

An Explainable Machine Learning Framework for Pre- Online Payment Fraud Risk Assessment and Prevention

Deepali Mahadev Shamane¹ and Ms. Rucha Mohite²

¹ Student, Master of Computer Applications

² Assistant Professor, Master of Computer Application

Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra, India.

Abstract: The work uses a PaySim-based transaction dataset and focuses on information that can be available before a transaction is complete including transaction step, amount, sender balance, receiver balance, fraud flag information and transaction type indicators. A 150-record working sample containing 135 15 fraudulent transactions was used for the main experiment. The Random Forest model was trained with class weighting and evaluated using five-fold stratified cross-validation. The resulting average accuracy was 94.67% precision was 80.00% recall was 46.67%, F1-score was 56.00% and ROC-AUC was 95.43%. SHAP was used to make the model easier to understand with sender balance, transaction step, amount and receiver balance emerging as the important features. A Gradio dashboard was also developed to show model performance, transaction distributions, SHAP importance, transaction-level fraud probability, risk category and recommended security action. The prototype converts a probability score into three risk levels: LOW, MEDIUM and HIGH. The system is intended as a research prototype showing how predictive modeling and explainable AI can support more transparent fraud prevention.

Keywords: Online Payment Fraud, Random Forest, Explainable AI, SHAP, Fraud Risk Assessment, Pre-Transaction Detection, PaySim, Gradio Dashboard

I. INTRODUCTION

Digital payments have become a part of everyday financial activity. At the time online payment platforms face the continuing challenge of detecting transactions that may be fraudulent. A useful fraud prevention system should not classify a transaction as normal or suspicious; it should also provide a risk estimate that can support an appropriate security response.

Traditional rule-based systems are useful for known patterns but fraudulent behavior can change over time and may involve several transaction characteristics at once. Machine learning can learn relationships among transaction variables. Provide a probability-based assessment. However a complex model can be difficult for users to trust if its decision cannot be explained. For this reason the proposed project combines Random Forest classification with SHAP-based explanations.

The main objective is to build a prototype that assesses fraud risk before transaction completion. Of giving only a binary output the system produces a fraud probability maps the probability to a risk category and recommends an action. The dashboard further allows the model results to be viewed through charts and a transaction simulation interface.



II. LITERATURE REVIEW

Machine learning has been widely studied for financial fraud detection because fraudulent and legitimate transactions can show patterns in amount, timing, balances and transaction type. Ensemble models such as Random Forest are attractive for this problem because they combine decision trees and can model non-linear relationships. Class imbalance is another issue in fraud detection because fraudulent transactions are usually much less frequent than legitimate transactions.

Explainable AI methods can improve the interpretation of machine learning predictions. SHAP (SHapley Additive exPlanations) provides feature contribution values that help explain why a prediction moves toward or away from the fraud class. In a fraud prevention setting this is useful because analysts can inspect the factors that contributed to a high-risk result than treating the model as a black box.

The present project applies these ideas to a controlled prototype using PaySim transaction data. The emphasis is not on claiming production-level performance. On demonstrating an end-to-end workflow from transaction inputs to prediction, explanation and risk-based action.

III. RESEARCH GAP

Many fraud detection demonstrations focus mainly on classification accuracy. A practical prevention workflow however also needs a decision layer and a user-facing interface. The project addresses this gap by combining three components: a machine learning prediction model, SHAP explanations and a dashboard that translates the predicted probability into a risk category.

Another important point is the timing of the prediction. The proposed prototype is designed around -transaction information rather than relying on variables that would only become available after a transaction has been completed. This makes the workflow more relevant to prevention, where the goal's to verify or hold a transaction before final completion.

IV. METHODOLOGY

The methodology follows a pipeline: dataset preparation, preprocessing, feature selection, model training, cross-validation transaction-level testing, explainability analysis and dashboard integration. The working experiment uses 150 records selected from the larger PaySim dataset. Of these records 135 are normal. 15 Are fraudulent giving a 10% fraud proportion in the experimental sample.

The Random Forest model uses 100 trees, a fixed state of 42 and balanced class weighting. The data are divided using an 80:20 stratified train-test split and five-fold StratifiedKFold cross-validation is used to obtain a stable estimate of model behavior.

