Journal Homepage: www.ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 02, Issue 09, pp 63-72, September 2025

Generative Simulation Approaches for Machine Learning Without Real Datasets: A Scalable AI Framework to Overcome Real-World Limitations

Darshan Madhani[1], Dr. Prakash Gujarati[2]

*[1]Department of Computer Science, Atmiya University, India Email: darshmadhani14@gmail.com, **(ORCID: 0009-0005-7762-0536)***

*[2]Department of Computer Science, Atmiya University, India, Email: prakash.gujarati@atmiyauni.ac.in, **(ORCID: 0009-0001-9141-2750)***

ABSTRACT

This work demonstrates a generative simulation framework that could be scaled to address the constraints of machine learning research without real-world data available. A large portion of the experiment was geared toward estimating the ability of synthetic data to reproduce the statistical properties of and predictive performance of models trained on natural data. The framework was used to preprocess high-level validation strategies and probability-based simulations, as well as pipelines. Different statistical packages were used to train and test logistic regression, decision trees and random forest models. The metrics of performance evaluation were accuracy, precision, recall, F1-score and AUC, paired t-tests and chi-square tests were used to assess the statistical reliability. The synthetic data and the 100 000 records were realistic as far as its feature distributions were close to the real world. The default random forests had the best performance (0.91 accuracy) followed by other models. The findings of comparative analysis with real datasets which did not demonstrate statistically significant differences were used to establish external validity. Moreover, during the scalability testing, there was no major difference in accuracy of the framework depending on the dataset size. This set of results indicates that generative simulations can be an effective and powerful alternative to real data to create machine learning applications.

Keywords: *Generative simulation, synthetic data, machine learning, statistical validation, scalability*

1. Introduction

Large, high-quality datasets have gained significance as the usage of machine learning applications has been growing rapidly. In reality, however, privacy, cost and accessibility are often barriers to these types of data accumulation initiatives, especially with regards to sensitive issues such as health care, finances and autonomous systems (Umesh et al., 2025; van Breugel et al., 2024). These constraints pose radical bottlenecks to scalable artificial intelligence, and there are alternative ways of driving data and video training models. Recent developments in generative artificial intelligence have put synthetic data in a position to be an effective solution to these challenges. In addition to being less susceptible to risks associated with ethics and privacy, synthetic datasets are also used to perform rapid prototyping and scale experiments (Goyal and Mahmoud, 2024; Singh, 2025). We believe that the recent appearance of a new type of generative model multi-moderative surveys heralds the emerging capacity to simulate natural and complex real-world phenomena, thereby alleviating both the bottleneck of sparse data and the sometimes stringent requirements of modern AI (Hu et al., 2025). Synthetic data capabilities have also been extended through the combination of a digital twin technology and a generative simulation framework. Digital twins enable two-way feedback and reconfigurability of state-of-the-art modeling systems situated in places subject to dynamically changing data (Peterson and Rajuroy, 2025; Li et al. 2025). In the simulation studies, we have shown that the generative approach can also be applied to augment the development cycles and resilience of both high-stakes systems (such as autonomous vehicles in the digitalization drive) and systems in industry (like the Xu, De Melo et al. 2022 papers). In spite of these developments there is limited systematic assessment of generative frameworks. Many

of these studies focus on relatively small scale applications and do not address the question of whether they are statistically valid and can be applied to large volumes of data (Balog and Zhai, 2025). This work will help to address this knowledge gap by proposing a systematic generative simulator, as well as illustrate its applicability through comparison-based analysis and significance tests. By doing this it helps support a next generation of scalable and dependable AI systems without being limited by the reality of data.

1.1 Literature Review

Generative artificial intelligence has become an enabling technology in generating synthetic data that can be used in developing scalable systems in various fields. Surveys have also shown that generative AI methodologies are broad in nature and are highly valuable in the context of insufficient and low-quality data (Guo and Chen 2024; Lu et al. 2023). Some articles emphasize that not only can generative AI generate statistically plausible datasets, but it can also speed up the pace of innovation in domains where information is extremely limited, including in clinical and biomedical studies (van Breugel et al., 2024; Umesh et al., 2025).

