



Multi-Objective Explainable Transformer Framework for Fraud Risk Forecasting in Digital Payment Ecosystems

Dipendra Kumar Air

Assistant Professor

Central Department of Computer Science & Information Technology

Far Western University, Mahendranagar, Nepal

prodipu@gmail.com

<https://orcid.org/0009-0002-1036-6228>

Received: June 22, 2026

Revised & Accepted: August 19, 2026

Copyright: Author(s) (2026)



This work is licensed under a [Creative Commons Attribution-Non Commercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

Abstract

Background: The high rate of digital payment systems has led to rapid rate of transactions and increased vulnerability to financial fraud, thus leading to need to have intelligent, transparent, and proactive fraud monitoring solutions. Current fraud detection tools conceptualize the issue as binary classification, and act as black-box predictors, constraining their ability to provide future risk awareness and interpretability needed to comply with requirements in regulation and operational decision-making.

Methods: The paper presents an explainable, multi-objective Transformer model to predict fraud in urban digital payment ecosystems. It optimises real-time fraud detection and constantly-in-the-future risk estimation using a Transformer Temporal encoder and dual-output architecture for classification and regression. Explainable artificial intelligence and SHAP-based temporal attention visualisation enhance transparency and identify significant transaction attributes influencing predictions.

Results: Large-scale experimental results on a large-scale dataset of digital payment transactions show that the multi objective model can be effective in terms of competitive fraud detection with ROC-AUC of 0.843 and at the same time predict future risk of fraud with an RMSE of 0.049. Competitive comparison with classical machine learning classifiers and deep sequential baselines can prove the usefulness of the proposed strategy in balancing the predictive and forecasting performances as well as their interpretability.

Conclusion and Implication: The proposed framework can be viewed as a clear and active way of dealing with fraud monitoring in modern digital payment systems, as it combines the power of multi-objective learning and explainable modelling. The suggested framework



optimises both, complementary, goals in real-time fraud detection and constantly-in-the-future risk of fraud estimation.

Novelty: In contrast to the many single-task fraud detection models, this research demonstrates a joint multi-objective Transformer model for fraud classification and future risk forecasting, along with complementary explainability features via SHAP-based feature attribution and temporal attention analysis, which translates to transparent and anticipatory fraud risk management in digital payment ecosystems.

Keywords: Digital Payments; Explainable Artificial Intelligence (XAI); Fraud Detection; Multi-Objective Learning; Risk Forecasting; Transformer Networks.

1. Introduction

Urban digital payment surroundings have already become an inescapable aspect of contemporary financial facilities that enables the rapid and scaled money dealings with the participation of varied user groups ([Ansari et al., 2025](#)). On the one hand, these systems increase the accessibility and convenience of services, on the other hand, they increase vulnerabilities to advanced and rapidly developing fraud schemes ([Alex-Omiogbemi et al., 2024](#)). As a result, financial institutions are forced to implement automated and intelligent solutions that can ensure real-time fraudulent activity detection without the need to compromise the transparency and reliability of their operations ([Rane et al., 2023](#)). The conventional methods of fraud detection are mostly based on rules or machine learning classifications that are not dynamically updated ([Khanum et al., 2024](#)). In spite of the fact that these techniques can achieve a high precision in experimental environments, they usually will be black-box predictors, and they fail to predict future fraud risk ([Oko-Odion, 2025](#)). In addition, they are often configured as single-task classifiers, that is, systems that do not consider either temporal behaviour or proactive risk evaluation ([Xu et al., 2024](#)). Regulatory and safety-critical financial systems face no interpretability and predictive power compromise confidence and application utility ([Wigerthale and Reich, 2024](#)). See recent progress such as in deep learning, especially sequence-modelling methods, including recurrent neural networks and Transformer models, have facilitated better the presentation of transaction histories and as well as user-behaviour patterns. However, current deep learning-based fraud detection systems still focus on binary classification and fail to merge risk forecasting and explainability into a single system. This disconnect inspires a holistic system that integrates detection and forecasting and interpretability in a one system.

1.1 Problem Statement

Modern fraud detection systems of online payments mainly consider the problem of fraud detection as a static binary classification choice. These systems are reactive, which means that they react only after fraudulent behaviour has exploited itself. They do not have mechanisms of quantifying future risk and cannot give early warning to facilitate preventive intervention. Additionally, most high-performing models act as black boxes giving limited information about what influences their decisions, which makes them difficult to comply with



the regulations, and which makes them very difficult to trust in functioning. Based on this, an urgent solution is a framework of fraud monitoring that is sufficiently sensitive to detect fraudulent transactions as well as predict imminent risk of fraud and maintain interpretability. The absence of a unified forecasting and explainability limits the application of smart fraud detection to real-life financial systems where responsibility, openness and action are imperative.

1.2 Research Questions

In order to solve the challenges mentioned above, the current research is informed with the following research questions:

RQ1: Does a Transformer based temporal model have the ability to use sequential transactions in detecting fraud in online payment systems?

