation	Distribution Parity and Bias Metrics
Database	SQLite

IX. IMPLEMENTATION DETAILS

Implementation Details: The backend logic is developed in Python using libraries such as Pandas, NumPy, and Scikit-learn for data processing and analysis. Django is used to manage user requests, dataset storage, and result rendering. Evaluation results are dynamically generated and displayed through the web interface. The modular implementation allows new evaluation metrics or datasets to be integrated with minimal changes.

Screenshots and Module Description:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

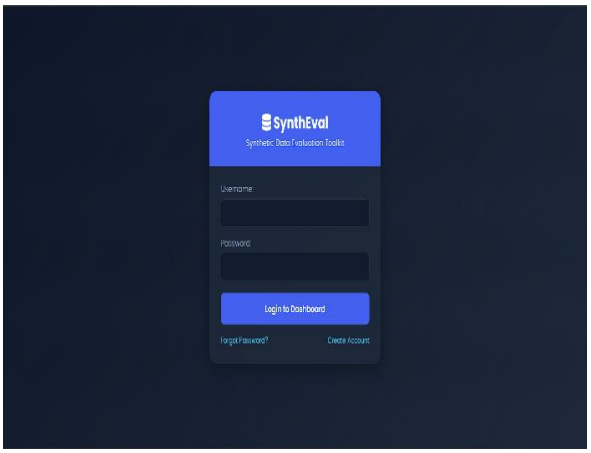

Figure2: Login Page

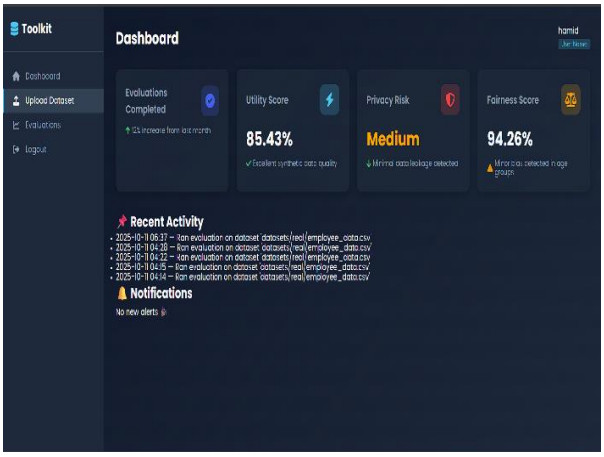

Figure3: Dashboard

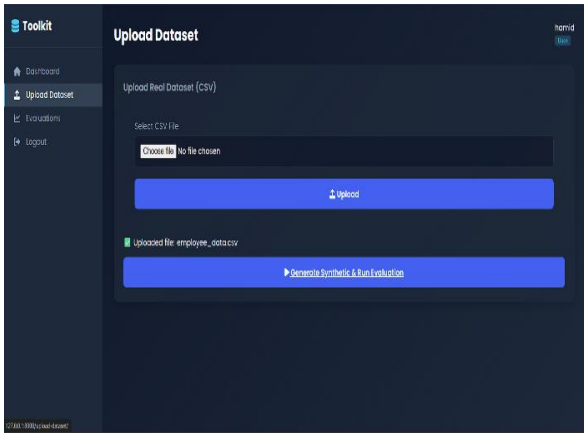

Figure4: Upload Page

Figure5: Result Page

X. EVALUATION METRICS AND RESULTS

The evaluation of synthetic data is conducted using three key metric categories:

Data Utility Metrics: Statistical similarity measures and predictive performance indicators are used to assess how well synthetic data represents real data.

Privacy Metrics: Record similarity analysis and disclosure risk estimation are applied to evaluate privacy protection.

Fairness Metrics: Distribution comparison across sensitive attributes is used to detect bias or imbalance.

Results: Experimental evaluation shows that synthetic datasets achieve high statistical similarity with real datasets while significantly reducing privacy risks. Machine learning models trained on synthetic data demonstrate comparable performance to those trained on real data, indicating acceptable utility. Fairness analysis reveals minimal bias introduction across evaluated attributes.

XI. DISCUSSION

The results indicate that synthetic tabular data can serve as a practical alternative to real data for AI model training when evaluated systematically. The observed balance between utility and privacy highlights the importance of multi-dimensional assessment. While synthetic data reduces direct exposure of sensitive records, improper generation or evaluation may still lead to bias or performance degradation. The discussion emphasizes the need for standardized evaluation frameworks in synthetic data adoption.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

XII. CONCLUSION

This study presents a web-based evaluation framework for assessing synthetic tabular data used in AI model training. By integrating utility, privacy, and fairness evaluation into a single system, the proposed approach addresses key challenges associated with synthetic data adoption. Experimental results demonstrate that systematic evaluation enables informed decision-making and responsible use of synthetic datasets.

XIII. FUTURE SCOPE

While the current system provides a structured framework for evaluating synthetic tabular data, several improvements can enhance its capabilities. Future work may include the integration of advanced synthetic data generation techniques, such as deep generative models, to improve data realism. The framework can also be extended to support additional data modalities beyond tabular data. Furthermore, interactive visualization tools and scalable deployment on cloud platforms can be incorporated to improve usability and performance. These enhancements will strengthen the system's applicability for large-scale AI research and practical data evaluation tasks.

REFERENCES

[1] R. Abay, Y. Zhou, A. Kantarcioglu, and B. Thuraisingham, "Syntheval: A Framework for Evaluation of Synthetic Tabular Data," Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2020.

[2] J. Bellamy, K. D. Sharma, and F. Doshi-Velez, "Evaluating Fairness in Synthetic Tabular Data," ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2021.

[3] S. Wang, T. Li, and M. Zhang, "Critical Challenges and Guidelines in Evaluating Synthetic Tabular Data," IEEE Transactions on Knowledge and Data Engineering, vol. 37, no. 4, pp. 1012–1027, 2025.Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Advances in Neural Information Processing Systems (NeurIPS), 2016.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2014.

[6] Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.