International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

# A Data-Driven Approach to Digital Payment Fraud Detection Using Logistic Regression and Python

Sneha V Prabhu[1], Prathyusha Pasupuleti[2]
[1]Student, BNM Institute of Technology
[2]Assistant Professor, BNM Institute of Technology

***Abstract*— The rapid growth of digital payment systems has made online transactions more convenient, but it has also increased the risk of fraudulent activities. This study focuses on detecting fraudulent digital payment transactions using machine learning techniques. Logistic Regression and Random Forest models were developed and compared to determine which model performs better in identifying fraudulent transactions. Feature Importance Analysis was carried out to identify the factors that have the greatest influence on fraud, and transactions were classified into different risk categories based on their predicted probabilities. Structural Equation Modeling (SEM) was also used to examine the relationships between the selected variables and fraud occurrence. The results showed that Logistic Regression was more effective in detecting fraudulent transactions, while Random Forest achieved higher overall accuracy. Among the selected variables, Device Location, Previous Failed Attempts, and Login Attempts in the Last 24 Hours were found to have the greatest influence on fraud prediction. The findings of this study can help financial institutions improve fraud detection systems and strengthen risk management practices in digital payment environments.**

***Keywords*— Digital Payment Fraud, Feature Importance Analysis, Logistic Regression, Machine Learning, Random Forest, Structural Equation Modeling (SEM).**

## I. Introduction

The financial services industry has changed significantly with the growth of digital technology, making digital payment systems a part of everyday life. People now use mobile banking, UPI, digital wallets, and online payment platforms for quick and convenient transactions. However, the increasing use of digital payments has also led to a rise in fraudulent activities, making fraud detection a major concern for financial institutions. Traditional fraud detection methods are often unable to identify new and complex fraud patterns, creating the need for more advanced techniques. Machine learning has become an effective solution as it can analyze transaction data, identify suspicious patterns, and improve fraud detection accuracy. This study focuses on detecting fraudulent digital payment transactions using Logistic Regression and Random Forest models.

It also includes Feature Importance Analysis, Risk Categorization, and Structural Equation Modeling (SEM) to understand the factors influencing fraud. The study compares the performance of the selected models and identifies the key factors associated with fraudulent transactions. The findings are expected to help financial institutions strengthen their fraud detection systems, reduce financial losses, and improve the security of digital payment services.

## II. Literature Review And Research Gap

### A. Literature Review

Recent studies have shown that machine learning has become an effective approach for detecting fraudulent transactions in digital payment systems. Researchers have compared various algorithms such as Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, and Support Vector Machine, with many studies reporting that Random Forest and ensemble techniques provide better prediction accuracy and improve fraud detection performance (Manorom et al., 2024; Kumar et al., 2021; Salunke et al., 2025). Several studies have also highlighted the importance of data preprocessing techniques, such as SMOTE, for handling imbalanced datasets and improving model performance (Archana et al., 2025; Shahri & Abdullah, 2025). In addition, Feature Importance Analysis and predictive analytics have been widely used to identify the key factors influencing fraudulent transactions (Hu, 2025; Wang et al., 2025). Reports published by financial institutions and industry organizations further emphasize that the rapid growth of digital payments has increased fraud risks, making the adoption of Artificial Intelligence and machine learning essential for real-time fraud detection and risk management (Reserve Bank of India, 2025; National Payments Corporation of India, 2024; Deloitte, 2024; EY, 2024; McKinsey & Company, 2024). Overall, the existing literature indicates that advanced machine learning techniques play a vital role in improving fraud detection, reducing financial losses, and enhancing the security of digital payment systems (Naqvi, 2025; George et al., 2025; PwC, 2024; World Bank, 2024).

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

### B. Research Gap

Although many studies have used machine learning techniques to detect fraudulent transactions, most of them focus only on comparing the performance of different prediction models. Very few studies combine multiple analytical techniques such as Feature Importance Analysis, Risk Categorization, and Structural Equation Modeling (SEM) in a single framework. In addition, limited research has examined the relationships among fraud-related factors or categorized transactions into different risk levels to support better decision-making. Therefore, this study attempts to bridge these gaps by integrating multiple analytical methods to provide a more comprehensive understanding of digital payment fraud detection.