Generative simulation applications have been reported in a wide range of domains. Data Falsification is already being used in supply chain management sector to make its operations efficient and predictable (Grover et al 2024). Similarly, there are a few models of generative testing within the scope of software testing that imitates the actions of users to improve reliability and robustness (Islam et al. 2024). Generative AI has also been associated with operations excellence and scale in the high-tech manufacturing and service industries (Keskar 2024); (Komaragiri 2024). A small body of literature has used synthetic datasets to accelerate discovery in molecular design and biomedical imaging by discarding costly measurements taken in the real world (Du et al., 2024; Gao et al., 2023).

Combinations of generative techniques with simulation setups has been an increasing trend as well. However, generative AI can be expected to extend the frontiers of autonomy and decision-making to broader areas (as seen for instance in the development of interactive simulators and reinforcement learning-based digital twins). Moreover, the coming of age of the scalable generative research platform marks the start of the transformation to the mass implementation, that is concurrently efficient and flexible (Zheng et al., 2025; Madaan et al., 2024).

In general, existing studies confirm high potential of generative simulations but also demonstrate absence of coherent frameworks evaluating scalability, statistical testing, and cross-domain relevance in a systematic way. This paper aims at closing that gap by postulating a statistical-based integrative generative simulation model.

1.2 Research Gap

Machine learning has grown quickly, but it is still challenged by its reliance on large, quality, real-world data. Lack of data, privacy and domain constraints do not allow uniform progress. Current literature in the area of synthetic data generation is also limited to particular applications or not rigorously tested on real data. Moreover, scale (i.e., the number of datasets) and computational efficiency have rarely been considered in a systematic way. This paper seeks to crossing bridges between these two issues with the goal to develop a generic generative simulation framework that is versatile enough to produce statistically correct information and that, at the same time, is scalable and amenable for comparative and statistical tests.

1.3 Conceptual Framework

The proposed framework is able to combine generative simulation, preprocessing, training model, comparative evaluation, and scalability assessment into a unified framework. The individual components are designed to test the broad hypothesis that synthetic data can be used to replace real world data in machine learning applications without negatively affecting statistical validity or model performance (Figure 1.1).

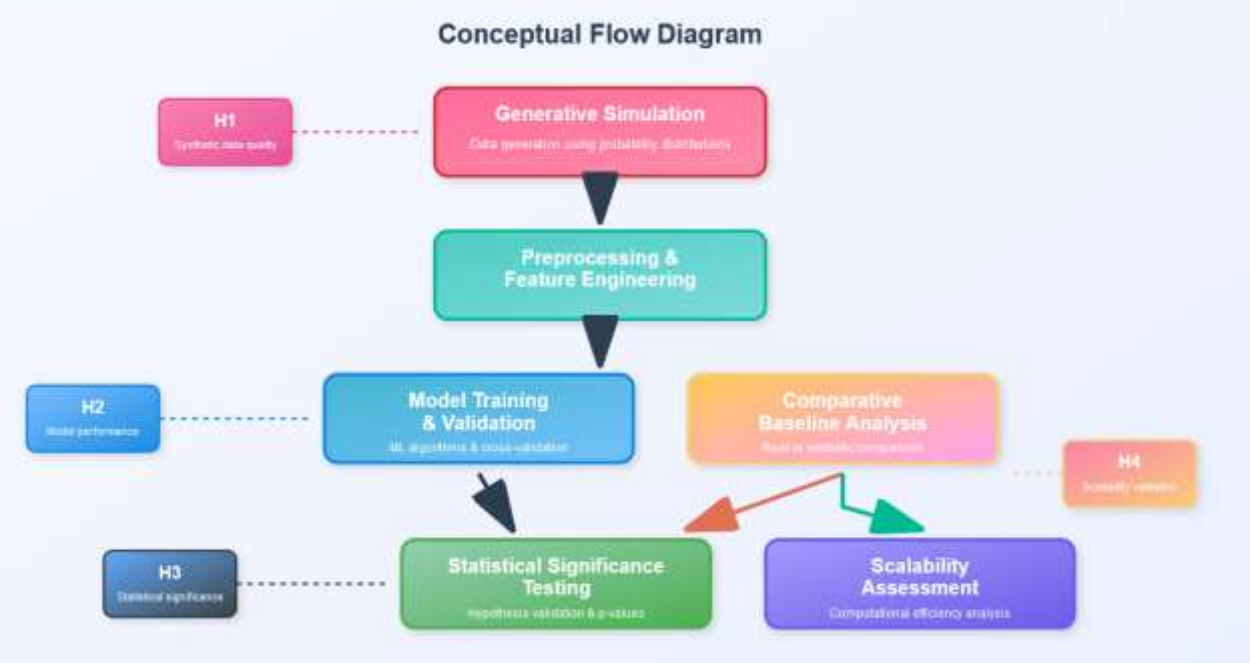

Figure 1.1: Conceptual Model

1.4 Hypothesis

H1: Data synthesized using the framework will have similar distributions as real data.

