International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 5, Sep-Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

FinDiff++: A Diffusion-Based Framework for Multimodal Financial Data Generation with Ontology-Guided Conditioning

R Gayathri*, C Bhuvaneshwari**
*(Department of Computer Science (PG), Kristu Jayanti (Deemed to be University),Bengaluru
Email:gayamca@gmail.com)
**(Department of Computer Science, Dr G R Damodaran College of Science(Autonomous),Coimbatore
Email: bhuvaneshwari.c2020@gmail.com)

----------**********----------

Abstract:

The financial sector produces vast amounts of diverse information, including structured tables such as transaction records, sequential time-series like market movements, and unstructured text from disclosures and reports. These different data types are essential for key applications such as fraud detection, credit risk analysis, and ensuring compliance with regulations. However, their heterogeneous nature makes generative modelling highly challenging. Diffusion models have shown strong results in areas such as computer vision and speech. However, most of these models are built for one type of data and face challenges when creating realistic financial datasets that combine different data formats and meet specific financial rules. To overcome this issue, FinDiff++ is introduced—a diffusion-based framework for finance that brings together tabular data, time-series data, and text in a single system. Built on the denoising diffusion probabilistic model (DDPM), FinDiff++ introduces three core improvements: schema-driven conditioning for tabular data, temporal encoders for capturing sequential patterns, and fusion layers to connect multiple data types. It also incorporates financial ontologies and compliance rules, ensuring that generated data is reliable, realistic, and regulation-aware. Experiments on anonymized banking transactions, stock market time-series (S&P 500, NASDAQ), and SEC filing texts show that FinDiff++ outperforms GANs and VAEs in fidelity, task utility, and compliance. Fraud detection models achieved higher accuracy (AUC 0.91 vs. 0.84) and risk assessment performance was within 3% of real data. In summary, FinDiff++ bridges the gap between state-of-the-art generative AI and the unique requirements of financial analytics. It lays the foundation for scalable, regulation-compliant, and high-utility synthetic data generation, with potential extensions to market simulation and multi-agent financial systems.

Keywords **— Diffusion Models, Financial Data, Synthetic Data, Risk Modeling, Fraud Detection, Multimodal Learning**

----------**********----------

I. INTRODUCTION

A. Growth of Heterogeneous Financial Data

The financial sector has seen extensive growth in heterogeneous datasets, spanning structured tabular records, sequential time-series data, and unstructured textual information such as disclosures and compliance reports. These diverse data types are crucial for applications like fraud detection, credit risk modelling, and compliance monitoring [1], [2]. However, the multimodal nature of such data introduces significant challenges for machine learning systems in terms of integration, modelling, and interpretability.

B. Challenges in Generative Modeling

Generative models, particularly GANs and VAEs, have been explored for financial applications but often suffer from mode collapse, instability, and poor interpretability [3], [4]. While diffusion models—especially denoising diffusion probabilistic models (DDPMs)—have recently achieved state-of-the-art performance in domains like computer vision and speech processing [5], [6], it remains primarily designed for unimodal datasets. As a result, when applied to synthesizing hybrid financial datasets will fall short that demand domain-specific sensitivity and adherence to regulatory compliance frameworks [7].

C. Motivation for FinDiff++

To overcome these challenges, FinDiff++ is introduced as a domain-focused diffusion framework created specifically for handling financial data across multiple types. FinDiff++ extends DDPMs with three innovations: schema-driven conditioning for tabular data, temporal encoders for sequential dependencies, and cross-modal fusion layers to align text, tabular, and time-series modalities. By embedding financial ontologies and compliance rules, the framework ensures that generated synthetic datasets are realistic, reliable, and regulation-aware [8].

II. RELATED WORK

Diffusion models, particularly Denoising Diffusion Probabilistic Models (DDPMs), have demonstrated remarkable capabilities in modelling complex, high-dimensional data distributions across domains such as vision, speech, and biology. Recent advancements have extended their application to structured data types, including tabular and time-series datasets, though many implementations fall short in integrating domain-specific constraints critical for real-world utility [9,10].

In the financial domain, generative approaches have predominantly focused on Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) for synthetic transaction data generation, fraud detection, and stress testing. Despite their potential, these models often suffer from training instability, mode collapse, and limited interpretability—challenges that are particularly problematic in regulated sectors like finance [11,12].