Item	Details
Dataset	PaySim transaction data
Working sample	150 transactions
Class distribution	135 normal 15 fraud
Classifier	Random Forest
Number of trees	100
Class handling	class_weight = balanced
Train/Test split	80% / 20% stratified
Cross-validation	5-fold StratifiedKFold
Random state	42

Table 1. Main configuration used in the fraud risk assessment prototype.



V. DATASET AND PREPROCESSING

The full PaySim dataset used during project exploration contains 6,362,620 records and 11 columns. For the model experiment a smaller 150-record sample was used so that the complete prototype, evaluation, SHAP analysis and dashboard could be demonstrated in a controlled environment. The sample contains 135 transactions and 15 fraudulent transactions.

The project examines transaction timing, transaction amount, sender balance, receiver balance, transaction type and the available fraud flag information. Categorical transaction types are converted into model- indicator variables. The final pre-transaction feature set used by the model includes step, amount, oldbalanceOrg oldbalanceDest isFlaggedFraud and transaction-type indicators such as type_CASH_OUT, type_DEBIT, type_PAYMENT and type_TRANSFER.

The preprocessing stage is important because the model must receive the feature representation during training and during dashboard simulation. Keeping the input structure consistent reduces errors when a user enters a transaction into the interface.

VI. FEATURE ENGINEERING AND MODEL DEVELOPMENT

Feature engineering in this project focuses on variables that describe the state and context of a transaction. The senders previous balance and the receivers previous balance provide information about the accounts involved while the transaction amount captures the size of the transfer. The step variable represents the time step in the PaySim data. Transaction type indicators allow the classifier to distinguish between payment operations.

Random Forest was selected because it can capture -linear relationships without requiring a linear decision boundary. The balanced class weighting is particularly useful for fraud detection because the minority fraud class should not be ignored simply because normal transactions are more numerous.

The prediction layer produces a probability for the fraud class. This probability is then passed to a decision layer. The project uses three thresholds: below 30% is risk and the transaction can proceed; 30% to below 70% is MEDIUM risk and the transaction should be verified; 70% or above is HIGH risk and the transaction should be held or warned.

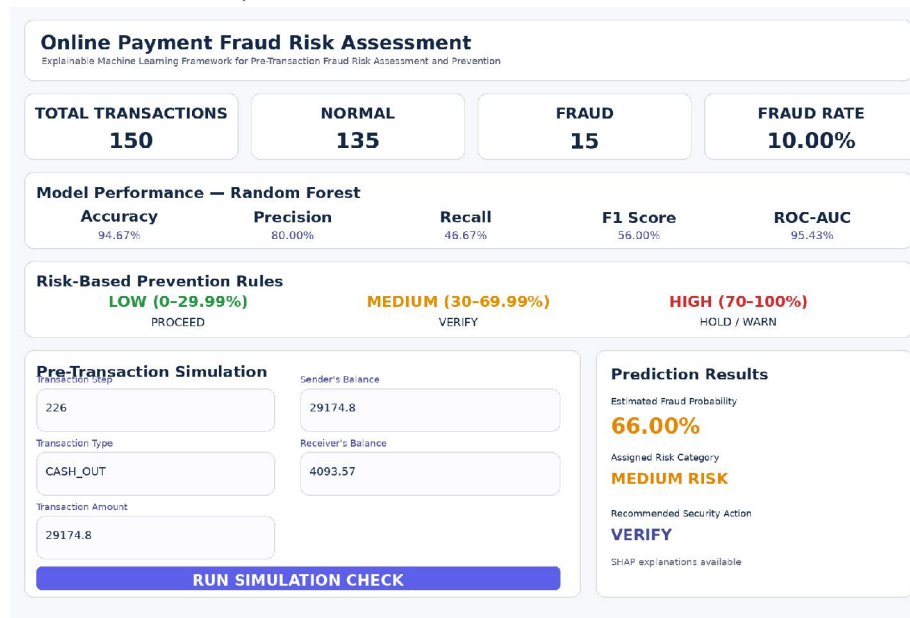


Figure 1. Integrated dashboard showing transaction summary, model performance, risk rules and pre-transaction simulation.



VII. DASHBOARD AND USER INTERFACE

A dedicated Gradio dashboard was developed as the presentation layer of the proposed fraud prevention system. The dashboard is designed to show the workflow in one place: dataset summary, model performance, risk rules, transaction simulation, prediction output and explainability. This makes the research model easier to demonstrate to a reviewer because the user can see both the results and the practical decision interface.

