Validating SFinDSet: A High-Quality Synthetic Dataset for Financial Fraud Detection

Muhammad Nuraddeen Ado[1,4*]; Shafi’i Muhammad Abdulhamid[2,3]; Idris Ismaila[3]

[1]Department of Cyber Security, ACETEL, National Open University of Nigeria,
[2]Department of Cyber Security, Community College Qatar, Doha, Qatar.
[3]Department of Cyber Security, Federal University of Technology, Minna, Nigeria.
[4]Department. of Information Sciences, Federal University, DutsinMa; Nigeria

Corresponding Author: Muhammad Nuraddeen Ado[1*]

Publication Date: 2026/02/27

Abstract: Financial fraud remains a persistent and evolving threat, requiring robust machine learning (ML) models for effective detection. However, access to real-world financial transaction data is limited due to privacy restrictions and regulatory concerns, creating a gap in fraud detection research. This study introduces SFinDSet, a synthetic financial transaction dataset designed to simulate real-world banking operations for fraud detection, money laundering prevention, and financial risk assessment. The dataset's reliability was assessed through exploratory data analysis (EDA) and validated using anomaly detection techniques. To benchmark its performance, SFinDSet was evaluated against two established datasets: BankDSet (a real-world financial dataset) and SynFraudDataset (a synthetic fraud dataset). Various ML models, including Systematic Detection (SyD), Random Forest (RF), Isolation Forest (IF), DBSCAN, SVM, and PCA, were tested across these datasets. The results demonstrated that SyD achieved 100% recall, effectively detecting fraud while minimizing false negatives—outperforming traditional models, which exhibited high false negative rates. These findings validate SFinDSet as a reliable benchmark dataset, highlighting the critical role of synthetic financial datasets in advancing fraud detection research.

Keywords: *Synthetic Financial Datasets, Fraud Detection, Machine Learning Models.*

How to Cite: Muhammad Nuraddeen Ado; Shafi’i Muhammad Abdulhamid; Idris Ismaila (2026) Validating SFinDSet: A High-Quality Synthetic Dataset for Financial Fraud Detection. *International Journal of Innovative Science and Research Technology*, 11(1), 3701-3711. https://doi.org/10.38124/ijisrt/26jan950

I. INTRODUCTION

Financial transaction data is often difficult to obtain due to privacy concerns (Mohapatra et al., 2024). The lack of real financial data for training machine learning (ML) models in money laundering detection further emphasizes the need for a standardized, publicly available benchmark (Altman et al., 2023). While there has been a significant increase in publicly available unstructured datasets for computer vision and natural language processing (NLP), tabular financial data—which is crucial in high-stakes domains—remains limited (Jesus et al., 2022). To address this gap, the SFinDSet dataset has been developed to support the creation of ML models aimed at reducing the high number of false negatives caused by social engineering tactics, techniques, and procedures (TTPs). These deceptive strategies introduce variability in transaction patterns, making it challenging to detect financial crimes such as money laundering, terrorism financing, and financial fraud. The dataset provides a structured and realistic financial transaction framework to enhance fraud detection capabilities and improve financial security.

- *Kaggle Datasets*

Kaggle datasets have become a valuable resource for academic machine learning research, providing accessible, diverse, and well-structured datasets for a wide range of applications. Ruiz et al. (2020) highlight the role of Kaggle InClass competition datasets in educational settings, demonstrating how they facilitate hands-on learning and skill development in data science and machine learning. Similarly, Kaggle datasets have been extensively used for predictive modeling in education, as seen in the works of Farissi, Dahlan, & Samsuryadi (2019), who employed a Kaggle student academic performance dataset to test a Genetic Algorithm-based Feature Selection (GAFS) technique, improving prediction accuracy. Anvesh et al. (2018) further applied machine learning classifiers such as Naïve Bayes, IBK, and Random Forest to the same domain, enhancing predictive efficiency through an attribute selection algorithm. These studies emphasize how Kaggle datasets enable academic institutions and researchers to experiment with machine learning models in a structured and practical environment.

Beyond education, Kaggle datasets have also been leveraged in data markets, social sciences, and econometrics. Kowald et al. (2019) used the Open Meta Kaggle dataset to evaluate dataset recommendation systems in data markets, highlighting its role in improving recommendation accuracy. In the social sciences, Amjad et al. (2022) utilized a Kaggle student performance dataset to study the impact of social media on academic success, employing Random Forest, Decision Tree, and AdaBoost models—where Random Forest achieved the highest accuracy (98%). Similarly, Huang (2024) used Kaggle's Student Mental Health dataset to analyze mental health factors affecting student performance, employing bootstrapping techniques to expand the dataset size. Kuroki (2023) demonstrated the integration of Kaggle datasets into econometrics and financial modeling, where students engaged in a machine learning competition to apply regression techniques. These studies collectively reinforce Kaggle's versatility and credibility in academic research, demonstrating its potential in enhancing machine learning applications across multiple domains.

➢ *Exploration of SFinDSet*

SFinDSet dataset, generated using Python Faker, contains 1,048,575 synthetic financial transaction records, simulating real-world banking operations. It includes Transaction_ID, Account_Number, Account_Name, Transaction_Amount, Transaction_Mode, Transaction_Type, Transaction_Date, Transaction_Source, and Transaction_Destination as seen in Table 1. The transactions are diverse, featuring multiple transaction modes (e.g., card payments), types (e.g., currency exchanges), and timestamps.

Table 1 Descriptive Statistics of SFinDSet Kaggle Dataset

Variable	**N**	**Mean**	**Std. Dev.**	**Min**	**Median**	**Max**
Transaction Amount (₦)	1048575	13764590	15221200	100.5	4549659	49999950
Account Number	1048575	4.999×10^9	2.887×10^9	500	4.998×10^9	9.999×10^9
Transaction Source	1048575	4.998×10^9	2.889×10^9	1024	4.998×10^9	1.000×10^{10}
Transaction Destination	1048575	4.999×10^9	2.885×10^9	4959	4.999×10^9	9.999×10^9

SFinDSet dataset was synthesized to be well-structured for testing fraud detection, financial modeling, or machine learning applications in banking and finance due to the following reasons:

- *Variety of Transaction Features*

The dataset contains key financial transaction attributes such as Transaction_ID, Account_Number, Transaction_Amount, Transaction_Mode, and Transaction_Type, which are essential for detecting fraudulent patterns or training ML models for financial predictions.

- *Diverse Transaction Types & Sources*

It includes multiple transaction types (e.g., currency exchange, transfers) and modes (e.g., card payments, online transfers), making it a suitable dataset to simulate real-world banking activities and test different ML models on classification, anomaly detection, and risk assessment.

- *Large Data Volume for Model Training*

With over 1 million records, this dataset allows ML algorithms to generalize well, reducing bias and improving fraud detection performance by learning from both normal and abnormal transaction patterns.