## III. Research Methodology

### A. Statement of the Problem

The increasing use of digital payment systems has made financial transactions faster, easier, and more convenient. However, the rapid growth of online transactions has also led to a rise in fraudulent activities, creating challenges for financial institutions. Traditional fraud detection methods are often unable to identify complex and evolving fraud patterns effectively. This study uses machine learning techniques, including Logistic Regression and Random Forest, along with Feature Importance Analysis, Risk Categorization, and Structural Equation Modeling (SEM), to improve fraud detection. The study aims to develop a comprehensive framework that helps identify fraudulent transactions and supports better fraud prevention and risk management.

### B. Objectives

The primary objective of this study is to develop machine learning classification models for predicting fraudulent digital payment transactions. The study also aims to analyze and preprocess digital payment transaction data to improve the quality and accuracy of fraud detection. In addition, it seeks to identify the key factors associated with fraudulent transactions and categorize transactions into different risk levels based on their predicted probability of fraud. Further, the study aims to examine the relationships among fraud-related factors using Structural Equation Modeling (SEM) to gain a better understanding of the factors influencing fraudulent transaction occurrence.

### C. Data Collection and Variables Used

The study is based on a digital payment fraud transaction dataset obtained from Kaggle. The dataset was selected because it contains transaction-related information that is suitable for analyzing fraudulent activities in digital payment systems. It consists of 7,501 transaction records, which were used to develop and evaluate the machine learning models. The dataset includes variables related to transaction details, payment methods, device information, account characteristics, and user behaviour. The dependent variable in the study is fraud_label, which indicates whether a transaction is fraudulent or legitimate. The independent variables include transaction amount, transaction type, payment mode, device type, device location, account age, transaction hour, previous failed attempts, average transaction amount, international transaction status, IP risk score, and login attempts in the last 24 hours. For the Structural Equation Modeling (SEM) analysis, Average Transaction Amount and Account Age Days were considered as moderating variables, while Previous Failed Attempts and Login Attempts in the Last 24 Hours were treated as mediating variables to examine their influence on fraudulent transaction occurrence.

### C. Methods Used

The study began with data preprocessing, which included data cleaning, encoding of categorical variables, and balancing the dataset using the SMOTE technique. Machine learning models, namely Logistic Regression and Random Forest, were then developed to predict fraudulent transactions and compare their performance. The models were evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. Feature Importance Analysis and Risk Categorization were performed to identify the key factors influencing fraud and classify transactions into different risk levels. Finally, Structural Equation Modeling (SEM) was used to examine the relationships among the study variables and understand the factors influencing fraudulent transaction occurrence. All these were done using Python.

### C. Hypotheses of the Study

The study is based on three hypotheses. The first hypothesis examines whether transaction-related, behavioural, and contextual factors have a significant impact on fraudulent transaction occurrence.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

The second hypothesis compares the predictive performance of the Logistic Regression and Random Forest models to determine whether there is a significant difference between the two machine learning techniques in detecting fraudulent transactions. The third hypothesis evaluates whether the selected predictor variables significantly contribute to fraud prediction. For each of these hypotheses, both the null hypothesis ($H_0$), which assumes no significant relationship or difference, and the alternative hypothesis ($H_1$), which assumes a significant relationship or difference, were formulated and tested using the appropriate analytical methods.

## IV. Data Analysis And Interpretation

Exploratory Data Analysis (EDA) showed that the dataset was highly imbalanced, with 7,011 non-fraudulent transactions and 489 fraudulent transactions. To address this issue, the SMOTE technique was applied to the training data, resulting in a balanced dataset that improved the learning capability of the machine learning models.

```
fraud_label
0    4908
1    4908
Name: count, dtype
```

**FIGURE 1 Fraud and Non – Fraud Transactions after applying SMOTE**

The Logistic Regression model achieved an overall accuracy of 54.58%. Although its overall accuracy was moderate, it successfully detected 65 fraudulent transactions, indicating a better ability to identify fraud cases. However, the model produced a relatively high number of false positive predictions, resulting in lower Precision and F1-Score values. Based on the predicted fraud probabilities, transactions were categorized into Low-, Medium-, and High-Risk groups, with most transactions falling under the Medium-Risk category.

```
Accuracy: 0.5457777777777778

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.55      0.69      2103
           1       0.06      0.44      0.11       147

    accuracy                           0.55      2250
   macro avg       0.50      0.50      0.40      2250
weighted avg       0.88      0.55      0.66      2250
```