The dashboard begins with transaction-summary cards and the main Random Forest metrics giving the reviewer a view of the experiment and model quality.

The risk-based prevention panel converts the models probability into three categories. A probability below 30% is treated as risk and can proceed 30% to below 70% is MEDIUM risk and should be verified and 70% or above is HIGH risk and should be held or warned. This converts a machine learning score into an action-oriented security workflow.

The pre-transaction simulation panel accepts the transaction step, transaction type, transaction amount, senders balance and receivers balance. When the user runs the simulation the Random Forest model returns the estimated fraud probability. The result panel then displays the probability, risk category. Recommended security action. In the demonstrated example the transaction receives a 66.00% fraud probability, which maps to MEDIUM RISK and the VERIFY action.

The dashboard is linked with SHAP explanations so that a prediction can be reviewed using both its probability and the main contributing features.

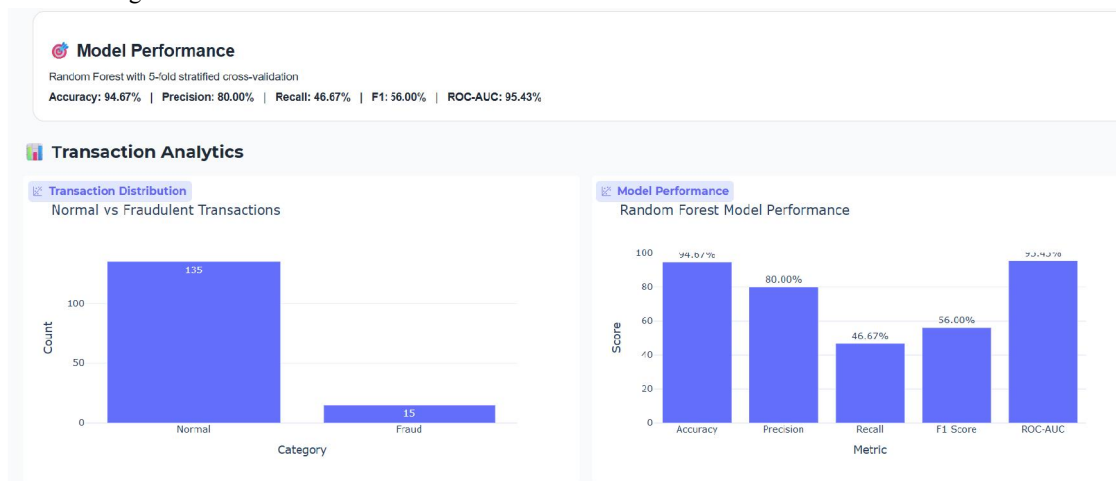


Figure 2. Dashboard evaluation view showing transaction distribution and Random Forest performance metrics.

VIII. MODEL EVALUATION AND RESULTS

The fivefold stratified cross-validation results provide the main evaluation reported in this study. The model achieved 94.67% accuracy, 80.00% precision 46.67% recall, 56.00% F1-score and 95.43% ROC-AUC. These values show that the model can separate normal and fraudulent cases, in the experimental sample while recall remains an important limitation because some fraudulent cases may not be detected.

Metric	Score	Interpretation
Accuracy	94.67%	Overall proportion of predictions
Precision	80.00%	Correctness of predicted fraud cases
Recall	46.67%	Fraud cases successfully identified
F1-score	56.00%	Balance between precision and recall
ROC-AUC	95.43%	Ability to rank the two classes

Table 2. Five-fold stratified cross-validation performance.