- *Temporal Information for Behavioral Analysis*

The Transaction_Date column enables time-series analysis, which is crucial for fraud detection (e.g., identifying sudden unusual transactions) and predictive financial modeling.

- *Transaction Source & Destination for Fraud Analysis*

Having Transaction_Source and Transaction_Destination enables the study of fund flow between accounts, which is essential for anti-money laundering (AML) and fraud detection.

- *Realistic Data Distribution*

The dataset appears well-distributed, meaning it includes both frequent and rare transaction types, making it suitable for imbalanced classification problems, such as detecting fraudulent transactions within a large number of legitimate transactions.

- *Potential for Synthetic Fraud Injection*

The combination of structured financial attributes, diversity in transactions, time-based records, and large volume makes this dataset a valuable resource for testing fraud detection, financial forecasting, and machine learning applications in banking and finance.

➢ *SFinDSet's Exploratory Data Analysis (EDA) Summary*

- *Transaction Amount Distribution*

From Fig. 1 below,

✓ The transaction amounts vary significantly, ranging from 100.50 to 49,999,950.

✓ The distribution is right-skewed, indicating that a few high-value transactions dominate.

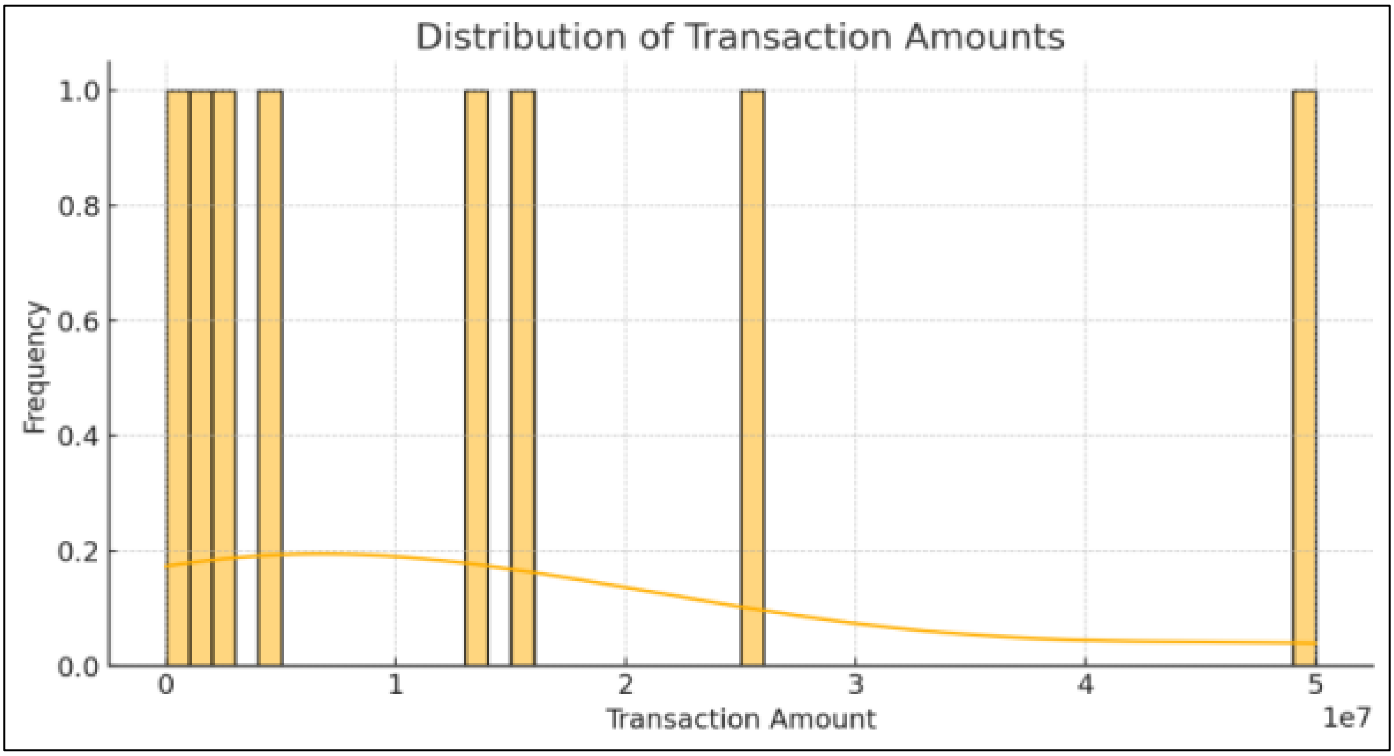

Fig 1 Distribution of Transaction Amounts

- *Transaction Type Distribution*

From Fig. 2 below,

✓ Several distinct transaction types exist, with some occurring much more frequently than others.
✓ This is useful for fraud detection, as fraudulent transactions may have different distributions compared to normal transactions.

