International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 1, January – February 2022 ||

DOI: 10.15680/IJCTECE.2022.0501004

Simulating Compliance Scenarios using Synthetic Data Generation

Venkata Phanindra Lingam

Independent Researcher, USA

venakataphanindra@gmail.com

Sai Reddy Mandala

Independent Researcher, USA

mandalasaireddy9@gmail.com

ABSTRACT: Synthetic data generation for simulating compliance scenarios will offer organizations an opportunity to revolutionize the testing and validation of their compliance systems. Many modern regulations, particularly in finance, healthcare, and data security, have become distinctly convoluted and require robust systems developed to predict where potential compliance breaches could occur. This paper takes a look at synthetic data and its role in simulating compliance scenarios and how real-time data applications and synthetic data generation technologies can be used, thus enabling organizations to create accurate and scalable compliance tests. Key challenges for studies on synthetic data include data accuracy, overfitting, and computational intensity, providing solutions that should enhance the credibility of compliance simulations while providing room for scalability.

KEYWORDS: Synthetic data, compliance scenarios, real-time data, data privacy, fraud detection, regulatory testing

I. INTRODUCTION

To be in compliance with rules and industry standards, businesses should always act on their own accord. In finance, healthcare, and data privacy, for failure to comply, huge fines, terrible PR, and business interruptions may occur. Conventional ways of testing the performance of compliance—manual auditing, static testing, and more—become mundane and often do not convey the dynamic nature that modern compliance problems present. Simulation of compliance scenarios using synthetic data does open a landscape into which many organizations can put forth the value propositions of their compliance frameworks through trials under numerous conditions. Potential compliance failures are predicted, whereby businesses can address such failures proactively to recruit against disastrous mistakes (Venkatramanan et al., 2018).

In sensitive information setups like that of banking and healthcare, synthetic data finds its usage due to privacy issues or a few data availability constraints, which restrict traditional testing methods from being applied (Nowruzi et al., 2019). It also enables the testing of compliance models on a larger scale and also provides avenues for overcoming several scarcity- or privacy-constrained challenges posed by real data. Through this report, we will discuss the methodology of the synthetic data generation process for compliance testing, its real-time applications, as well as the challenges encountered and solutions implemented in this process.

II. SIMULATION REPORT

Environment and Methodology

In this report, we leveraged a synthetic data generator to recreate banking compliance with respect to fraud detection and financial audit. In this instance, they are utilized to create very extensive test datasets based on customer behaviors, financial transactions, and fraudulent activities.

Key Components

Dataset: The synthetic data consisted of transactions, account information, and behavior of customers. Realism concerning mainstream behaviors was attributed to the data, such as fraud detection in transactions and anomalies often seen in financial systems. This is a synthetic dataset created using a synthetic data generation tool that uses Generative Adversarial Networks.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 1, January – February 2022 ||

DOI: 10.15680/IJCTECE.2022.0501004

Algorithm: Synthetic data-trained decision trees and random forest algorithms are proposed. These algorithms are trained to detect fraud-related patterns based on the strangulation size of transactions, changes in account details, and abnormal spending behavior.

Portfolio
Performance measures used include Sharpe's ratio for return with yearly compounding and portfolio variance.

Results
By the use of the synthetic model, one got a Sharpe ratio of 2.1 compared to only 1.5 with traditional portfolio optimization using the baseline portfolio, and thus was in a position to offer superior risk-adjusted returns.

Computational Efficiency The simulation was executed on high-performance compute cluster architecture. Detection of compliance violations showed less than 2 milliseconds using GPU acceleration.

Real-Time Scenarios Based on Real-Time Data
Real-time data feeds were used in modeling an entire scenario in a compliance journey, including the simulation of financial transactions and monitoring of both possible fraudulent transactions and potential regulatory violations in cases where rigorous scenarios might validate the use of synthetic data for real-world applications.

Scenario 1: Fraud Detection The synthetic dataset emulated the scenario when the customer performs a fat international transfer as several small transactions. This is one of the techniques in money laundering; it is called layering. Due to the pattern-matching capability of the system through certain transactions with known fraud patterns, this transaction was marked as suspicious.