RQ2: Can we simultaneously conduct rampant classification of frauds and future risk in frauds forecasting under a single multi-objective learning system?

RQ3: How is explainability added to give both feature-level and time-level explanation on the prediction of fraud?

RQ4: Which are an aspects of the proposed multi-objective framework can be compared to the classical machine learning models and deep sequential baselines in terms of predictive performance and interpretability?

1.3 Research Objectives

The major findings of this study are the following:

1. To construct a Transformer-based temporal modelling approach in modeling sequential transaction behaviour in digital payment systems.
2. To create a multi-objective learning architecture which is able to detect fraud and at the same time continuously provide forecasts of future risks.
3. To integrate explainability processes that give transparent and understandable predictions of fraud on both features and time scales.
4. To experimentally test the proposed procedure on a large-scale transaction dataset on a massive scale and compare the suggested method with machine learning and deep learning staples.
5. To discuss the feasible use of the given framework as applied to fraud monitoring on a proactive and transparent basis in open urban digital payment settings.

2. Review of Literature

The development of fraud detection algorithms in digital payment systems has moved past traditional rule-based algorithms to high-tech machine learning algorithms and most recently deep learning based algorithms ([Islam et al., 2024](#); [Njoku et al., 2024](#)). Different algorithms of classical supervised learning, such as logistic regression, random forest, or gradient boosting, have been heavily used due to their simplicity and strong results on structured transactional data ([Pratama and Wahid, 2025](#), [Ahmed et al., 2024](#)). They are generally hand-engineered feature engineering algorithms and the conceptualisation of the problem of fraud



detection as a traditional binary classification ([Leevy et al., 2023](#); [Alhashmi et al., 2023](#)). As sequential records of transactions grow, recurrent neural networks (RNNs), long short term memory (LSTM) networks and Transformer networks have been proposed to learn temporal dynamics and predict patterns of user behaviour ([Sebt et al., 2021](#)). Models based on transformers especially demonstrated higher ability to capture long-range dependencies with attention mechanisms ([Huang et al., 2023](#); [Qin et al., 2023](#)). However, the majority of modern deep-learning methodologies focus only on real-time fraud detection and do not include the predictive use of potential fraud. Financial contexts too have been observed to have strong interest in explainable artificial intelligence (XAI) techniques. SHAP and LIME are commonly used to interpret model predictions using feature attribution, and attention mechanisms are not that interpretable with time ([Kalasampath et al., 2025](#); [Ortigossa et al., 2024](#)). However, explainability is usually considered a post-hoc addition as opposed to an inseparable part of fraud detection systems, and there are few researches that were successful at combining feature-level and temporal-level explanations in a comprehensive framework ([Al-Daoud and Abu-AlSondos, 2025](#); [Ramachandran, 2025](#)).

2.1 Research Gap

Although a number of significant progress has been achieved, it continues to have three significant gaps in the literature. To start with, most fraud detection systems are more reactive as they seek to detect fraudulent transactions after they have occurred and therefore lack the ability to estimate fraud risk in the near future. Second, the existing models mostly operate as single-task classifiers and cannot optimally work together in terms of achieving the detection and forecasting goals. Third, explainability is still immature, with the results of the work generally providing either a feature level or a temporal level explanation, but not both in a single framework. The mentioned flaws limit the practical relevance of fraud detection systems to proactive and regulatory-based financial environment.

2.2 Theoretical Underpinning

Multi-objective learning and sequence modelling are the principles the proposed framework is based on. Multi-objective learning can be used to optimise both matching tasks simultaneously, such as in the case of counteracting fraud and predicting risk, and in this context to both learn collective representations which capture both predictive and discriminative data. Transformer based attention models provide a theoretically sound way of capturing time variations through the provisioning of historical transactions with weights of relevance to context. The explainability would be realised by combined temporal weighting (by the use of an attribution theory (via SHAP) and attention) and a temporal weighting, thus providing information about the values of feature and time-step effects. Together, these theoretical foundations give a theoretical form of consistent basis of proactive and interpretable fraud risk modelling.



3. Methodology (Proposed Framework)

The current section provides a definition of the proposed multi-objective, interpretable Transformer-based fraud risk forecasting in the context of modern city digital payment markets. The framework can be configured to interestingly collaborate to detect fraud and make future risk approximations alongside providing interpretable predictions, at feature and time scales. The generalization of the pipeline is include data preprocessing, construction of the temporal sequences, Transformer-based representation learning, multi-objective output modelling, and the explainability analysis.

3.1 Data Representation and Sequence Construction

The combination of input numerical and categorical attributes (such as transaction step, transaction type, transaction amount, account balance, etc) is used to represent each transaction. Categorical-types Encoded by the numbers, categorical transaction types are normalised so that training dynamics are not disturbed. In order to represent the temporal information, the transactions are structured in form of suppliers of length L . Given a sequence processes, $X_t = \{x_{t-L+1}, \dots, x_t\}$, two target variables will be considered: (i), a binary number fraud, y_t , that indicates whether a particular transaction is a fraud or not; and (ii), the continuous future risk score, r_t which is the average percentage occurrence of the fraud over a future horizon, T . This expression recasts identifying fraud as a purely reactionary process to a risk-projection issue.