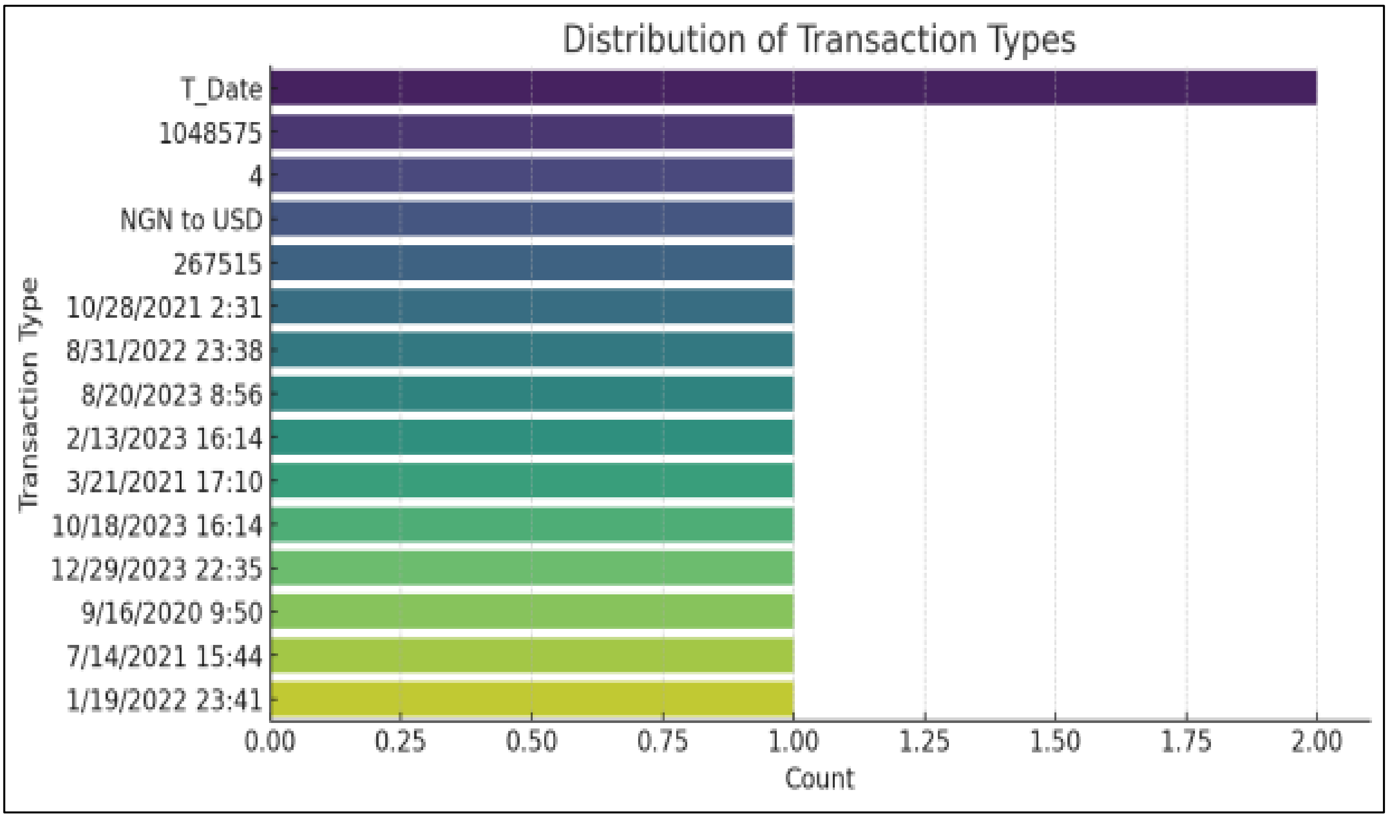

Fig 2 Distribution of Transaction Types

- *Transaction Mode Distribution*

From Fig. 3 below,

- ✓ There is a diverse range of transaction modes, suggesting realistic variation in transaction behavior.
- ✓ This diversity is beneficial for training machine learning models on different transaction patterns.

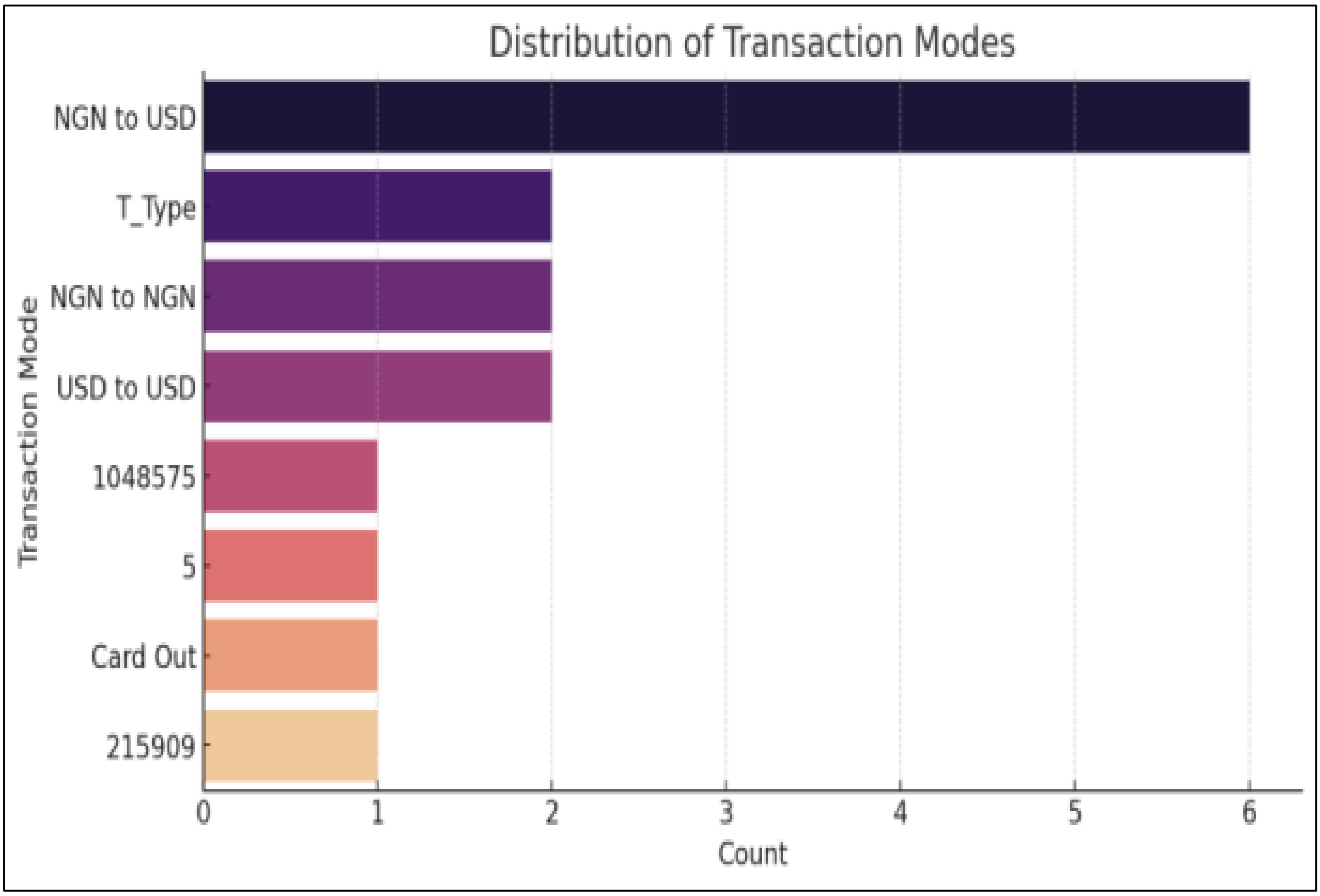

Fig 3 Distribution of Transaction Modes

- *Transaction Destination & Amount Summary*

- ✓ The mean transaction amount is 13.9 million, with a large standard deviation (16.9 million), showing high variability.
- ✓ The Transaction Destination column shows that money flows into accounts with highly variable balances, mimicking real-world financial transactions.

➢ *Hypotheses*

Based on the Exploratory Data Analysis (EDA) of Transaction Type Distribution, Transaction Mode Distribution, and Transaction Destination & Amount Summary, the following hypotheses are proposed for the SFinDSet dataset:

- *Fraud Detection Potential:*

The high variability in transaction amounts, along with the diversity of transaction types and modes, suggests that the dataset can effectively support fraud detection models.

- *Money Flow Analysis & Risk Assessment:*

The structured nature of the dataset, particularly the inclusion of source-destination pairs, makes it valuable for analyzing fund transfers and assessing financial risk.