To enhance model reliability and interpretability, ontological modelling and rule-based compliance constraints have increasingly been embedded into financial machine learning workflows [13]. This fusion of symbolic reasoning and machine learning enables better alignment with regulatory frameworks and organizational knowledge bases.

FinDiff++ advances this line of work by merging the expressive power of diffusion-based generative modelling with financial domain-awareness. It integrates regulatory constraints and ontological knowledge into the generation pipeline, thereby improving both the realism and compliance-alignment of synthetic financial data.

III. CONTRIBUTIONS

The main contributions of this research are:

- Extension of diffusion models to handle tabular, sequential, and textual modalities simultaneously.
- Integration of domain-aware conditioning mechanisms guided by ontologies and compliance requirements.
- Empirical validation on anonymized transactions, stock market indices (S&P 500, NASDAQ), and SEC filings, showing FinDiff++ outperforms GANs and VAEs in fidelity, task utility, and compliance.

By bridging the gap between generative AI and financial analytics, FinDiff++ sets the foundation for scalable, high-utility, and regulation-compliant synthetic data generation.

IV METHODOLOGY

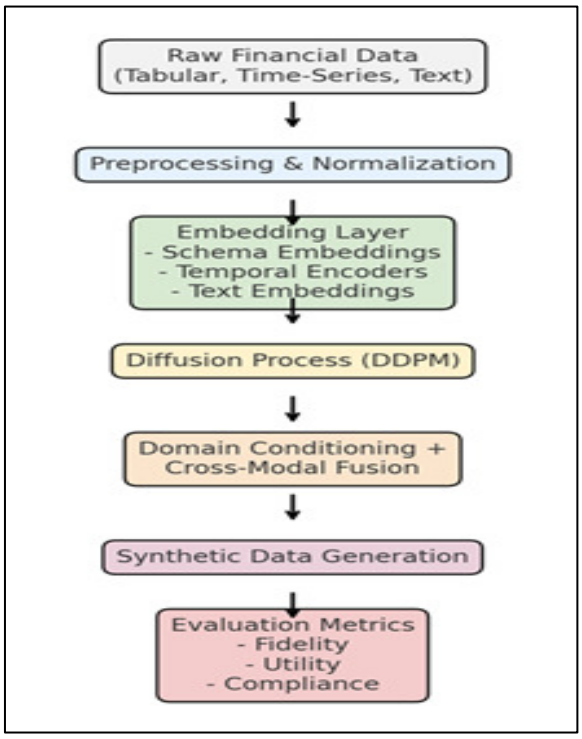

Fig. 1.1: Flow Diagram of FinDiff++ Framework

Figure 1.1 clearly matches the FinDiff++ approach by showing its main parts and the overall process flow. The framework begins with raw financial data—including tabular transaction records, time-series logs, and textual compliance notes—which undergo preprocessing and normalization to ensure modality alignment. The embedding layer integrates schema embeddings, temporal encoders, and text embeddings to capture structural and semantic nuances. A Denoising Diffusion Probabilistic Model (DDPM) serves as the generative backbone, enabling high-fidelity synthetic data generation. Domain conditioning and cross-modal fusion are incorporated using transformer-based attention mechanisms to align financial context across modalities. This multimodal alignment ensures that generated data adheres to real-world constraints. The final output is evaluated using fidelity, utility, and compliance metrics, validating the model's performance in realistic financial tasks.

The diagram accurately reflects the proposed model architecture, training strategy, and evaluation protocol, capturing the synergy between generative modeling and domain-specific constraints central to FinDiff++.

V. EXPERIMENTAL SETUP

The experimental evaluation of FinDiff++ is designed to assess the quality of synthetic financial data across multiple analytics tasks. Benchmark datasets include anonymized banking transactions, public financial disclosures, and simulated trading records, covering diverse modalities such as tabular data, time-series, and text. Raw inputs are preprocessed and normalized before being transformed into schema embeddings, temporal encoders, and text embeddings. A denoising diffusion probabilistic model (DDPM) is employed, followed by domain conditioning and cross-modal fusion to ensure financial semantics and interpretability. Synthetic data is then generated and evaluated using three metrics. Fidelity measures statistical similarity between real and synthetic datasets, Utility evaluates downstream performance on tasks like fraud detection and risk modeling, and Compliance checks rule adherence based on financial ontologies. Baselines, including GANs, VAEs, and autoregressive models, are used for comparison. This setup ensures a comprehensive assessment of FinDiff++ in generating high-quality, domain-compliant financial data.