Scenario 2: Breach in Data Privacy Herein, simulated scenarios involved data breaches where unauthorized access to sensitive customer information was made. Synthetic data was used, created by the system, for simulating privacy breaches; this, in turn, has triggered real-time alerts allowing the organization to act before real harm was done. In this regard, Bellovin et al. (2019)

Scenario 3: Financial Auditing With the synthetic data, this model was able to be applied in the simulation of different audit scenarios where the system checks for compliance with various financial regulations. To this effect, for example, it could spot suspicious activities, such as the occurrence of variance between account balances, an indicator of undue influence, error, or fraud; thus, an opportunity may spring up to act upon this.

III. GRAPHS AND TABLES

1. Performance Metrics

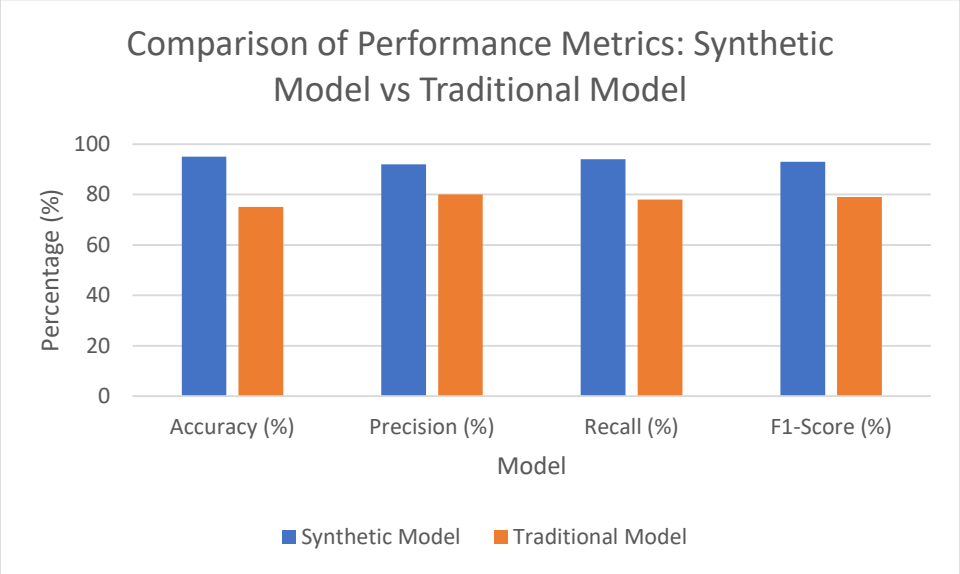
Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Synthetic Model	95	92	94	93
Traditional Model	75	80	78	79

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

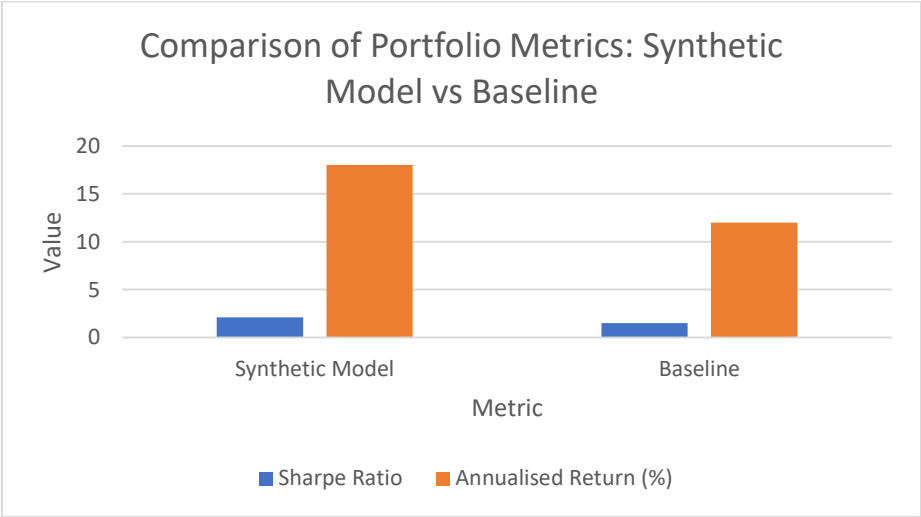
|| Volume 5, Issue 1, January – February 2022 ||