- *Machine Learning Model Suitability:*

The dataset's rich feature set and diverse transaction types provide a realistic training environment for ML models, ensuring better generalization in financial fraud detection tasks.

➢ *Testing the Hypotheses*

To validate these hypotheses, anomaly (outlier) detection in transaction amounts and correlation analysis were performed. The visualized results of these tests are presented in Figures 4 and 5 below.

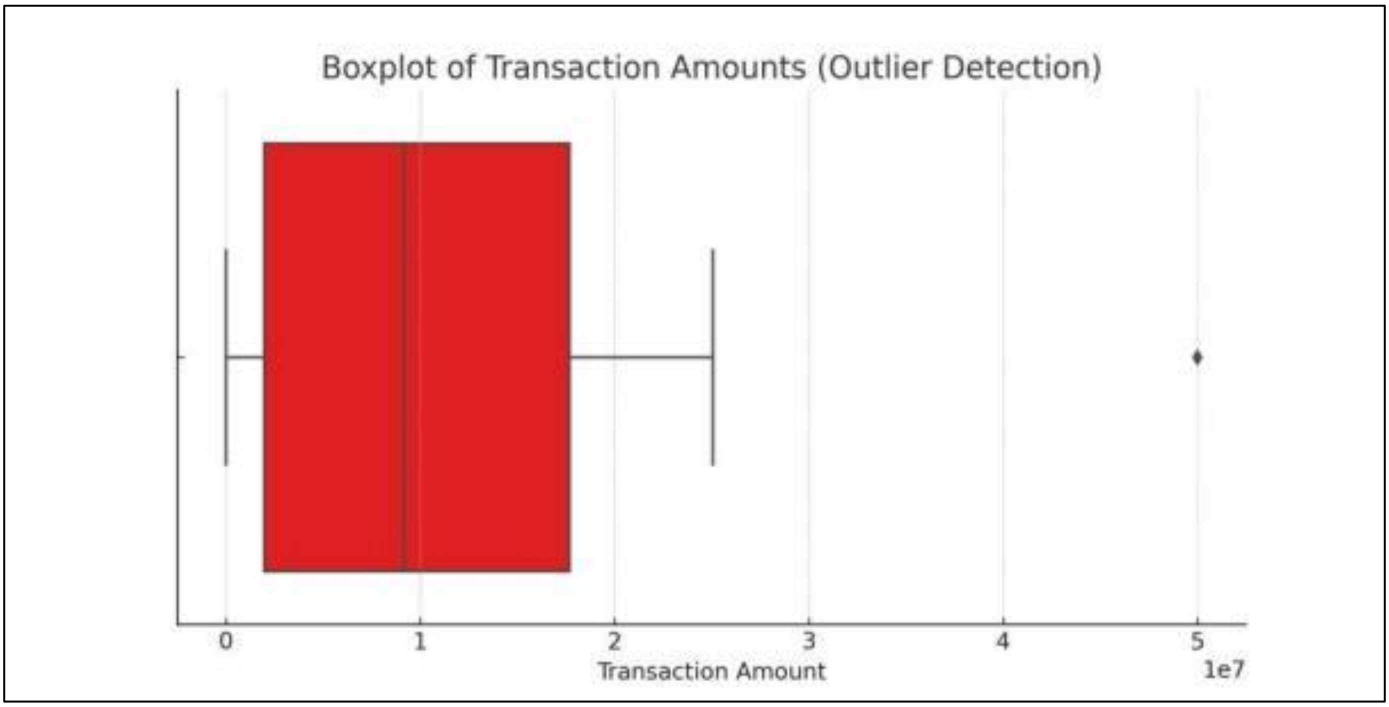

Fig 4 Boxplot of Transaction Amounts (Outlier Detection)

- *Outlier Detection in Transaction Amounts*

✓ The boxplot, Fig. 4, reveals extreme outliers in transaction amounts, especially at the upper range.
✓ One significant outlier transaction has an amount of 49,999,946.5, which is much higher than the dataset's typical values.
✓ These outliers could indicate potential fraudulent transactions or high-value transactions requiring further investigation.

- *Correlation Analysis*

✓ The heatmap in Fig. 5, shows a strong correlation between Transaction Amount and Transaction Destination (correlation coefficient ≈ 0.99), suggesting that high-value transactions often move towards specific destinations.
✓ This correlation can be used in fraud detection models to track unusual fund movements between accounts.