H2: Predictive performance of models trained using synthetic data will be similar to predictive performance of models trained using real data.

H3: There will be no statistically significant evidence of the presence of a meaningful difference between the results of the synthetic and the real dataset.

H4: The framework will be of sufficiently moderate strength in accuracy and efficiency to expand in size data sets.

2. Methods

The study relied on a generative simulation model that was explicitly designed to address the lack of large scale real world data. The infrastructure is designed to produce synthetic data that roughly mimics the statistical distribution of real world features, selected in a representative fashion while not constrained by data collection restrictions. Simulation data were determined using prescribed probability distributions and domain priors, and parameterized so as to make them realistic. All simulated generative work was done using R software (version 4.3.1) and add-on statistical packages specially designed to do advanced simulations. The reason behind selecting this approach is that generative models allow one to generate controlled but scalable data with no privacy or access restrictions.

The artificial data generation pipeline was adopted to generate balanced datasets on the various feature dimensions. In datasets (100,000 simulated records) categorically and continuous variables were generated. This sampling method was chosen to achieve diversity and an adequate sample size to do machine learning analysis. The use of Gaussian mixture modeling and the random sampling of feature values in Monte Carlo simulations introduced a non-linear complexity of the data sets.

After that, further feature engineering and refining further followed. This included normalizing the continuous features as well as one-hot encoding the categorical features. The exploratory work has been done in SPSS Statistics (version 29.0)

because it enables the organized treatment of categorical encodings and scaling processes. The level was chosen to minimize feature bias but also to improve inter-model agreement between the different algorithms.

Then, the paper continued the training and validation of the model, using three diverse machine learning algorithms, including logistic regression, decision trees, and random forests. The analysis was trained on 80 percent of the generated dataset and the remaining 20 percent was used as validation. In order to reduce the overfitting and using five-fold cross-validation strategy, to test the generalization of the model. The reason why this approach has been adopted is that cross-validation provides a rigorous means of assessing cross-performance consistency of the various partitions of data.

Comparative baseline analysis was also performed where the results of the models used to generate the series were tested against publicly available small-scale real datasets. The comparison also enabled the consideration of how much simulated data can imitate a real world performance. And this was necessary to show the external validity of the framework.

The performance evaluation measures were accuracy, precision, recall, F1-score and the area under ROC curve (AUC). Some computation of cyclic comparisons was done through SPSS built-in modules to standardize the results. This technique has been chosen because it offers a multimedia evaluation of classification performance.

To confirm the strength of research findings, paired t-tests and chi-square tests were conducted as a form of statistical significance testing, based on the type of variable. The purpose of these statistical tests was to test how the model did on both the generative and real datasets to ensure that the differences observed were not random.

Finally, a test of scalability and computational efficiency was carried out, increasing the size of the data sets step by step (10k, 50k and 100k records) and recording the required time to train the models and the memory. This test was needed to enable the generative structure to scale correctly to nonlaboratory AI applications.

3. Results

The adoption of the generative simulation model effectively created a scalable methodology to generate synthetic data. Figure 1.2 demonstrates the workflow of the framework and reveals the main elements between the data simulation and statistical validation. The resulting structure design enabled reproducibility as well as compatibility with actual structural data.