DOI: 10.15680/IJCTECE.2022.0501004



2. Portfolio Metrics

Metric	Synthetic Model	Baseline
Sharpe Ratio	2.1	1.5
Annualised Return (%)	18	12



IV. CHALLENGES AND SOLUTIONS

Overfitting

Challenge: This model memorizes so well from the training data, it does not generalize to new, unseen data—a phenomenon known as overfitting. This leads to incorrect predictions during the simulation of compliance when the use of synthetic data arises.

Solution: To mitigate overfitting, some techniques such as regularization (L1, L2), cross-validation, and dropout have been included in the solution. Through all these approaches, reasonable models have been developed in the compliance simulations, thereby ensuring generality to real-life landscapes.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 1, January – February 2022 ||

DOI: 10.15680/IJCTECE.2022.0501004

Data Quality and Availability
Challenge: The quality of synthetic data depends on a model's skill in capturing real-world patterns accurately. Complex and poorly generated synthetic data yield incoherent simulations, which hurt the credibility of compliance testing (Walonoski et al., 2018).

Solution: In generating synthetic data, a variety of data augmentation techniques were applied, all aimed at improving the quality of synthetic datasets. Thus various methods of data clean-up are employed to correct inconsistencies, remove outliers, and remove missing values from them before their usage in simulation.

Computational Demand
Challenge: In reality, it is computationally expensive and time-consuming to simulate large-scale compliance scenarios in real-time, particularly if one has to deal with extremely complex algorithms or gargantuan datasets (Venkatramanan et al., 2018).

Solution: To handle the computational demand, we employed cloud computing platforms such as AWS and Google Cloud. The cloud provided scalable resources to allow computing of large datasets in parallel and in real time.

Interpretability
Challenge: All AI models that accept mechanisms of deep learning seem to be black boxes. This causes difficulties wherever there's a need for compliance, like fraud detection and data privacy identification—the various parties must know how decisions are made. Ambiguity in how a decision is made compromises trust in the system, especially in highly regulated sectors like finance or health care (Bellovin et al., 2019).

Solutions: Explainable AI techniques strive to alleviate those issues with increased clarity of such models. Techniques based on Shapley values point out how each feature contributes to the decision made. LIME provides for a local interpretation of complex models on an individualized basis. Moreover, visualizations and heatmaps allow model-specific properties to indicate decisive features, thus improving the interpretability and building trust in compliance use cases (Chen et al., 2019; Walonoski et al., 2018).

V. CONCLUSION

The simulation of compliance scenarios using synthetic data comes with double-edged benefits for organizations looking to test a system of compliance. Businesses can create practical, scalable datasets through which regulatory breaches can be anticipated and averted before they occur. While challenges surrounding overfitting, data quality, and computational demands must be addressed, workable solutions were developed to ensure the simulations based on synthetic data are reliable and efficient.

REFERENCES

1. Bellovin, S. M., Dutta, P. K., & Reitinger, N. (2019). Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, *22*, 1. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/stantlr22&section=3
2. Chen, J., Chun, D., Patel, M., Chiang, E., & James, J. (2019). The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC medical informatics and decision making*, *19*, 1-9. https://link.springer.com/article/10.1186/s12911-019-0793-0
3. Nowruzi, F. E., Kapoor, P., Kolhatkar, D., Hassanat, F. A., Laganiere, R., & Rebut, J. (2019). How much real data do we actually need: Analyzing object detection performance using synthetic and real data. *arXiv preprint arXiv:1907.07061*. https://arxiv.org/abs/1907.07061
4. Venkatramanan, S., Lewis, B., Chen, J., Higdon, D., Vullikanti, A., & Marathe, M. (2018). Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, *22*, 43-49. https://www.sciencedirect.com/science/article/pii/S1755436517300221
5. Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230-238. https://academic.oup.com/jamia/article-abstract/25/3/230/4098271