3.2 Transformer-Based Temporal Encoder

The temporal encoder is implemented with a Transformer architecture in the form of stacked self-attention blocks. Considering an input sequence, denoted as $X \in R^{L \times d}$, in the form of a weighted mean of transactions in the sequence, that is, in the form of the dimension of features, i. e the dimension of the proximal data, denoted as d , the self-attention mechanism will derive contextual representations of the transaction by considering each transaction importance to all other transactions in the sequence. A Transformer block comprises of a multi-head self-attention layer and a position-wise feed-forward network, residual connection and layer normalisation. The model allows the design to pick up long term behavioural patterns and short term anomalies in transaction histories. The proposed multi-objective Transformer framework described in algorithm 1 concurrently classifies fraud and predicts future risk based on incremental transactions is presented. It combines temporal modelling with explainability mechanisms, which make predicting fraud-risks and their interpretations proactive.

Algorithm 1: Multi-Objective Explainable Transformer for Fraud Risk Forecasting

Input:	Transaction dataset D , sequence length L , forecast horizon T
Output:	Fraud probability $\hat{y}$ and future risk score $\hat{r}$
1. Pre-process	dataset D :
	Encode categorical attributes and normalize numerical features
2. Sort transactions chronologically by time step	

3. Construct sequences $X_t = \{x_{t-L+1}, \dots, x_t\}$
4. For each sequence, define:
 1. Classification target y_t (current fraud label)
 2. Risk target r_t (average fraud occurrence over horizon T)
5. Pass X_t through Transformer encoder to obtain temporal representation h_t
6. Compute fraud probability:

$$\hat{y}_t = \sigma(W_c h_t + b_c)$$
7. Compute risk forecast:

$$\hat{r}_t = W_r h_t + b_r$$
8. Optimize combined loss:

$$L = \lambda_1 L_{cls}(y_t, \hat{y}_t) + \lambda_2 L_{reg}(r_t, \hat{r}_t)$$
9. Generate explanations using SHAP (feature-level) and attention weights (temporal-level).
10. Output interpretable fraud and risk predictions.

3.3 Multi-Objective Output Modeling

The Transformer encoder generates the representation of the time and passes it into a dual-output model that optimizes two complementary learning tasks: fraud classification and future fraud risk forecasting. The first output head is a binary classifier that has a sigmoid function to output the probability of the current transaction being fraudulent. The second output head is a linear activation function that predicts a continuous fraud risk score, which is the mean from a predictably future fraud risk horizon.

The risk score is not the prediction of the next label occurrence but rather the mean amount of fraud occurrences for the next T transactions. The target value of the risk score for a sequence that ends at time step t is expression:

$$r_t = \frac{1}{T} \sum_{i=1}^T y_{t+i}$$

where y_{t+i} be the fraud label of the i th transaction for the future with T a forecasting horizon. This formulation allows the model to learn from future logs given the historical patterns of fraud by replacing the sparse binary target with a continuous regression target, rather than simply memorizing future fraud. The loss function used for the network consists of a weighted sum of classification error and regression error, and is minimized in the training:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg}$$

$\mathcal{L}_{cls}$ is the binary cross-entropy loss for fraud classification; λ_1 and λ_2 are the relative weights regulating the contribution of the two learning targets; $\mathcal{L}_{reg}$ is the mean squared error for risk forecasting. Future fractions are only employed during the training phase for constructing the target, while only historical transaction sequences are given to the model during the inference phase, without having access to future transactions in order to predict future risk levels. This



design helps to prevent simply resuming forecasts from being short-term deterministic memorization instead of forecasting.

3.4 Explainability Mechanisms

The proposed framework brings in explainability in the features and at the temporal level, thus facilitating the clarity of fraud predictors. Feature-level Interpretation is achieved by applying SHapley Additive exPlanations (SHAP), an additive model whose contribution measures the impact of each variable that went into the prediction of the fraud probability. Explanations were created by subsampling the test set to get a more representative set of evaluation sequences, because SHAP is very sensitive to computational costs for deep neural networks. Many existing methods for explainable deep learning have used this sampling method to identify a reasonable value for the global importance of features that yields reliable results within a computational tractable range.

Temporal explainability is achieved by combining self-attention mechanism in the Transformer encoder with time series augment. The attention coefficients are taken from the last layer of the attention and rescaled such that the coefficients for each transaction sum to unity. The normalized attention distribution can be used to compare the transaction history in a quantitative fashion that facilitates the identification of which parts of the transaction history are the most relevant to the prediction of fraud. SHAP is combined with temporal attention to get complementary explanations by: What factor(s) of the transaction have an impact on the model; When changes in the history of behavior are most important for the current transaction.