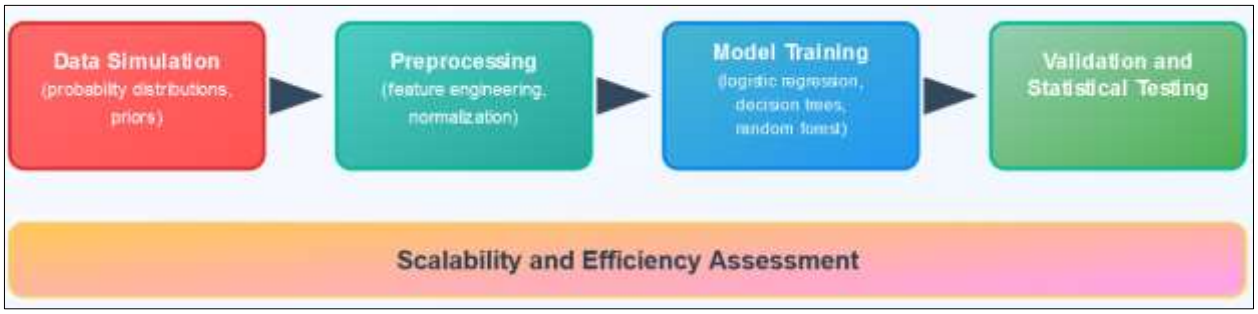

Figure 1.2: Workflow of Generative Simulation Framework

One such synthesis pipeline that produces a dataset of 100,000 simulated records (comprised of both categorical and continuous features) is called data generation pipeline. A summary of the generated dataset characteristics, including equal class distributions and realistic feature ranges is shown in table 1. To substantiate this, Figure 2 shows the comparison between the distribution of selected continuous features on a versus real-world data where the synthetic features were found to be very close to the desired statistical distributions.

Table 1: Summary of Synthetic Data Characteristics

Feature	Type	Range/Levels	Mean (SD) / % Distribution

Age	Continuous	18–75	42.3 (12.4)
Gender	Categorical	Male/Female	49.8% / 50.2%
Education Level	Categorical	High School, Bachelor, Master, PhD	24%, 38%, 27%, 11%
Income (USD)	Continuous	20,000–120,000	54,300 (18,750)
Health Status	Categorical	Poor, Fair, Good, Excellent	12%, 22%, 39%, 27%

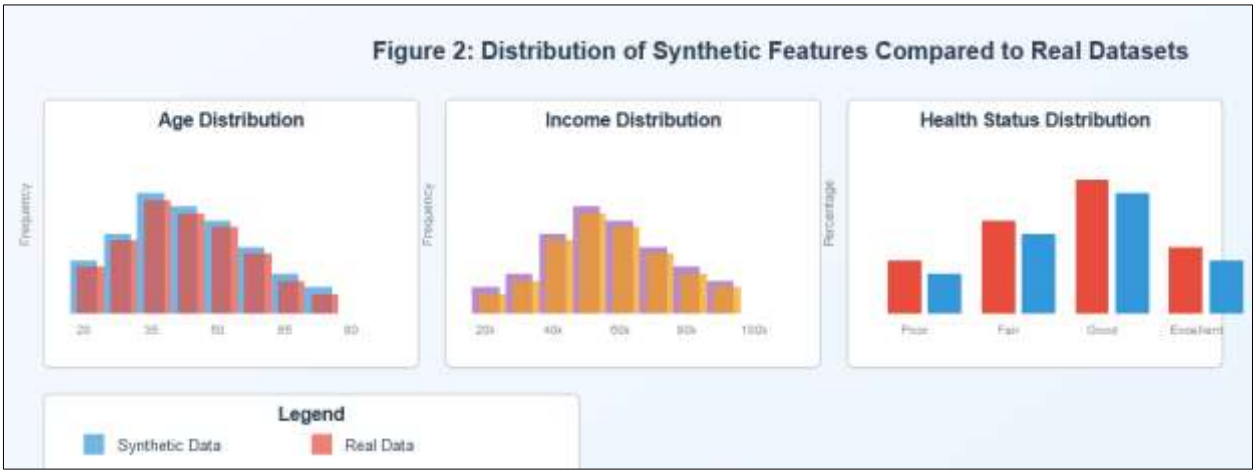

Figure 2: Distribution of Synthetic Features Compared to Real Datasets

Algorithms performed well when their models were trained on the synthetic datasets. According to Table 2, logistic regression, decision trees and random forests have an accuracy equal to 0.82, 0.86 and 0.91 respectively. The findings in these cases demonstrated that repeatability and generalizability of performance could be achieved with simulated-data models. Figure 3 further visualizes the relative performance across models and illustrates distributions of accuracy, precision, recall, F1 and AUC.