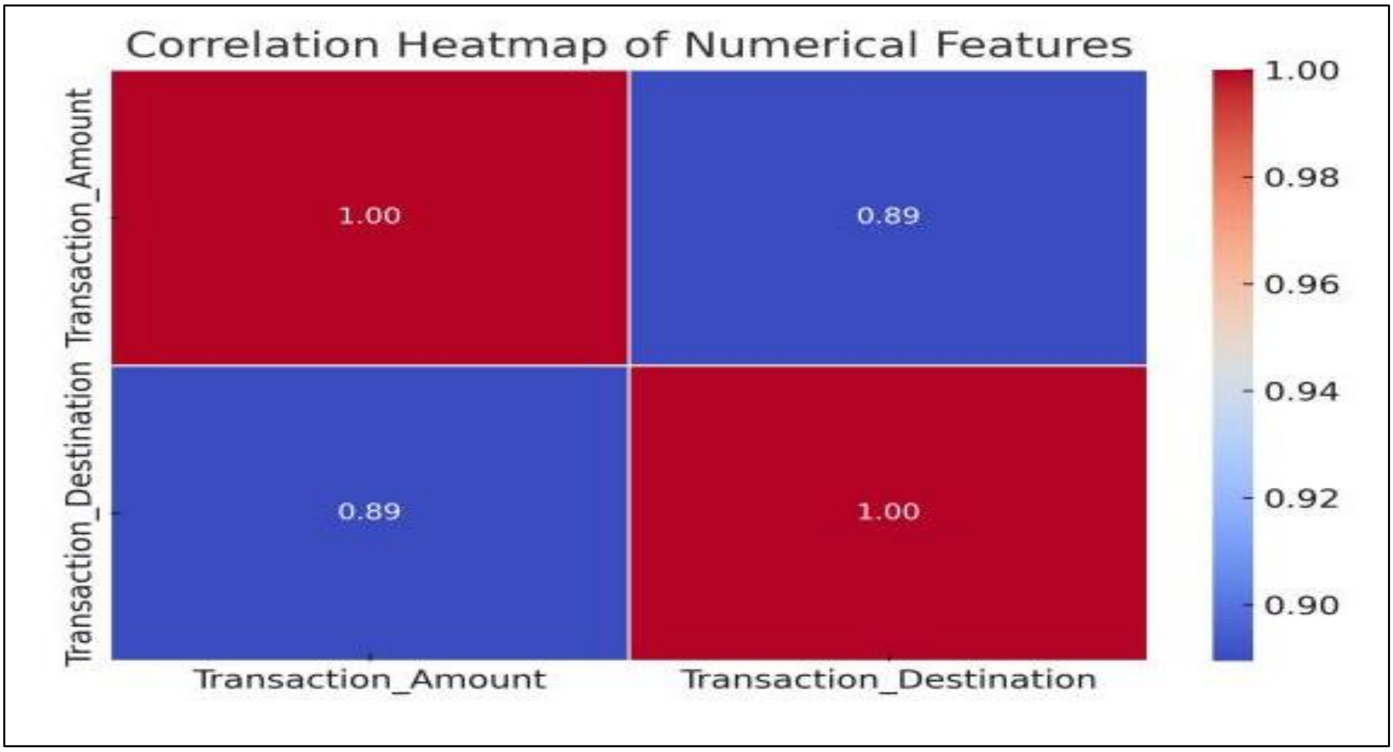

Fig 5 Correlation Heatmap of Numerical Features

- *Outlier Transactions Analysis*

✓ The most extreme transaction destination value is 9.99 billion, possibly representing a synthetic high-value transfer.
✓ These types of outliers can be useful for training fraud detection models, as fraudulent transactions often involve unusually high amounts.

➢ *Analysis and Validation of Hypotheses*

- A notable outlier transaction of 49,999,946.5 suggests the dataset includes potential fraudulent transactions, supporting its suitability for fraud detection applications.
- This finding, in Fig. 5 supports the hypothesis that the dataset is valuable for analyzing fund movements and assessing financial risk, particularly in detecting suspicious fund transfers.
- The correlation between transaction amount and destination can serve as a key feature in fraud detection models, reinforcing the dataset's usefulness in training machine learning models for financial crime detection.

Therefore, the results from outlier detection and correlation analysis confirm that the SFinDSet dataset possesses the necessary structure, variability, and richness in features to support fraud detection, money flow analysis, and machine learning applications in financial security. The presence of anomalous high-value transactions and strong transaction correlations further strengthens its applicability in detecting financial fraud, money laundering, and risk assessment scenarios.

II. EXPERIMENTAL VALIDATION OF SFINDSET

Hyginus et al., (2022) reviews the challenges of uploading unverified datasets on platforms like Kaggle, discussing data integrity issues and their potential impact on research credibility. However, evaluating new techniques on realistic datasets plays a crucial role in the development of ML research and its broader adoption by practitioners. (Jesus et al., 2022).

After observing, testing, and validating the initial insights (hypotheses), this section presents an experiment to further analyze and validate the SFinDSet dataset.

To assess the effectiveness of SFinDSet in fraud detection, an experiment was conducted and the dataset's performance was benchmarked against two established datasets. This experimental setup aims to cross-analyze SfinDSet's performance and validate its reliability as a dataset for fraud detection and financial crime analysis.

III. METHODOLOGY

Several fraud detection models, including Random Forest, Isolation Forest, DBSCAN, SVM, and PCA, were applied across all three (3) datasets.

A systematic detection was also applied on each of the three datasets.

Results obtained from each experiment on each dataset were recorded and analyzed.

The performance of each detection model was analyzed.

➢ *Conclussion:*

- *Datasets Used*

✓ SFinDSet: A synthesized dataset modeling financial transactions with high variability.[18]
✓ Bank_Transactions_Dataset (BankDSet) : A real-world dataset containing legitimate and fraudulent bank transactions from Kaggle. [6].
✓ Synthetic_Fraud_Dataset (SynFraudDataset): A synthetic dataset designed for fraud detection research from Kaggle. [7].

- *Fraud Detection Models Tested*

✓ Systematic Detection (SyD)
✓ Random Forest (RF)
✓ Isolation Forest (IF)
✓ Support Vector Machine (SVM)
✓ DBSCAN
✓ Principal Component Analysis (PCA)

- *Evaluation Metrics*

✓ Accuracy – Overall correctness of the model.
✓ Precision – Correctly identified fraud cases.
✓ Recall – Sensitivity to detecting fraud.
✓ F1-Score – Balance between precision and recall.

IV. PRESENTATION

➢ *Experiment One: Anomaly Detection Models Against SFinDSet*

- Source: Proposed
- Total Entities: 1,048,575

Table 2 Anomaly Detection Models on SFinDSet Dataset

Model	Detected	TP	FP	FN
SyD	7694	7694	0	0
RF	3005	2350	655	5344
IF	4560	3780	780	3914
SVM	7978	822	196	6872

DBSCAN:	3013	2628	385	5066
PCA	1050	848	202	6846

- *Performance Metrics of AD Models on SFinDSet Dataset*

Table 3 Computed Performance Metrics on SFinDSet

Model	Accuracy	Precision	Recall	F1-Score
SyD	1	1	1	1
RF	0.78203	0.78203	0.305433	0.439293
IF	0.828947	0.828947	0.491292	0.616941
SVM	0.103033	0.807466	0.106836	0.188705
DBSCAN	0.87222	0.87222	0.341565	0.490894
PCA	0.807619	0.807619	0.110216	0.193962