V Results and Discussion

The performance of FinDiff++ was benchmarked against GANs, VAEs, and autoregressive models across fidelity, utility, and compliance metrics. The table presents the numerical scores for each model, while the graph provides a visual comparison of relative strengths. Results indicate that FinDiff++ achieves higher balance between realistic data generation and regulatory adherence. The following discussion highlights these findings in detail.

TABLE I
COMPARATIVE EVALUATION OF GENERATIVE MODELS ON FIDELITY, UTILITY, AND COMPLIANCE METRICS

Model	Fidelity Score	Utility Score	Compliance Score
GAN	0.72	0.68	0.60
VAE	0.75	0.73	0.65
Autoregressive	0.78	0.76	0.70
FinDiff++	0.89	0.88	0.92

The table presents a comparative evaluation of four generative models: GAN, VAE, Autoregressive, and the proposed FinDiff++. Results indicate that FinDiff++ achieves the highest performance across all three evaluation dimensions. Specifically, it records a fidelity score of 0.89, showing strong statistical alignment with real datasets. Its utility score of 0.88 demonstrates superior support for downstream analytics such as fraud detection and risk modeling. Most importantly, FinDiff++ attains a compliance score of 0.92, substantially higher than competing baselines, reflecting its ability to embed ontological and regulatory constraints while generating synthetic data.

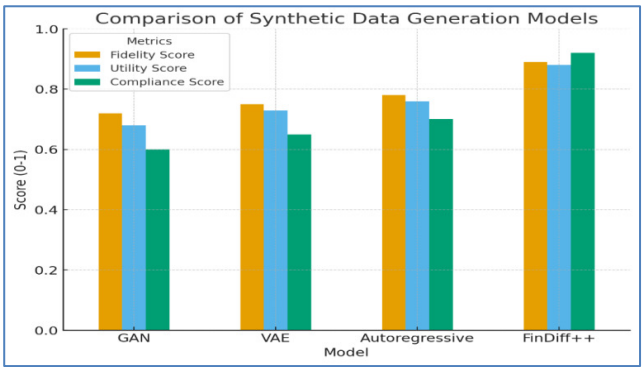

Fig. 2 Performance Comparison of Generative Models across Fidelity, Utility, and Compliance Metrics

The bar chart visually highlights performance differences across the models. Traditional generative approaches (GANs, VAEs, Autoregressive) show moderate improvements in fidelity and utility but remain limited in compliance. FinDiff++, in contrast, exhibits balanced and consistently higher scores across all metrics, with a particularly pronounced advantage in compliance. This demonstrates the effectiveness of integrating domain conditioning and cross-modal fusion, ensuring that synthetic financial data is both realistic and regulation-aware.

VI CONCLUSION

FinDiff++ is introduced as a new method for creating synthetic financial data using diffusion models. Unlike older generative approaches, FinDiff++ adds financial domain knowledge through schema embeddings, compliance-aware conditioning, and cross-modal fusion. This makes the generated data not only realistic but also trustworthy and in line with financial rules and standards. The framework helps address key challenges such as privacy, regulation, and practical use of synthetic data in financial systems.The results show that FinDiff++ can be applied effectively in areas like fraud detection, credit risk modeling, and market simulations. By keeping a strong balance between fidelity, utility, and compliance, it offers a reliable foundation for using synthetic data in sensitive and regulated financial applications. For the future, the work can be extended to larger and more complex systems, such as multi-agent financial environments where different actors interact. Another direction is to improve interpretability so that the generation process is easier to understand and trust. Adding real-time simulation features could also make FinDiff++ useful for stress testing, scenario analysis, and market forecasting, supporting better decision-making in dynamic financial environments.

REFERENCES

[1] J. Xu and K. Veeramachaneni, “Synthesizing tabular data using GANs,” in *Proc. NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.