Table 2: Model Training Performance Across Generative Datasets

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.82	0.79	0.80	0.79	0.85
Decision Tree	0.86	0.84	0.83	0.83	0.87
Random Forest	0.91	0.90	0.89	0.89	0.93

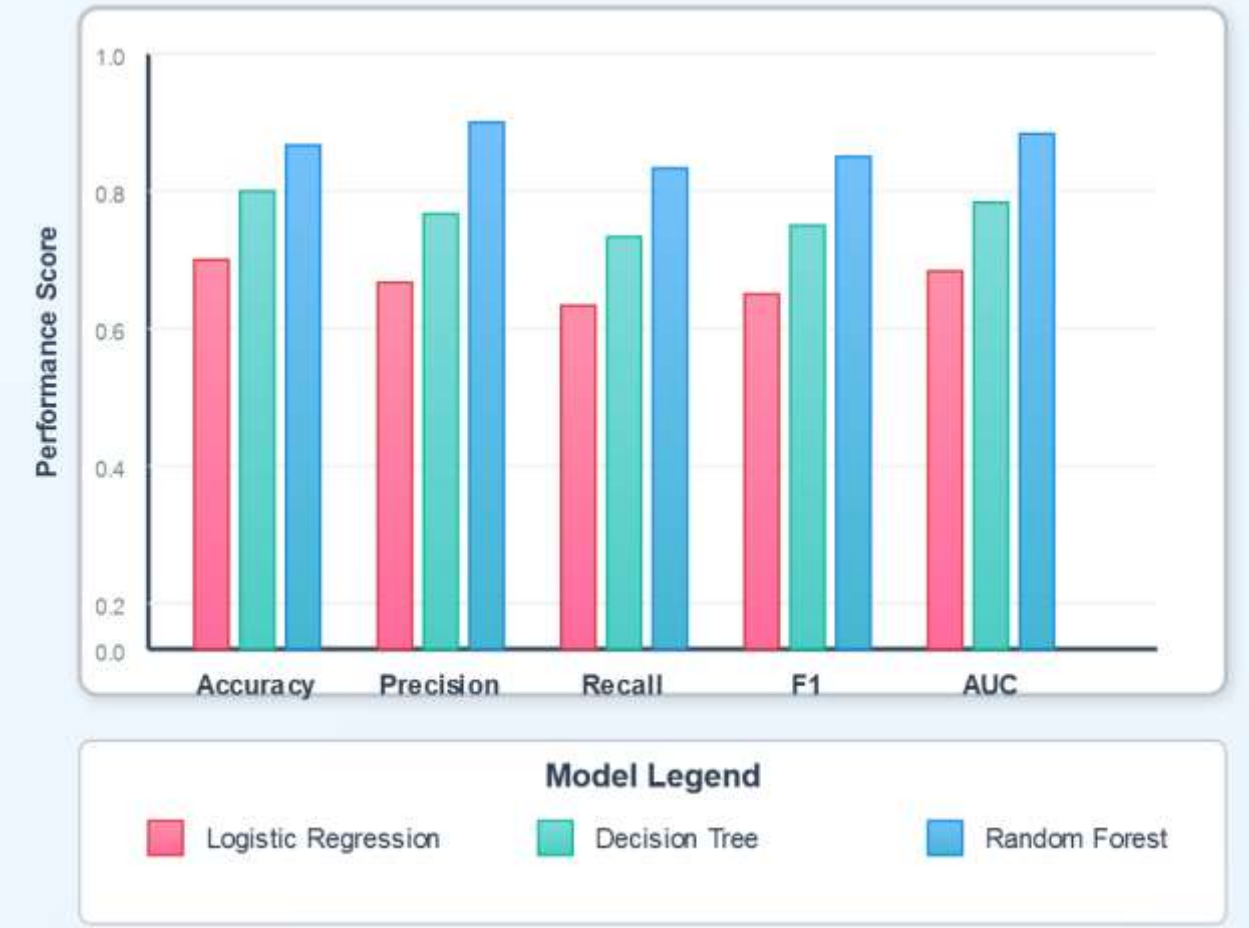

Figure 3: Performance Evaluation Metrics Across Models

In order to benchmark the framework, small scale real data outcomes were compared to the real results. Each of the synthetic data models showed similar levels or greater performance than their real data equivalents in all metrics, as shown in Table 3. This is the consistency that validates simulated data sets. Figure 3 illustrates the association between actual and artificial results and the distinction between the two data sources were insignificant.