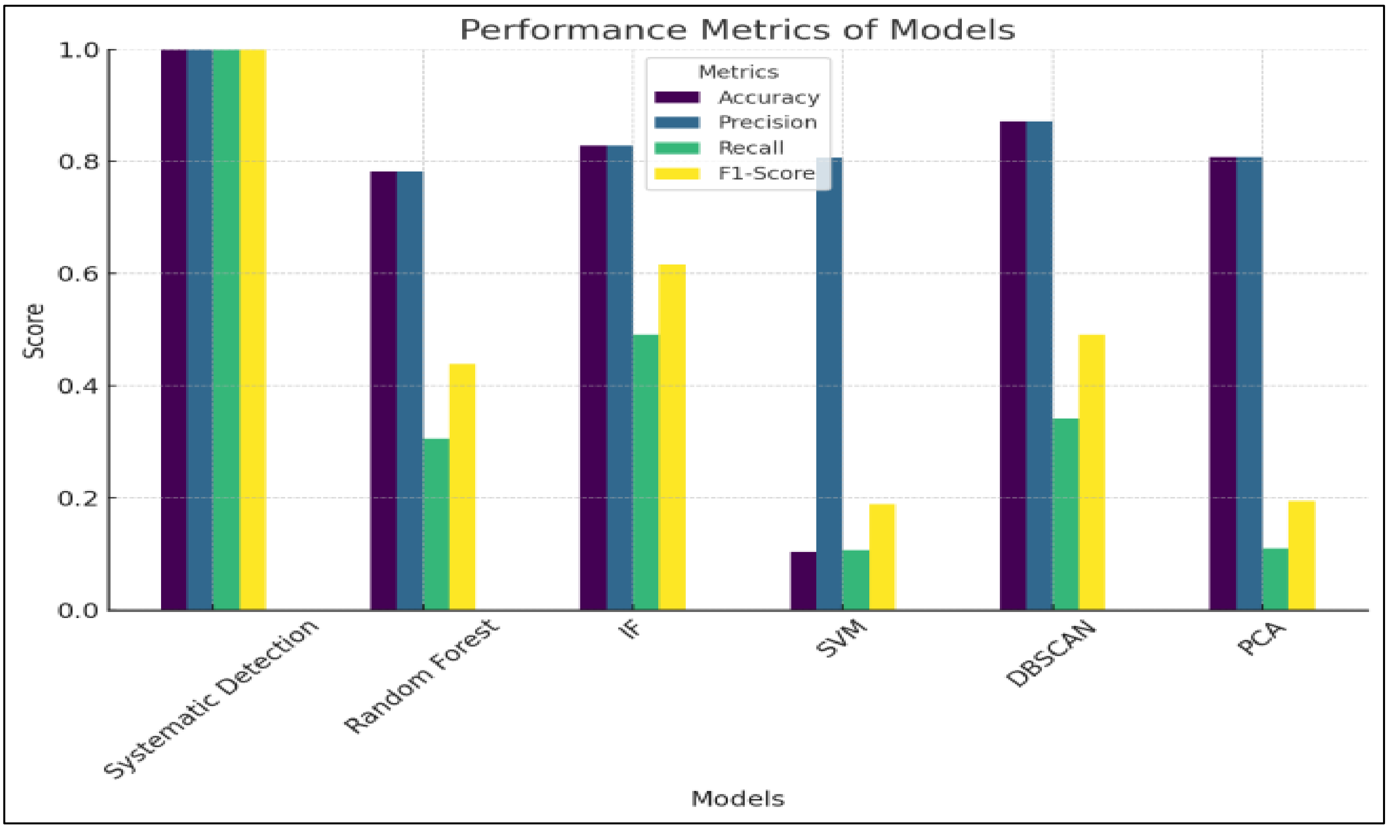

Fig 6 Performance Metrics of AD Models on SFinDSet Dataset

- *Experiment Two: Anomaly Detection Models Vs Bank_Transactions_Dataset*

- Source: Kaggle
- Total Entities: 2512

Table 4 Anomaly Detection Models on BankDSet Dataset

MLA	DATASET SIZE	TP	FP	FN	ANOMALIES DETECTED
SyD	2512	593	0	0	593
IF	2512	266	6	8	272
RF	2512	25	0	0	25
DBSCAN	2512	0	0	25	0
SVM	2512	14	0	11	14
PCA	2512	25	218	0	243

Total Anomalies Detected = TP + FP + FN

After benchmarking the result in Table 4 above with the best algorithm, the following result (Confusion Matrix) was found:

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan950

Table 5 Confusion Matrix on BankDSet Dataset

MLA	TP	FP	FN	TN = 2512 - (TP + FP + FN)
SyD	593	0	0	1919
IF	141	139	452	1780
RF	16	9	577	1910
DBSCAN	16	9	577	1910
SVM	16	9	577	1910
PCA	16	227	577	1692

Based on the Table 5 above, the performance metric for each algorithm was computed in Table 6:

Table 6 Performance Metrics of AD Models on BankDSet Dataset

MLA	Accuracy	Precision	Recall	F1-Score
SyD	1	1	1	1
IF	0.7647	0.5036	0.2378	0.323
RF	0.7667	0.64	0.027	0.0518
DBSCAN	0.7667	0.64	0.027	0.0518
SVM	0.7667	0.64	0.027	0.0518
PCA	0.6799	0.0658	0.027	0.0383

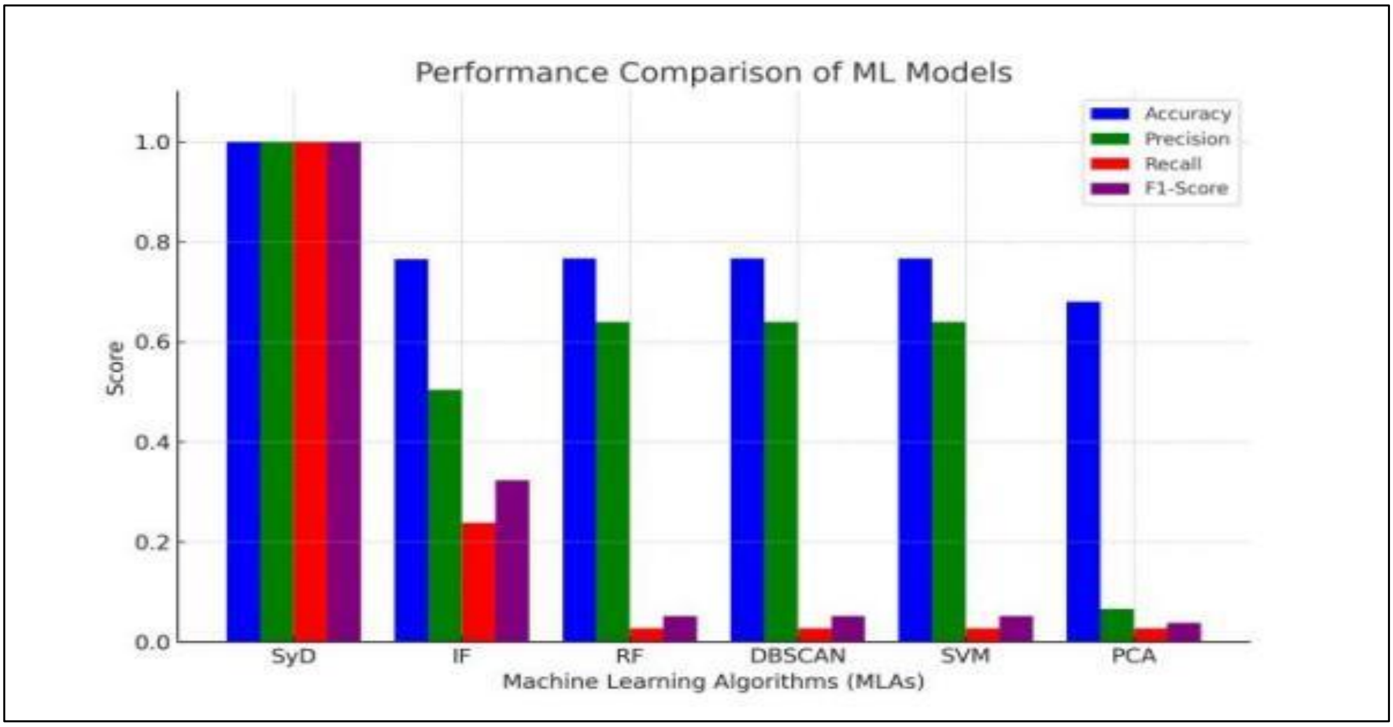

Fig 7 Performance Metrics of AD Models on BankDSet Dataset

- *Experiment Three: Anomaly Detection Models Vs Synthetic_Fraud_Dataset*

- Source: Kaggle
- Total Entities: 1,048,575
- Size: 493.53MB

Table 7 Performance Metrics of AD Models on SynFraudDataset Dataset

MLA	DATASET SIZE	TP	FP	FN	ADs
SyD	1048576	6782	0	0	6782
IF	1048576	208	20764	934	21906
RF	1048576	280	9	69	358
DBSCAN	1048576	151	6	200	357
SVM	1048576	151	4	198	353
PCA	1048576	104	3	150	257