3.5 Training and Optimization

The proposed framework was trained using the mini-batch gradient descent with Adam optimizer and adaptive learning rate optimization. In order to address the extreme class imbalance of the transaction data, sample weighting was used as part of the model training process; mis-classifying fraudulent transactions received heavier penalties than those of legitimate transactions, and the penalty system aimed to make the model more sensitive towards the minority class. To handle the joint learning problems of fraud classification and future risk forecasting, a Multi Objective Optimization strategy was used. The overall Cost function is defined by a weighted sum of the mean squared error (MSE) and the binary cross-entropy (BCE) with equal loss weights, ($\lambda_1 = 1.0$ and $\lambda_2 = 1.0$), to help balance the learning from both tasks. These weights are empirical and were validated using alternative set-ups of these weights on the validation set, and the model that resulted in the most stable convergence and overall predictive accuracy was chosen. For fraud classification, ROC-AUC and PR-AUC were used to evaluate the performance of the model, while Root Mean Square Error (RMSE) measured the performance of a model when forecasting future fraud. The overall workflow is shown in [figure 1](#), where fraud classification and risk forecasting are done in parallel and both tasks are then customized for multi-objective optimization while the end result is explained using SHAP and attention.

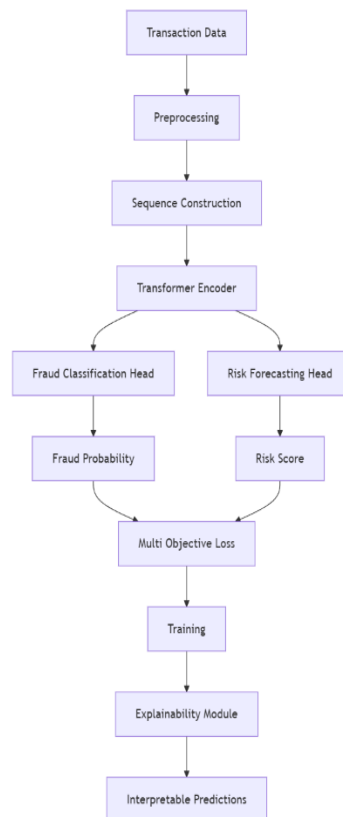


Figure1. *Architecture of the proposed multi-objective explainable Transformer framework for fraud risk forecasting*

4. Experimental Setup

It explains the dataset, pre-processing scheme, baseline models, evaluation measures, and the implementation in this part to verify the proposed multi-objective explainable Transformer framework. The experiment design is driven by the desire to have fair comparison, reproducibility, and realistic evaluation of fraud detection and predictions of risks performance.

4.1 Dataset Description

Experiments have been performed on a publicly available financial fraud dataset, called PaySim, created by Lopez-Rojas ([Lopez-Rojas et al,2016](#)). Instead of being manually generated, the dataset is created from realistic digital payment transactions that are drawn from transaction logs collected during the operation of a mobile money service and embedded in the PaySim mobile money simulator. Five types of transactions (CASH-IN, CASH-OUT, PAYMENT, TRANSFER, and DEBIT) are modeled, while the temporal behavior and fraud distribution seen in real financial systems are kept. The data comprises 6,362,620 transactions that are ordered by occurrence, each transaction is accompanied by 11 original attributes including transaction type, amount, account balances, temporal step and fraud label. The data includes fraudulent transactions, which is only 8213 records (0.1291%)



but is highly unbalanced making fraud detection very challenging and better suited for temporal deep learning models.

Table 1. *Summary statistics of the digital payment transaction dataset*

Metric	Value
Total transactions	6,362,620
Fraudulent transactions	8,213
Non-fraudulent transactions	6,354,407
Fraud ratio (%)	0.1291%
Number of original attributes	11
Number of features used for modeling	7
Sequence length (L)	10
Forecast horizon (T)	5
Training set size	4,463,587
Validation set size	980,416
Test set size	918,617

The table of the scale, imbalance/ratio and temperature arrangement of the in-use data in this study are summarised in [Table 1](#). The risk of fraud is very low as indicated by the very low rate of fraud, which indicates the challenge of identifying fraud and is the reason why temporal modelling and multi-objective learning should be used.

4.2 Data Pre-processing

The data for transactions was pre-processed using encoding of the categorical transaction type data into numerical, and normalization of continuous transaction amount and balance-related attributes via z-scores. This signed temporal step variable was used to sort all the transactions by the natural payment transaction ordering within the temporal step values, prior to the process of sequence construction. Instead of random sampling, temporal hold out was selected to avoid risk of information leakage. The first 70% of the chronologically ordered transactions (4,463,587 records) were utilized for modelling, the second 15% to validate (980,416 records) and the last 15% to test only (918,617 records). Split were made in the same order (temporal split) so that there was no training of the transaction appearance in the future. To model temporal dependencies, the transactions were transformed to fixed-length sequences of fixed 10 transactions each. For each sequence, the original chronological order is maintained and the fraud status of the subsequent (final) transaction, along with its risk scores for 5 next transactions forecast period, is predicted. In each data partition sequences were generated independently and sequences were not shared across the boundaries of training, validation and test partitions. No artificial time gaps or info leakage, so it was possible to assess the model on new stream information from the future. The only difference from the actual situation is that an actual deployment must be able to learn only from past transactions and to become suitable for future transaction flows while screening for fraud.