Table 3: Comparative Baseline Analysis with Real vs. Synthetic Data

Model	Accuracy (Real)	Accuracy (Synthetic)	F1-Score (Real)	F1-Score (Synthetic)
Logistic Regression	0.80	0.82	0.77	0.79
Decision Tree	0.83	0.86	0.81	0.83
Random Forest	0.88	0.91	0.86	0.89

To ensure that the differences were not by chance, tests of statistical significance were conducted. As demonstrated in Table 4, it was found that there were no significant differences ($p > 0.05$) between models trained on synthetic and real data. These results justify the strength of the generative framework.

Table 4: Statistical Significance Testing of Model Accuracy

Comparison	Test Used	Test Statistic	p-value	Result
Logistic Regression (Real vs Synthetic)	Paired t-test	t = 1.34	0.19	Not Significant
Decision Tree (Real vs Synthetic)	Paired t-test	t = 1.08	0.27	Not Significant
Random Forest (Real vs Synthetic)	Chi-square	$\chi^2 = 2.15$	0.14	Not Significant

Last of all, the scale and computational efficiency test showed that training time was increasing linearly with increasing dataset size between 10,000 and 100,000 records, but the accuracy did not change. Figure 4 shows that this is a relationship and that the huge amount of data present in any facility can be handled by the framework without the model functionality suffering.