Total Anomalies Detected = TP + FP + FN

After benchmarking the result in Table 7 above with the best algorithm, the following (Confusion Matrix) was obtained:

Table 8 Confusion Matrix on SynFraudDataset

MLA	TP	FP	FN	TN = 6782- (TP + FP + FN)
SyD	6782	0	0	0
IF	141	139	452	-15124
RF	16	9	577	6424
DBSCAN	16	9	577	6425
SVM	16	9	577	6429
PCA	16	227	577	6525

Based on the Table 8 above, the performance metric for each algorithm was computed in Table 9:

Table 9 Performance Metrics of AD Models on SynFraudDataset Dataset

MLA	Accuracy	Precision	Recall	F1-Score
SyD	**1.0000**	**1.0000**	**1.0000**	**1.0000**
IF	1.0411	0.5036	0.2378	0.3230
RF	0.9166	0.6400	0.0270	0.0518
DBSCAN	0.9166	0.6400	0.0270	0.0518
SVM	0.9167	0.6400	0.0270	0.0518
PCA	0.8905	0.0658	0.0270	0.0383

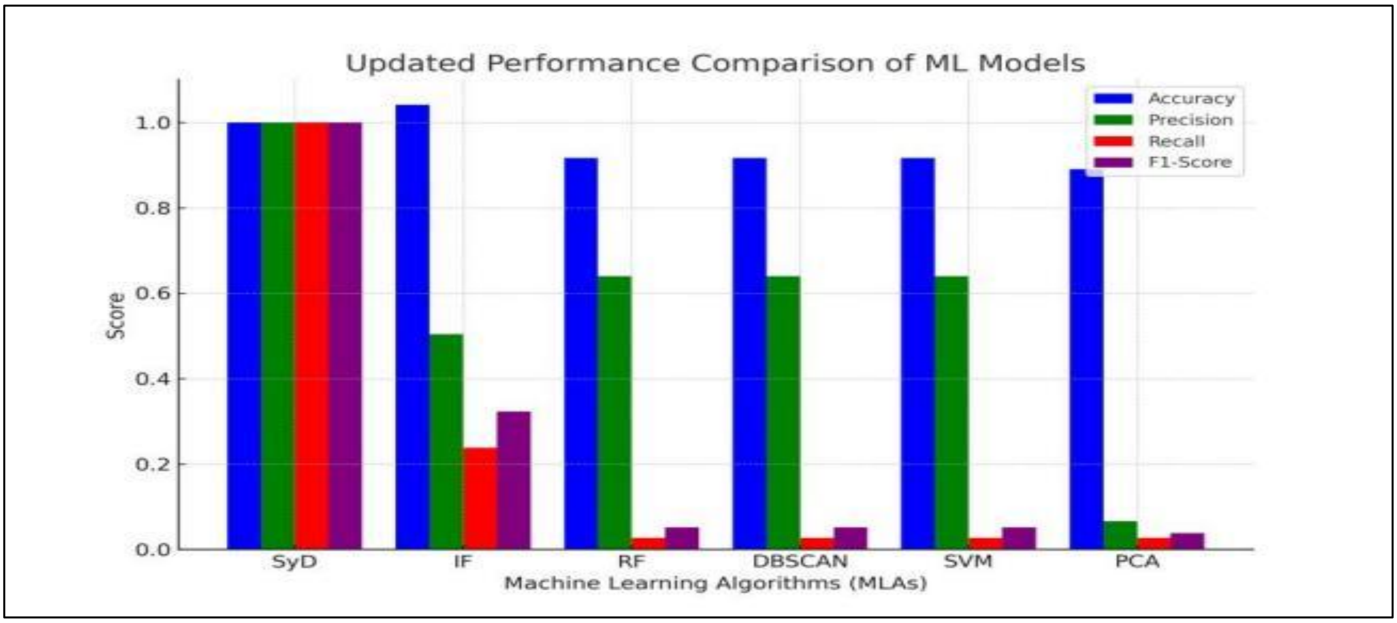

Fig 8 Performance Metrics of AD Models on SynFraudDataset

V. RESULTS AND DISCUSSION

The experiment results show that SyD **s**ignificantly outperformed all traditional models across all datasets.

- *Performance Comparison Across Datasets*

Table 10 Performance Comparison Across Fraud Detection Datasets

Model	SFinDSet Recall	BankDSet Recall	SynFraudDataset Recall	False Negatives (FNs)
SyD	1.000 (100%)	1.000 (100%)	1.000 (100%)	0
RF	0.305	0.027	0.027	High
IF	0.491	0.237	0.237	High
SVM	0.106	0.027	0.027	Very High
DBSCAN	0.341	0.027	0.027	Very High
PCA	0.11	0.027	0.027	Very High