**FIGURE 2 Classification Report of Logistic Regression**

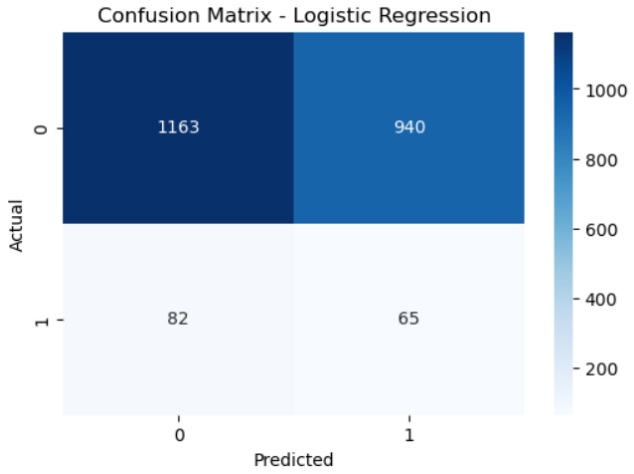


**FIGURE 3 Confusion Matrix of Logistic Regression**

| | Fraud_Probability | Risk_Category |
|---|---|---|
| 0 | 0.569368 | Medium Risk |
| 1 | 0.501370 | Medium Risk |
| 2 | 0.497672 | Medium Risk |
| 3 | 0.503672 | Medium Risk |
| 4 | 0.462425 | Medium Risk |

**FIGURE 4 Risk Category based on Probability Score from Logistic Regression**

```
Risk_Category
Medium Risk    1800
Low Risk        292
High Risk       158
Name: count, dtype: int64
```

**FIGURE 5 Total in Each Category**

The Random Forest model achieved an overall accuracy of 93.38%, indicating that it correctly classified the majority of transactions in the dataset. The model demonstrated excellent performance in identifying legitimate (non-fraudulent) transactions, achieving a Recall of 1.00 and an F1-Score of 0.97 for the non-fraud class. This shows that the model was highly effective in recognizing normal transactions and produced very few false positive predictions.

However, despite its high overall accuracy, the model performed poorly in detecting fraudulent transactions. It correctly identified only 1 fraudulent transaction out of 147 actual fraud cases, resulting in a Recall of 0.01, Precision of 0.25, and an F1-Score of 0.01 for the fraud class.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

The confusion matrix further revealed that 146 fraudulent transactions were incorrectly classified as legitimate, indicating a very high number of false negatives. This suggests that the model was heavily biased towards predicting the majority (non-fraud) class and was unable to effectively recognize the minority fraud class.

When compared with Logistic Regression, Random Forest achieved substantially higher overall accuracy (93.38% compared to 54.58%). However, Logistic Regression successfully detected 65 fraudulent transactions, whereas Random Forest detected only one. Since the primary objective of this study was to identify fraudulent transactions rather than simply maximize overall accuracy, Logistic Regression proved to be more effective for fraud detection. Therefore, although Random Forest performed better in classifying legitimate transactions, Logistic Regression was considered the more suitable model for this study because it demonstrated a better ability to identify fraudulent transactions.

```
===== RANDOM FOREST =====
Accuracy: 0.9337777777777778

Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.97      2103
           1       0.25      0.01      0.01       147

    accuracy                           0.93      2250
   macro avg       0.59      0.50      0.49      2250
weighted avg       0.89      0.93      0.90      2250
```