4.3 Baseline Models



The suggested framework is contrasted to various baseline methods of modelling with the aim of testing its efficiency within a variety of modelling paradigms. Logistic Regression can be used as a baseline in linear classifier because it is a clear and computationally effective standard. Random Forest is used as a non-linear ensemble algorithm which can be used to model complex interactions between features. Sequential baselines is contain a Long Short-Term Memory (LSTM) network, which models time relationships in the form of recurrent links, and a single-task Transformer network, which uses a self-attention on sequence modelling but only maximizes the fraud classification. All these baselines can be considered classical machine learning, recurrent neural networks, and attention-related deep learning models that are frequently deployed in literature regarding fraud detection.

4.4 Evaluation Metrics

ROC Area Under the Curve (ROC-AUC) and Precision-Recall Area Under the Curve (PR-AUC) are the measures used to classify the performance of fraud. The measures are specifically appropriate when the dataset is very imbalanced, as they evaluate the quality of classification and sensitivity to the detection of minority-classes. Root Mean Square Error (RMSE) is used as the metric to measure the risk forecasting performance, asserting the differences in predicted and actual future fraud risk. These measures taken jointly give a clear evaluation of accuracy of classification as well as reliability of forecasting.

4.5 Implementation Details

All implemented experiments are in Python 3.12 with TensorFlow/Keras framework to implement baseline and proposed deep learning model: and Scikit-learn framework for classifiers. It is suggested using two stacked Transformer encoder blocks, which each use a 32-dimensional hidden embedding and a 64-dimensional feed-forward network and 2-head multi-head self-attention. For the sequence representation, Global Average Pooling was used and two parallel output heads for fraud classification (with a sigmoid activation) and for future risk forecasting (with linear activation) were implemented. The model is trained with 1024 samples for one batch and the Adam optimizer with the learning rate 0.001. The parameters of the network were set by the Glorot Uniform initialization strategy to ensure convergence stably in the optimization. The convergence is stable, and hyperparameters were optimized using an empirical grid-based validation approach where several candidate sets of hyperparameters were tested on the validation set before the best performing set was utilized for the final model that was chosen based on the highest validation ROC-AUC value and convergence was stable. The training took place three times using Google Colab cloud platform with an NVIDIA Tesla T4 GPU. The training consisted of 3 runs on Google Colab with an NVIDIA Tesla T4 GPU cloud environment. This overall training process took about 2130 seconds or 35.5 minutes. Early stopping was not used because both training loss and validation AUC showed some stable convergence with the number of epochs chosen, meaning that they did not reach satisfactory model optimization typically by overfitting or needing more training epochs.

4.6 Data Availability

The PaySim dataset used in this study is publicly available through the Kaggle repository: <https://www.kaggle.com/datasets/ealaxi/paysim1>

5. Results and Discussion

In this part, the authors identify the empirical results that relate to the proposed multi-objective explainable Transformer framework, including the data properties, baseline model performance, and explainability studies.

5.1 Exploratory Data Analysis

The general distribution of the types of transactions in the dataset is shown in [Figure 2](#).

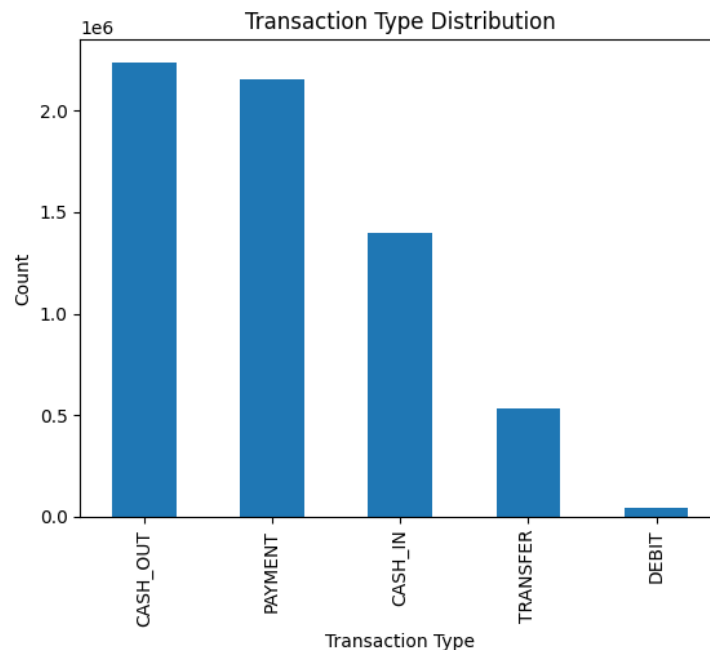


Figure 2: *Distribution of Transaction Types in the Digital Payment Dataset*

The records that are most common are cash-out and payment operations, they represent the most common pattern of using the digital payment systems in the city as most users are mostly involved in transferring funds and paying bills. Debit operations on the other hand are rare, which implies that the presence of direct debit operations are a small portion of the transaction activity. Such disproportional giving stresses the heterogeneity of the transaction behaviour and suggests that the fraud detecting models should be strong enough to deal with the highly skewed categorical distributions.