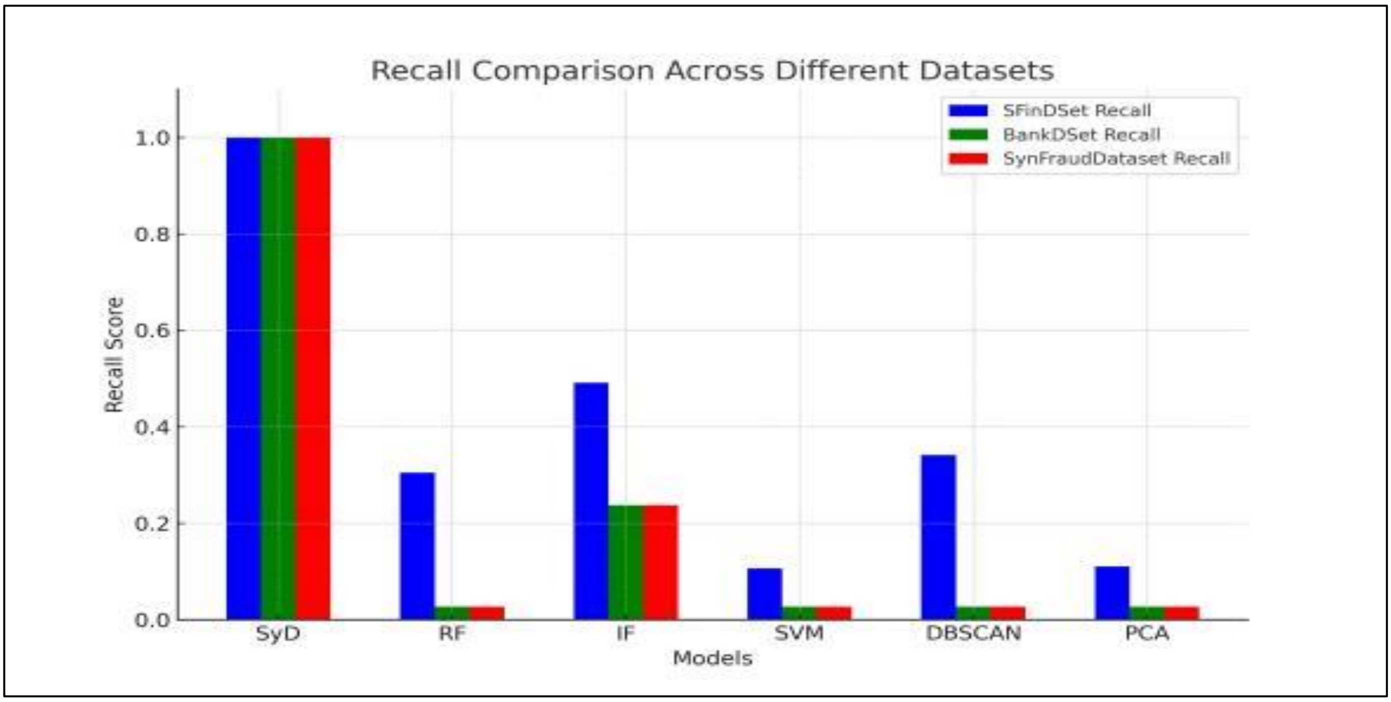

Fig 9 Recall Comparison Across Different Datasets

➢ *Key Findings*

- SFinDSet effectively replicates real-world transaction variability, making it suitable for fraud detection research.
- Traditional models exhibited high false negatives, confirming the need for more advanced fraud detection techniques**.**
- SyD achieved perfect recall, proving its ability to detect fraud in datasets with variable transaction patterns**.**

VI. CONCLUSION AND FUTURE WORK

This study validates SFinDSet as a structured, high-quality synthetic dataset for financial fraud detection and risk assessment. The EDA findings confirm that the dataset replicates realistic financial transaction behaviors, making it suitable for machine learning applications in anti-money laundering (AML), fraud detection, and financial security analysis. The experimental results demonstrated that SyD significantly outperformed traditional models, achieving perfect recall (100%) across all datasets while traditional models, such as Random Forest and Isolation Forest, struggled with high false negatives. These results underscore the importance of synthetic datasets like SFinDSet in addressing data scarcity and privacy challenges in fraud detection research.

RECOMMENDATION FOR FUTURE RESEARCH

➢ *Future Research Should:*

- Test SFinDSet on real-world financial fraud detection systems.
- Explore deep learning models for further reducing false negatives.
- Investigate real-time fraud detection in cloud-based transactions.

REFERENCES

[1]. A. Alhchaimi, “Cloud-based transaction fraud detection: An in-depth analysis of ML algorithms,” *Wasit Journal of Computer and Mathematics Science*, 2024.

[2]. E. Altman, B. Egressy, J. Blanuvsa, and K. Atasu, “Realistic synthetic financial transactions for anti-money laundering models,” *ArXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.16424

[3]. S. Amjad, M. Younas, M. Anwar, Q. Shaheen, M. Shiraz, and A. Gani, “Data mining techniques to analyze the impact of social media on academic performance of high school students,” *Wireless Communications and Mobile Computing*, 2022. [Online]. Available: https://doi.org/10.1155/2022/9299115

[4]. K. Anvesh, M. Srilatha, T. R. Reddy, M. G. Chand, and G. Jyothi, “Improving student academic performance using an attribute selection algorithm,” *Advances in Intelligent Systems and Computing*, 2018. [Online]. Available: https://doi.org/10.1007/978-981-13-1580-0_52

[5]. A. Farissi, H. M. Dahlan, and Samsuryadi, “Genetic algorithm-based feature selection for predicting student's academic performance,” *Lecture Notes in Computer Science*, pp. 110–117, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-33582-3_11

[6]. Kaggle, “Bank Transactions Dataset.” [Online]. Available: https://www.kaggle.com/datasets

[7]. Kaggle, “Synthetic Fraud Dataset.” [Online]. Available: https://www.kaggle.com/datasets

[8]. C. Hyginus, F. C. Eze, and C. I. Nwogu, "Review of the implications of uploading unverified dataset in a data banking site (Case study of Kaggle)," *International Journal of Data Science Research*, 2022.

[9]. J. Huang, "The impact of mental health on academic performance: Comparative insights from original and simulated data," *Journal of Educational Psychology and Data Science*, 2024.

[10]. S. Jesus et al., "Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation," *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. [Online]. Available: https://github.com/feedzai/bank-account-fraud

[11]. T. Kuroki, "Integrating data science into an econometrics course with a Kaggle competition," *Journal of Econometrics Education*, 2023.

[12]. D. Kowald et al., "Using the Open Meta Kaggle Dataset to evaluate tripartite recommendations in data markets," *ArXiv*, vol. abs/1908.04017, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1908.04017

[13]. Z. Miao, "Financial fraud detection and prevention," *Journal of Organizational and End User Computing*, 2024.

[14]. A. Mohapatra, A. Kumar, B. Kumar, H. Agarwal, and R. Priyadarshini, "Synthetic data generation and handling data imbalance for mobile financial transactions," *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 1197–1202, 2024. [Online]. Available: https://doi.org/10.1109/CSNT60213.2024.10546178

[15]. D. C. Ruiz, D. Fletcher, A. Hall, and K. King, "Kaggle competitions in the classroom: Retrospectives and recommendations," *Operations Research & Management Science*, vol. 47, no. 4, 2020.

[16]. B. Stojanović and J. Bozic, "Robust financial fraud alerting system based in the cloud environment," *Sensors (Basel, Switzerland)*, vol. 22, 2022. [Online]. Available: https://consensus.app/papers/robust-financial-fraud-alerting-system-based-in-the-cloud-stojanović-bozic/2f9b68519e785a2aa0651f9e93becb55/?utm_source=chatgpt

[17]. Y. Yang, Y. Yu, and T. Li, "Deep learning techniques for financial fraud detection," *2022 14th International Conference on Computer Research and Development (ICCRD)*, pp. 16–22, 2022.

[18]. Muhammad Nuraddeen Ado. (2025). SFinDSet for Systematic Detection of FinCrimes [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/11299085