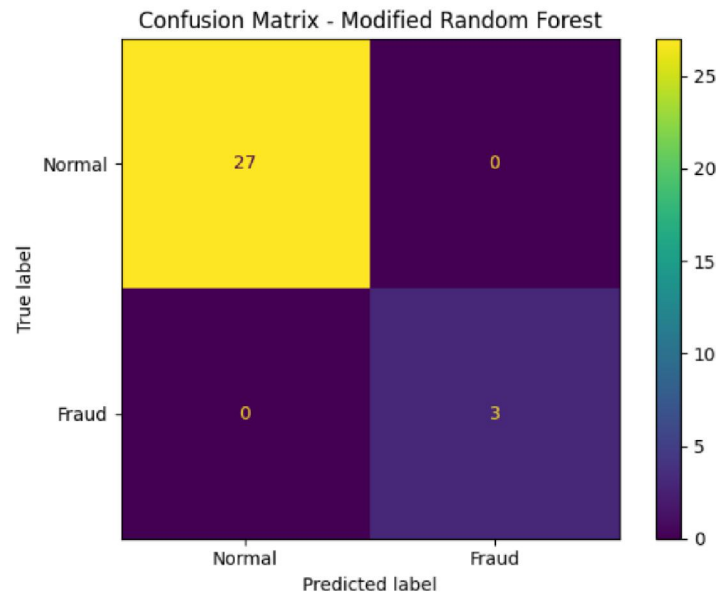


Figure 3. Confusion matrix for the 30-record hold-out test.

9. HOLD-OUT TEST OBSERVATION

The project also shows a hold-out test result on 30 records. The displayed confusion matrix contains 27 classified normal transactions and 3 correctly classified fraud transactions, with zero false positives and zero false negatives in that particular split. The associated visualization therefore shows a result for this small test subset.

This result should be interpreted cautiously. The hold-out subset contains 30 records and only 3 fraud cases so it is not sufficient evidence of production-level performance. The five-fold cross-validation results are more appropriate for the discussion because they average performance across multiple stratified folds.

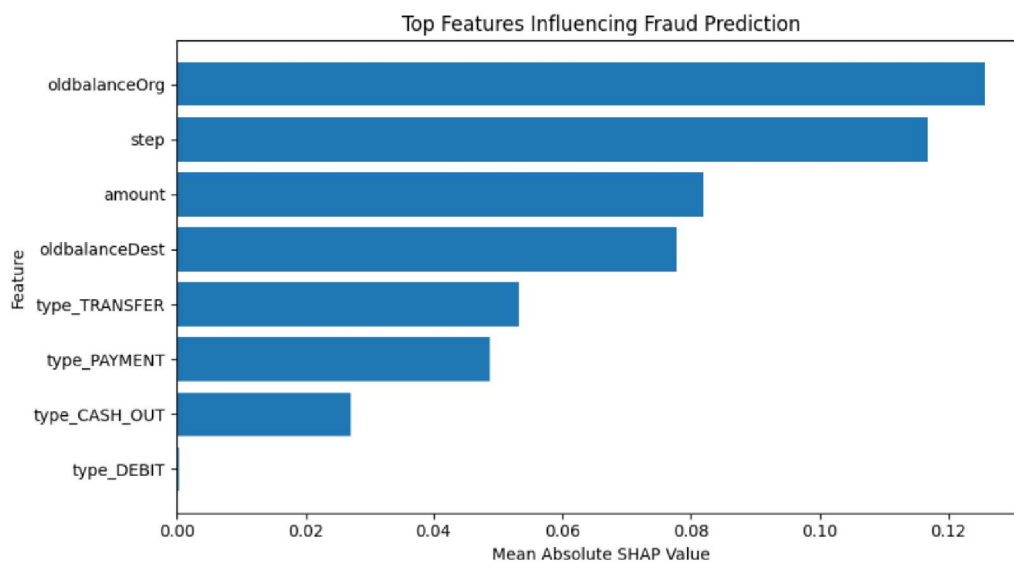


Figure 4. Global SHAP feature importance showing the factors influencing fraud prediction.



X. EXPLAINABLE AI USING SHAP

Explainability is a part of the proposed framework. SHAP values are used to estimate how individual features contribute to the models prediction. The global feature-importance chart shows that oldbalanceOrg has the mean absolute SHAP value, followed by step, amount, oldbalanceDest, type_TRANSFER, type_PAYMENT and type_CASH_OUT. Type_DEBIT has small contribution in the displayed analysis.

Feature	Mean Absolute SHAP
oldbalanceOrg	0.125702
step	0.116833
amount	0.082001
oldbalanceDest	0.077828
type_TRANSFER	0.053243
type_PAYMENT	0.048711
type_CASH_OUT	0.026910
type_DEBIT	0.000301

Table 3. Global SHAP feature importance from the project model.

For the displayed transaction-level example the estimated fraud probability is 66.00%, which places the transaction in the MEDIUM risk category. The recommended action is VERIFY than immediately holding the transaction. The projects individual SHAP analysis shows balance-related variables and transaction-type information as contributors to the prediction.