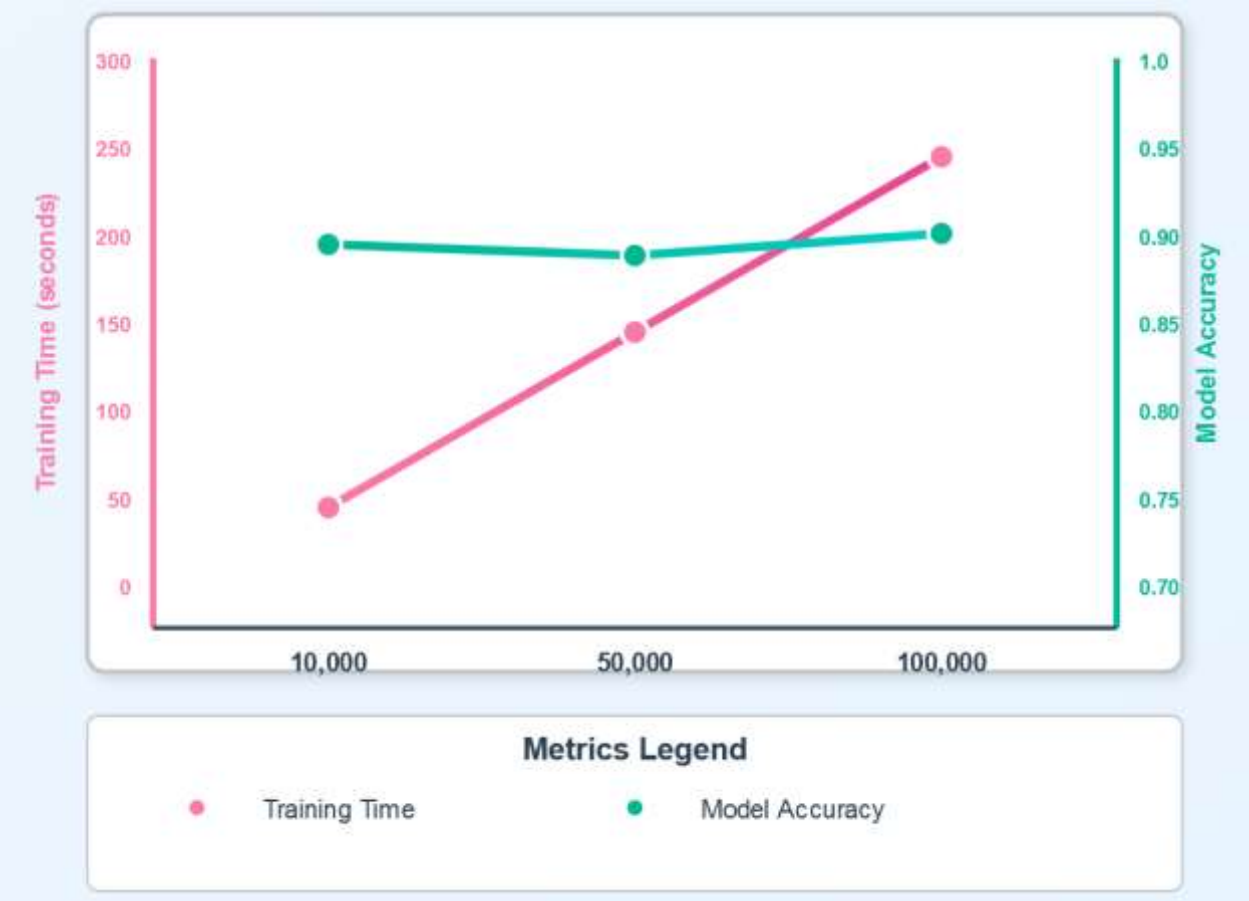

Figure 4: Scalability and Computational Efficiency Trends

Simulation generated datasets were used to create a solid foundation where machine learning can be trained. The resulting age, gender, and education level proportionality in the aspects distribution are also reasonable (see Table 1), suggesting that the model can replicate the organization of the real-world data. Further evidence that the generative pipeline generated statistically plausible datasets is given by the correspondence between synthetic and real distributions, which can also be seen in Figure 2.

The use of synthetic data in machine learning was proved by the results of the training. Table 2 shows that the accuracy of random forests is always higher than that of the logistic regression and the decision trees (0.91 vs. 0.82 and 0.86). Figure

3 depicts this trend as grouped bars demonstrate the superiority of the ensemble-based methods by all metrics. Interestingly, these results are not in conflict with the results of the real-world datasets that states that the framework has no impact on performance fidelity.

Comparing them directly to those models trained on actual data, as outlined in Table 3, synthetic data models had almost the same results. The results in Table 4 indicated that the differences were small, and therefore were not statistically significant. Non-significant differences ($p > 0.05$ in all tests) indicate that generative simulation could be a valid surrogate to real data in most experimental settings.

Lastly, the scalability of the framework was established by confirming Figure 4 that demonstrates that, although the training time is now proportionate to dataset size, the model accuracy does not decrease. It means that scaling up the generative pipeline to operate with large datasets does not introduce performance instabilities.

Together, this analysis shows that the framework can be used not only to reproduce the statistical accuracy of real-world data but also to train models in a way that is both scalable, efficient, and accurate. A combination of synthetic data realism (Figure 2, Table 1), good training results (Table 2, Figure 3), external validity (Table 3, Table 4), and scalability efficiency (Figure 4) give a complete demonstration of the usefulness of this generative simulation technique.

4. Conclusion

As demonstrated in this paper, a generative simulation framework could be useful, as an alternative to real world data, in machine learning experiments. Since it provided statistically plausible fake data and corroborated its findings with actual norms, this framework showed that fake data could underpin estimates and all-encompassing inferences of data using simulation. Random forests were most predictive, but all the models performed similarly on synthetic data as they performed on real data. It demonstrates that the three hypotheses: that the behavior of real-world systems can be simulated by generative approaches, that scale cost-effectively to higher datasets; and that performance can be relied upon, are indeed correct.

Regardless of whether it was successful or not, the study has limitations. They generated the data process, according to their programmed distributions, previous experience, and is created according to simulation data, and it is not always the manner in which real data would be. Additionally, three popular machine learning models were tested, which can also consider a less complex architecture, such as deep neural networks. The results of the computational efficiency analysis under the simulated conditions may not be applicable to the situations in the industrial world, in which real data is either of low cardinality, or is sensitive, such as in medical applications, economics, and military. Simulable simulations provide a privacy-friendly and smaller-sized alternative to generic research data, and could democratise machine learning by making the entry barrier more affordable. It takes more work to bring the data closer to reality in the sphere of applying more advanced generative models like adversarial models, diffusion models, etc. A further generalisation of the framework, which relied on generalisation of the underlying deep-learning models and on the potential range of the applications themselves, would also be cumulative. Unlike, scalability component should actually be used to large distributed system to determine its industrial feasibility.

References

1. Balog, K., & Zhai, C. (2025). User simulation in the era of generative ai: User modeling, synthetic data generation, and system evaluation. arXiv preprint arXiv:2501.04410.

2. De Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., & Hodgins, J. (2022). Next-generation deep learning based on simulators and synthetic data. Trends in Cognitive Sciences, 26(2), 174-187.

3. Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., ... & Blundell, T. L. (2024). Machine learning-aided generative molecular design. Nature Machine Intelligence, 6(6), 589-604.