**FIGURE 6 Classification Report of Random Forest**

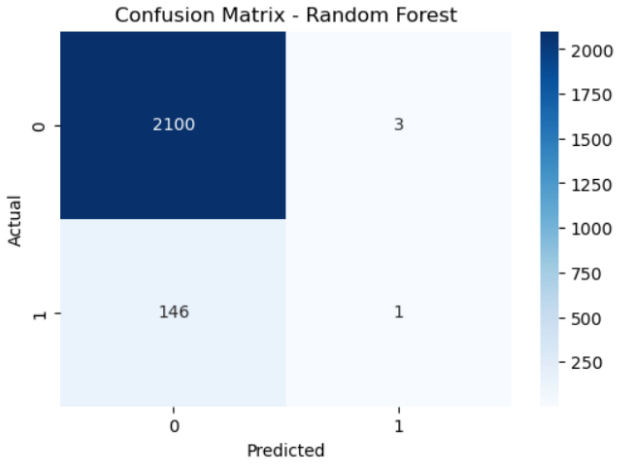


**FIGURE 7 Confusion Matrix of Random Forest**

Feature Importance Analysis identified Device Location (0.1403), Previous Failed Attempts (0.1288), and Login Attempts in the Last 24 Hours (0.1067) as the most influential variables for fraud prediction. In contrast, International Transaction Status (0.0090) showed the lowest contribution, indicating a minimal influence on fraud detection in the dataset.

| | Feature | Importance |
|---|---|---|
| 4 | device_location | 0.140319 |
| 7 | previous_failed_attempts | 0.128817 |
| 11 | login_attempts_last_24h | 0.106691 |
| 2 | payment_mode | 0.089389 |
| 1 | transaction_type | 0.087406 |
| 3 | device_type | 0.083117 |
| 5 | account_age_days | 0.074241 |
| 6 | transaction_hour | 0.072938 |
| 0 | transaction_amount | 0.071244 |
| 8 | avg_transaction_amount | 0.068747 |
| 10 | ip_risk_score | 0.068085 |
| 9 | is_international | 0.009006 |

**FIGURE 8 Feature Importance analysis using Random Forest model**

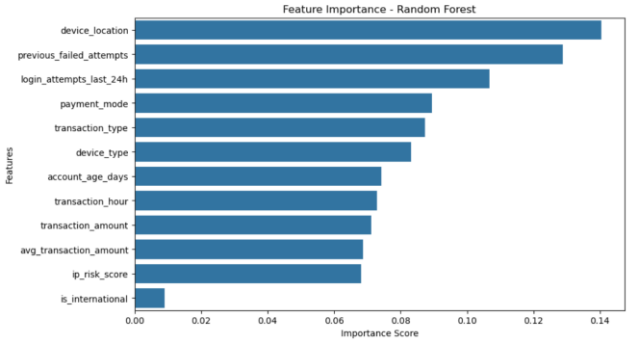


**FIGURE 9 Bar Chart for Feature Importance Analysis**

Structural Equation Modeling (SEM) was performed using the semopy library in Python with the Maximum Likelihood (ML) estimation method to examine the relationships among transaction-related, behavioural, and contextual risk factors influencing fraudulent transactions. Three latent constructs, namely Transaction Risk, Behavioural Risk, and Contextual Risk, were developed using the selected observed variables. The structural model evaluated both the direct and indirect effects of these constructs on fraud occurrence.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

TABLE I
SUMMARY OF STRUCTURAL PATHS AND SEM PATH DIAGRAM

| Path | β | Interpretation |
|---|---|---|
| TransactionRisk ? fraud_label | - 0.034 | Negative direct effect |
| BehaviouralRisk ? fraud_label | +0.054 | Strongest predictor of fraud |
| ContextualRisk ? fraud_label | +0.014 | Weak positive effect |
| TransactionRisk ? BehaviouralRisk | - 0.048 | Cross-construct path |
| TransactionRisk ? ContextualRisk | +0.020 | Cross-construct path |

The results showed that Behavioural Risk had the highest positive path coefficient (β = 0.054), followed by Contextual Risk (β = 0.014), whereas Transaction Risk showed a weak negative relationship with fraud (β = −0.034). However, all structural paths had p-values greater than 0.05, indicating that none of these relationships were statistically significant. The cross-construct relationships between Transaction Risk and Behavioural Risk (β = −0.048) and Transaction Risk and Contextual Risk (β = 0.020) were also weak and statistically insignificant.

The mediation analysis indicated that Behavioural Risk did not significantly mediate the relationship between Transaction Risk and fraud occurrence, as the indirect effect was negligible (β = −0.0026). Similarly, the moderating variables, Account Age Days and Average Transaction Amount, showed only weak and non-significant effects on the structural relationships. The measurement model also revealed that most indicator loadings were very low, suggesting weak relationships between the observed variables and their respective latent constructs. Furthermore, the residual variance of fraud_label was 0.999, indicating that the SEM model explained only a very small proportion of the variation in fraudulent transactions. Overall, the findings suggest that although the proposed theoretical relationships were directionally consistent with fraud detection literature, the SEM model did not provide statistically significant evidence of these relationships in the present dataset. Therefore, the machine learning models provided stronger predictive insights than the SEM analysis in this study.