This approach makes the prediction easier to communicate. Of simply saying that a transaction is risky the system can indicate that the models decision is associated with specific transaction characteristics. This is particularly valuable for analysts who need to review or justify a security decision.

XI. PROPOSED FRAUD PREVENTION WORKFLOW

- Receive transaction information available before completion.
- Prepare the transaction in the feature format used during model training.
- Generate the Random Forest fraud probability.
- Use the 30% and 70% thresholds to assign LOW, MEDIUM or HIGH risk.
- Proceed with LOW-risk transactions.
- Request additional verification for MEDIUM-risk transactions.
- Warn for HIGH-risk transactions.
- Use SHAP explanations to inspect the factors behind the prediction.

XII. MAJOR FINDINGS

The project demonstrates that a Random Forest model can be integrated with a risk-based prevention layer. The cross-validation accuracy of 94.67% and ROC-AUC of 95.43% indicate class separation within the experimental setting. At the time the recall of 46.67% shows that fraud detection remains challenging and that accuracy alone should not be used as the primary measure of a fraud prevention system.

The SHAP analysis identifies sender balance, transaction step, transaction amount and receiver balance as the influential features in the model. The dashboard successfully brings these results with transaction simulation allowing a user to test an input and immediately see the predicted risk and recommended action.

XIII. DISCUSSION

The proposed system illustrates how fraud detection can be reframed as a risk assessment problem. A binary fraud label is useful for evaluation. A probability and risk category are more useful for operational decision-making. For example



the displayed 66% prediction is not treated as a block; it is classified as medium risk and sent for verification. This provides a flexible response than applying one identical action to every suspicious transaction.

The dashboard also demonstrates the value of combining analytics with explainability. Model metrics help assess behaviour while SHAP helps interpret individual or global feature influence. Together these components create a connection between machine learning output and a potential fraud prevention workflow.

XIV. LIMITATIONS

The main limitation is the size of the working experiment. Although the complete PaySim dataset contains millions of records the primary model demonstration uses 150 transactions, including 15 fraud cases. Therefore the reported metrics should be treated as prototype-level results than evidence of production readiness. The low recall also indicates that the model may miss a portion of fraudulent cases.

The 30-record hold-out result is based on three fraud cases so its perfect classification outcome should not be generalized. Further validation on a larger and more representative dataset would be required before deployment. Thresholds as 30% and 70% are also prototype decision rules and would need to be calibrated using real operational costs, false-positive tolerance and fraud-loss data.

XV. FUTURE SCOPE

- Validate the model on a much larger transaction sample.
- Compare Random Forest with gradient boosting, XGBoost, LightGBM and other suitable classifiers.
- Introduce richer behavioural and temporal features using historical transaction sequences.
- Calibrate fraud. Optimize risk thresholds using business costs.
- Add authentication and step-up verification workflows for medium-risk transactions.
- Monitor model. Retrain when fraud patterns change.
- Extend the dashboard with transaction history, alerts, audit logs and analyst feedback.
- Evaluate fairness, stability, latency and security before any real-world deployment.

XVI. CONCLUSION

This research presents a machine learning framework for pre-transaction online payment fraud risk assessment and prevention. A Random Forest classifier was trained on a controlled PaySim-based sample. Evaluated using stratified cross-validation. The model achieved 94.67% accuracy, 80.00% precision 46.67% recall, 56.00% F1-score and 95.43% ROC-AUC in the reported experiment.

The project goes beyond classification by adding SHAP-based explanations and a Gradio dashboard. The dashboard allows users to inspect model performance, transaction distributions, feature importance, transaction-level fraud probability, risk category and recommended action. The resulting workflow demonstrates how predictive modelling, explainable AI and a practical decision layer can work together to support fraud prevention.

Overall the prototype provides a foundation for future development. The next step is not simply to increase the accuracy number but to validate the model on data improve fraud recall calibrate risk thresholds and evaluate the system under realistic operational conditions.

REFERENCES

1. Breiman, L. Random Forests. Machine Learning, 45 5–32, 2001.
2. Lundberg, S. M. And Lee S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 2017.
3. PaySim: A financial mobile money simulator, for fraud detection research and transaction modelling.
4. Scikit-learn documentation: RandomForestClassifier, model evaluation and cross-validation methods.
5. SHAP documentation: SHapley Additive. Treeexplainer.
6. Gradio documentation: building machine learning interfaces.