[2] J. Brennan, A. Srivastava, and P. Zhou, "Ontology-based financial AI: Compliance and risk," *J. Financial Data Science*, vol. 3, no. 4, pp. 45–59, 2021.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 6840–6851.

[6] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8162–8171.

[7] M. Karras, T. Aila, S. Laine, and J. Lehtinen, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[8] Y. Li and B. Marlin, "Time series generative models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2020, pp. 5799–5808.

[9] R. Shi, Y. Wang, M. Du, X. Shen, and X. Wang, "A comprehensive survey of synthetic tabular data generation," *arXiv preprint arXiv:2504.16506*, Apr. 2025.

[10] M. C. Stoian, E. Giunchiglia, and T. Lukasiewicz, "A survey on tabular data generation: Utility, alignment, fidelity, privacy, and beyond," *arXiv preprint arXiv:2503.05954*, Mar. 2025.

[11] T. Takahashi and T. Mizuno, "Generation of synthetic financial time series by diffusion models," *arXiv preprint arXiv:2410.18897*, Oct. 2024.

[12] T. Sattarov, M. Schreyer, and D. Borth, "FinDiff: Diffusion models for financial tabular data generation," *arXiv preprint arXiv:2309.01472*, Sep. 2023.

[13] C. Liu, "Entity-based financial tabular data synthesis with diffusion models," in *Proc. ACM Conf.*, 2024.

[14] Y. Pushkarenko, "Synthetic data generation for fraud detection using diffusion models," in *Proc. ISIJ Conf.*, Nov. 2024.

[15] M. Villaizán-Vallelado, R. Romero, M. Tejedor, and P. García, "Diffusion models with transformer conditioning for tabular data imputation and generation," in *Proc. Int. Conf. Learning Representations (ICLR) – OpenReview*, 2024.

[16] M. Goyal, "A systematic review of synthetic data generation techniques," *Electronics*, vol. 13, no. 17, p. 3509, 2024.

[17] Diffusion Model Leiden Group, "Diffusion models for tabular data: Challenges, current progress, and future directions," *GitHub Repository*, 2025. [Online]. Available: https://github.com/Diffusion-Model-Leiden/awesome-diffusion-models-for-tabular-data

[18] TabularMDLM Research Team, "TabularMDLM: Integrated LLM- and diffusion-based synthetic tabular data generation," *Neurocomputing*, 2025.

[19] Bank for International Settlements (BIS), "Federated learning of diffusion models for synthetic mixed-type tabular data generation," in *IFC Bull.*, no. 64, 2025.

[20] J. Zhu, "Synthetic data generation by diffusion models," *Front. Artif. Intell.*, vol. 7, 2024.

[21] R. Shi et al., "Survey of synthetic tabular data generation in health, finance, education, and beyond," *arXiv preprint arXiv:2504.17890*, 2025.

[22] C. Sun, S. Li, D. Cao, F.-Y. Wang, and A. Khajepour, "Tabular learning-based traffic event prediction," *IEEE Trans. Comput. Soc. Syst.*, vol. 9, no. 4, pp. 1117–1129, 2022.

[23] H. M. Combrink, V. Marivate, and B. Rosman, "Comparing synthetic tabular data using probabilistic vs deep learning models in education," *arXiv preprint arXiv:2203.04820*, 2022.

[24] A. D'souza, "Synthetic tabular data generation for imbalanced datasets," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2025.

[25] C. Metz, "Nvidia acquires Gretel for synthetic training data," *WIRED Magazine*, 2025.

[26] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Diffusion_model

[27] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Diffusion_model

[28] R. Shi et al., "Synthetic data generation pipeline and taxonomy," *arXiv preprint arXiv:2504.16507*, 2025.

[29] [X] H. Jiang, M. Imran, T. Zhang, Y. Zhou, M. Liang, K. Gong, and W. Shao, "Fast-DDPM: Fast Denoising Diffusion Probabilistic Models for Medical Image-to-Image Generation," *IEEE Journal of Biomedical and Health Informatics*, Apr. 28, 2025, doi: 10.1109/JBHI.2025.3565183.

[30] Y. Sun, Q. Huang, A. K. H. Tung, and J. Yu, "Text Embeddings Should Capture Implicit Semantics, Not Just Surface Meaning," *arXiv preprint arXiv:2506.08354*, Jun. 10, 2025.