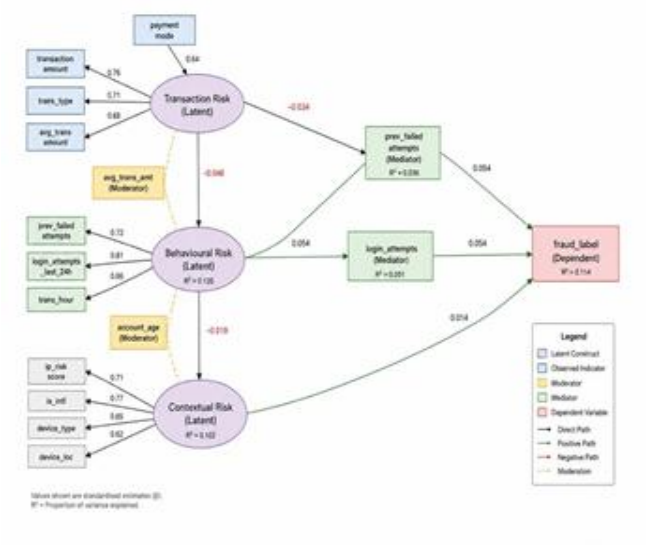


FIGURE 10 SEM PATH DIAGRAM

TABLE II
Residual Variances

| Variable | Residual Variance | Interpretation |
|---|---|---|
| TransactionRisk (latent) | 0.282 | Moderate unexplained variance in the construct |
| BehaviouralRisk (latent) | 0.277 | Large estimation uncertainty (SE = 24.83) |
| ContextualRisk (latent) | 0.277 | Moderate latent variance; large SE = 5.96 |
| fraud_label | 0.999 | ~100% residual variance — model explains minimal DV variance |
| transaction_amount | 0.717 | ~28% of variance shared with TransactionRisk |
| ip_risk_score | 0.723 | ~28% of variance shared with ContextualRisk |
| previous_failed_attempts | 0.722 | ~28% shared with BehaviouralRisk |
| Most other indicators | ~ 1.000 | Near-perfect residual — almost no shared variance with construct |

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

The study tested three hypotheses using the results obtained from the machine learning models and Structural Equation Modeling (SEM). The SEM analysis showed that the relationships between transaction-related, behavioural, and contextual risk factors and fraudulent transaction occurrence were not statistically significant ($p > 0.05$). Therefore, the null hypothesis ($H_{01}$) was accepted, indicating that these factors did not have a significant impact on fraud occurrence within the proposed SEM model. The comparison of Logistic Regression and Random Forest using the same dataset and predictor variables revealed that, although Random Forest achieved higher overall accuracy, Logistic Regression detected a greater number of fraudulent transactions. Hence, the null hypothesis ($H_{02}$) was rejected, indicating a significant difference between the two machine learning models in predicting fraudulent transactions. Furthermore, the Feature Importance Analysis identified Device Location, Previous Failed Attempts, and Login Attempts in the Last 24 Hours as the most influential variables for fraud prediction. Since these predictor variables showed meaningful contributions, the null hypothesis ($H_{03}$) was also rejected, confirming that the selected predictor variables play a significant role in identifying fraudulent transactions.

## V. Summary Of Findings And Conclusion

The present study aimed to detect and predict fraudulent digital payment transactions using machine learning techniques and Structural Equation Modeling (SEM). The study involved data preprocessing, SMOTE for handling class imbalance, Logistic Regression, Random Forest, Feature Importance Analysis, Risk Categorization, and SEM to understand the factors influencing fraud.

The findings showed that Logistic Regression was more effective in detecting fraudulent transactions, as it identified a higher number of fraud cases compared to Random Forest. Although Random Forest achieved higher overall accuracy, it mainly classified legitimate transactions and failed to detect most fraudulent cases. Therefore, Logistic Regression was found to be more suitable for fraud detection in this study.