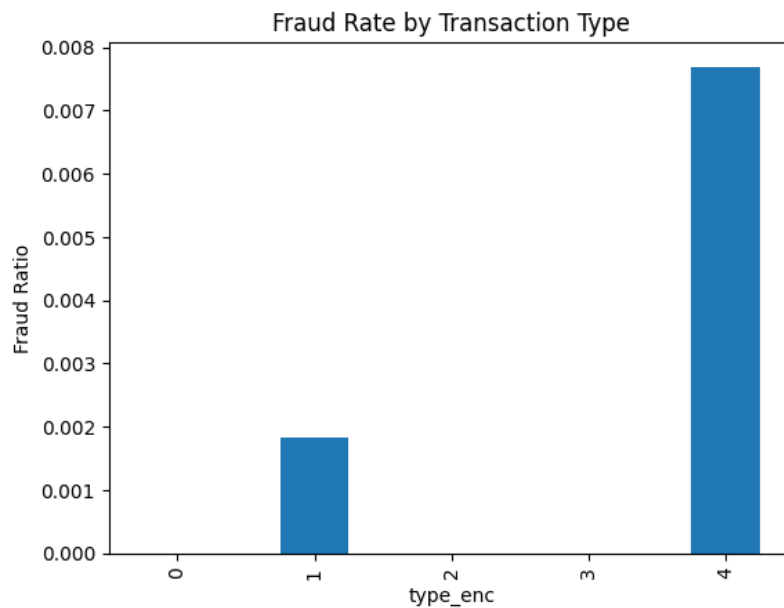


Figure 3: *Fraud Incidence Across Different Transaction Categories*

More on the connection between fraudulent activity and type of transaction is investigated in [Figure 3](#). Fraudulent transactions are also dissimilar in certain line of transaction, especially the transfer and cash-out transactions. This finding indicates that some transaction mechanisms are more susceptible to exploitation as they have immediacy and less authentication. In turn, the type of transaction becomes a significant predictive variable, and the models that fail to take into account this difference might fail when faced with the high-risk groups.

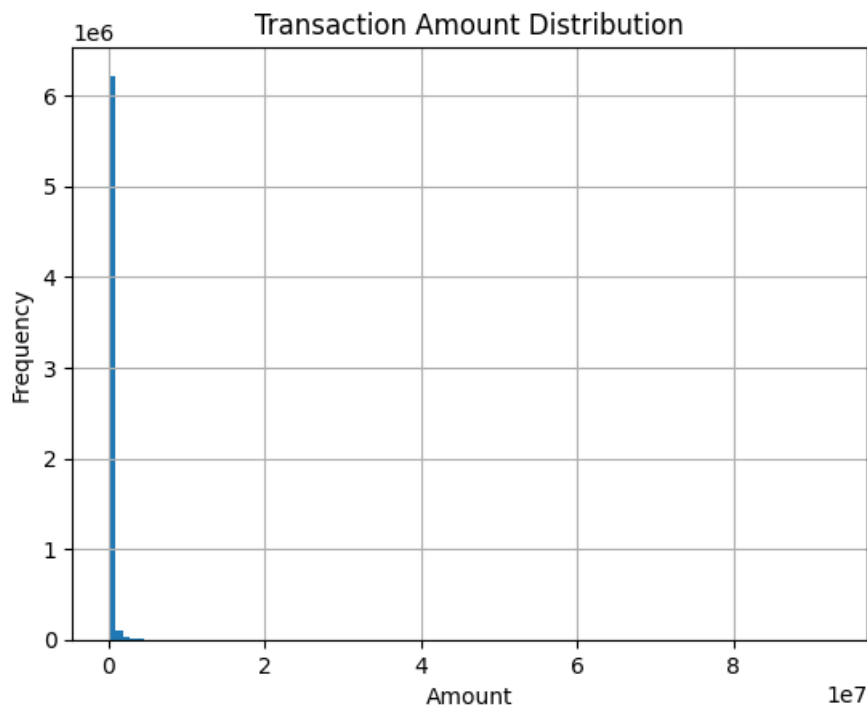


Figure 4: *Statistical Distribution of Transaction Amounts*

The figure of monetary nature of transactions is represented in [Figure 4](#). The skewness of the distribution is strong to the right with most of the transactions taking up small sums of money, and a few embodying very huge sums. This type of long-tailed distribution is characteristic of the data of financial transactions and presents a difficulty to statistical modelling, whereby extreme values have a disproportional impact on the algorithms used to learn. The heavy tail of the amount of transactions also suggests that abnormal transfers of high values can be used as one of the possible evidences of fraudulent behaviour.