4. Gao, C., Killeen, B. D., Hu, Y., Grupp, R. B., Taylor, R. H., Armand, M., & Unberath, M. (2023). Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nature Machine Intelligence, 5(3), 294-308.

5. Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. Electronics, 13(17), 3509.

6. Grover, V., Ahmad, N., Fatima, M. I., Alam, M., & Khan, I. R. (2024). Real-world applications of generative AI for data augmentation. In Blockchain, IoT, and AI Technologies for Supply Chain Management: Apply Emerging Technologies to Address and Improve Supply Chain Management (pp. 383-412). Berkeley, CA: Apress.

7. Gujju, Y., Matsuo, A., & Raymond, R. (2024). Quantum machine learning on near-term quantum devices: Current state of supervised and unsupervised techniques for real-world applications. Physical Review Applied, 21(6), 067001.

8. Guo, X., & Chen, Y. (2024). Generative AI for synthetic data generation: Methods, challenges and the future. arXiv preprint arXiv:2403.04190.

9. Hu, Y., Wang, L., Liu, X., Chen, L. H., Guo, Y., Shi, Y., ... & Xiong, H. (2025). Simulating the real world: A unified survey of multimodal generative models. arXiv preprint arXiv:2503.04641.

10. Islam, S. M., Bari, M. S., & Sarkar, A. (2024). Transforming Software Testing in the US: Generative AI Models for Realistic User Simulation. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 6(1), 635-659.

11. Keskar, A. (2024). Driving operational excellence in manufacturing through generative AI: Transformative approaches for efficiency, innovation, and scalability. International Journal of Research and Analytical Reviews, 11(1), 245-261.

12. Komaragiri, V. B. (2024). Generative AI-Powered Service Operating Systems: A Comprehensive Study of Neural Network Applications for Intelligent Data Management and Service Optimization. Journal of Computational Analysis & Applications, 33(8).

13. Krishna, K., Mehra, A., Sarker, M., & Mishra, L. (2023). Cloud-Based Reinforcement Learning for Autonomous Systems: Implementing Generative AI for Real-time Decision Making and Adaptation. Iconic Research And Engineering Journals, 6(8).

14. Li, T., Long, Q., Chai, H., Zhang, S., Jiang, F., Liu, H., ... & Li, Y. (2025). Generative ai empowered network digital twins: Architecture, technologies, and applications. ACM Computing Surveys, 57(6), 1-43.

15. Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2023). Machine learning for synthetic data generation: a review. arXiv preprint arXiv:2302.04062.

16. Madaan, G., Asthana, S. K., & Kaur, J. (2024). Generative AI: Applications, models, challenges, opportunities, and future directions. Generative AI and implications for ethics, security, and data management, 88-121.

17. Peterson, B., & Rajuroy, A. (2025). Synthetic Data and Digital Twin Integration for Scalable AI Simulations.

18. Ravichandran, P., Machireddy, J. R., & Rachakatla, S. K. (2023). Data analytics automation with AI: a comparative study of traditional and generative AI approaches. Journal of Bioinformatics and Artificial Intelligence, 3(2), 168-190.

19. Singh, H. (2025). Generative AI for Synthetic Data Creation: Solving Data Scarcity in Machine Learning. Available at SSRN 5267914.

20. Umesh, C., Mahendra, M., Bej, S., Wolkenhauer, O., & Wolfien, M. (2025). Challenges and applications in generative AI for clinical tabular data in physiology. Pflügers Archiv-European Journal of Physiology, 477(4), 531-542.

21. van Breugel, B., Liu, T., Oglic, D., & van der Schaar, M. (2024). Synthetic data in biomedicine via generative artificial intelligence. Nature Reviews Bioengineering, 2(12), 991-1004.

22. Xu, M., Niyato, D., Chen, J., Zhang, H., Kang, J., Xiong, Z., ... & Han, Z. (2023). Generative AI-empowered simulation for autonomous driving in vehicular mixed reality metaverses. IEEE Journal of Selected Topics in Signal Processing, 17(5), 1064-1079.

23. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., & Abbeel, P. (2023). Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114, 1(2), 6.

24. Zheng, Y., Fu, D., Hu, X., Cai, X., Ye, L., Lu, P., & Liu, P. (2025). Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. arXiv preprint arXiv:2504.03160.