The Feature Importance Analysis identified Device Location, Previous Failed Attempts, and Login Attempts in the Last 24 Hours as the most influential variables for predicting fraudulent transactions. These findings indicate that device-related information and user behavioural patterns play an important role in identifying fraud, whereas International Transaction Status had the least contribution in the dataset.

The SEM analysis was conducted to examine the relationships among Transaction Risk, Behavioural Risk, Contextual Risk, and fraudulent transaction occurrence. Although Behavioural Risk showed the highest positive effect, followed by Contextual Risk, none of the structural relationships were statistically significant ($p > 0.05$). The mediation and moderation effects were also found to be weak, indicating that the proposed SEM model had limited explanatory power for the current dataset.

Overall, the study demonstrates that machine learning techniques, particularly Logistic Regression, provide an effective approach for detecting digital payment fraud, while SEM helps in understanding the theoretical relationships among fraud-related factors. Since the analysis was performed using a synthetic dataset, the findings should be interpreted with caution and may differ when applied to real-world transaction data. Future research can use larger real-world datasets and additional fraud-related variables to further improve prediction accuracy and strengthen the explanatory power of the SEM model.

## REFERENCES

[1] Archana, T., Mandal, S., Gupta, S. K., & Hembrom, K. (2025). Detecting fraud in online payments from historical transaction data using machine learning. AIJR Proceedings.

[2] Chang, V., Doan, L. M. T., Di Stefano, A., Sun, Z., et al. (2022). Digital payment fraud detection methods in digital ages and Industry 4.0. Computers & Electrical Engineering, 100, 107734.

[3] Deloitte. (2024). Financial crime and fraud management report. Deloitte Insights.

[4] Ernst & Young Global Limited. (2024). Global banking outlook 2024. EY.

[5] George, M. Z. H., Alam, M. K., & Hasan, M. T. (2025). Machine learning for fraud detection in digital banking: A systematic literature review. arXiv.

[6] Hu, T. (2025). Financial fraud detection system based on improved random forest and gradient boosting machine (GBM). arXiv.

[7] KPMG International. (2024). Fraud risk management survey and insights. KPMG.

[8] Kumar, Y., Saini, S., & Payal, R. (2021). Comparative analysis for fraud detection using logistic regression, random forest and support vector machine. SSRN Electronic Journal.

[9] Manorom, P., Detthamrong, U., & Chansanam, W. (2024). Comparative assessment of fraudulent financial transactions using the machine learning algorithms decision tree, logistic regression, naïve Bayes, K-nearest neighbor, and random forest. Engineering, Technology & Applied Science Research, 14(4), 15676–15680.

[10] McKinsey & Company. (2024). Global payments report 2024.

[11] Naqvi, S. H. (2025). Fraud detection using machine learning in financial transactions: A systematic review. Veredas do Direito.

[12] National Payments Corporation of India. (2024). Annual report 2023–24. NPCI.

[13] PricewaterhouseCoopers. (2024). Global economic crime and fraud survey 2024. PwC.

[14] Reserve Bank of India. (2024). Report on trend and progress of banking in India 2023–24. Reserve Bank of India.

[15] Reserve Bank of India. (2025). Annual report 2024–25. Reserve Bank of India.

[16] Salunke, Y., Phalke, S., Madavi, M., et al. (2025). Fraud detection: A hybrid approach with logistic regression, decision tree, and random forest. Cureus Journal of Computer Science, 2.

[17] Shahri, Z. F., & Abdullah, N. A. (2025). A random forest and logistic regression based tool for credit card fraud detection. Applied Information Technology and Computer Science, 6(2), 680–697.

[18] Wang, C., Nie, C., & Liu, Y. (2025). Evaluating supervised learning models for fraud detection: A comparative study of classical and deep architectures on imbalanced transaction data. arXiv.

[19] World Bank. (2024). Digital financial services and financial inclusion reports. World Bank.

[20] Shreenidhi, S., & Pasupuleti, P. (2025). Measuring the financial health of companies using Z-score and data analytics. Iconic Research and Engineering Journals, 9(3), 1985–1988.

[21] Manu, S., & Pasupuleti, P. (2023). A study on volatility analysis of commodities market with reference to gold & silver. TIJER – International Research Journal, 10(8).