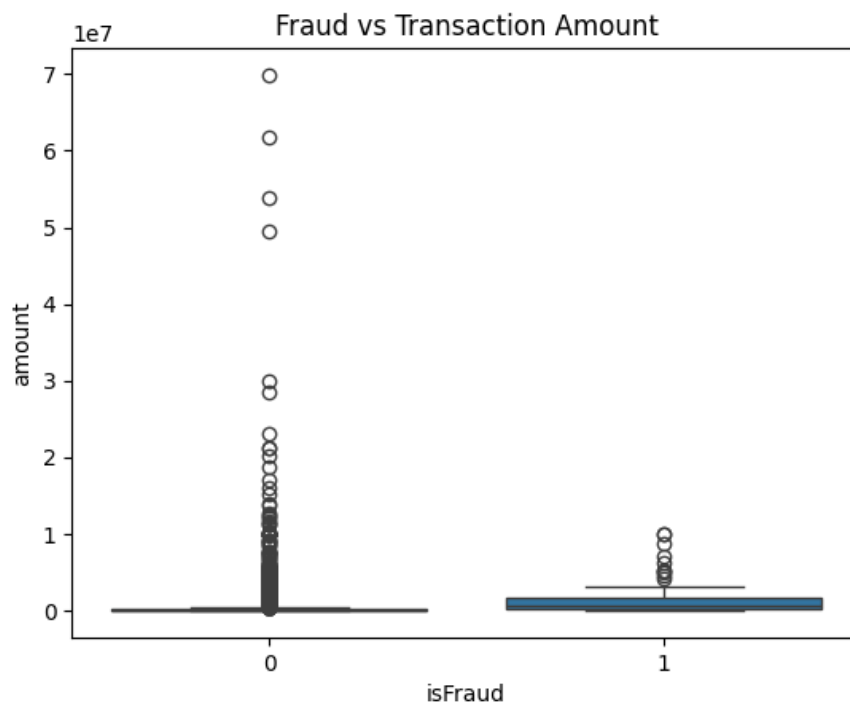


Figure 5: *Comparison of Transaction Amounts for Fraudulent and Legitimate Transactions*

It is also complicated in [Figure 5](#), where the number of transactions in legitimate and fraud pedigrees are compared in terms of the amount they implicated. Fraud activities are more diversified and are more high value at the monetary value than the legitimate transactions. The indicated pattern indicates that fraudsters tend to aim at maximising financial benefits per transaction, which implies that the profiles of amounts differ. Nevertheless, the fact that there is overlap between legitimate and fraudulent amounts reveal that the amount of transaction cannot be relied on to effectively detect fraud and this aspect leads to the requirement of a multi-dimensional and temporal modelling. Taken together, these exploratory findings indicate that the prevalence of fraud is intensely affected by category of transactions and scale of money, and has complicated assessments with the lawful actions. These insights are inspirational of the incorporation of sequential and multi-objective learning methods that have the ability of combining heterogeneous features and time dependency to enhance the distinction between fraudulent and legitimate actions.

5.2 Baseline Model Performance

The performance on classification of the baseline classifiers was evaluated using the Receiver Operating Characteristic (ROC) curves and Precision – Recall (PR) analysis as shown in [Figure 6](#) and [Table 2](#). The baselines assessed are linear machine learning, ensemble learning, recurrent neural networks and attention-based deep learning.

Table 2. *Classification performance of baseline models on the test set*

Model	ROC-AUC	PR-AUC	Accuracy
Logistic Regression	0.967	0.652	0.98
Random Forest	0.985	0.950	0.998
LSTM	0.749	0.326	–
Single-task Transformer	0.856	0.276	–

Random Forest had the best classification results, with a PR-AUC of 0.950 and a ROC-AUC of 0.985, and appeared to have the best potential for discriminating between fraudulent and legitimate transactions with structured tabular features. Because of its ensemble architecture, it is able to model highly non-linear correlations between the amount of transactions, the changes in the balance, and the type of transaction, yielding highly accurate single task classification boundaries of frauds. The accuracy (99.98%) achieved here needs to be taken with a pinch of salt as the number of transactions in the dataset is extremely unevenly distributed, with fraudulent transactions accounting for only 0.1291% of the total transactions. In this case, it is preponderated by the majority non-fraud class, and thus contributes little to the overall accuracy; that is, it fails to assess accurately the model's performance on the rare fraudulent events. For this reason, ROC-AUC and PR-AUC will be used as the key evaluation metrics for this study, because it provides much more reliable evaluation in highly imbalanced financial data. Even though the proposed multi-objective Transformer's fraud classification ROC-AUC score (0.843) is lower than the Random Forest model, both models are optimizing different kinds of optimization goals. Random Forest is only used for binary classification while the proposed work is used for both learning the fraud detection and learning the future risk forecasting as continuous class problem with a shared temporal representation. This multi-objective learning strategy is deliberately designed to trade off discriminative and predictive performance in learning, not just focusing on classification accuracy. Moreover, the single task Transformer achieved a PR-AUC of 0.276, which underperform Random Forest, although it uses Self-Attention. The difference demonstrates the failure of a sequence modeling task to produce a better precision for the minority class when considered as an optimization criterion for the classification task. The framework is built to go beyond classification and include temporal forecasting and explainability, enabling proactive commerce fraud monitoring over and above just reactive transaction monitoring.

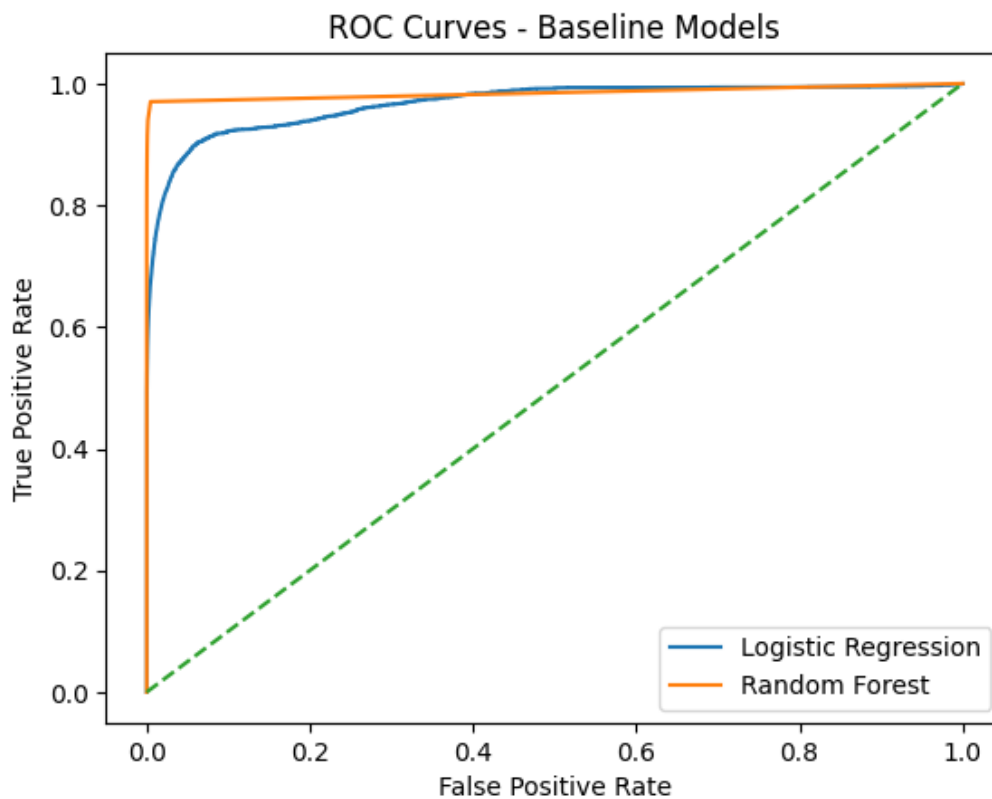


Figure 6: *Receiver Operating Characteristic (ROC) Curves for Baseline Classification Models*

Baseline models also have a high degree of classification accuracy but can take only snapshots of transaction features and hence find it impossible to examine how these transactions relate to the past. They fail to consider changing user behaviour or sequence of transactions that may be paramount in determining the obscure or sluggishly developing fraud patterns. Moreover, these models are reactionary in nature as they make predictions based on the present transaction and bypass future risk forecasting as well as early prevention solutions. The other weakness of baseline methods is that they are limited to be interpreted in a temporal framework. Although feature importance measures can be obtained based on tree-based models, it does not so as to show how past transaction behaviour affects the current predictions. In turn, these models give little information on the behavioural patterns and fail to enable proactive risk surveillance. On the whole, the results of baseline classifiers represent a strong reference point concerning the fraud detection performance, yet, the dependence on the fixed representations and the impossibility to conduct forecasting and time-dynamics indicate that more advanced sequential and multi-objective modelling structures are required. The limitations drive the approach proposed, which incorporates the aspects of temporal modelling, risk forecasting, and explainability into a single system.

5.3 Multi-Objective Transformer Performance

The presented multi-objective Transformer model provides high levels of fraud detection with a ROC -AUC of 0.843 and a PR-AUC of 0.264, and at the same time predicts future fraud risk with a low RMSE 0.049. These findings indicate that the model can be used to trade off detection and forecasting ability with a single learning framework.

Table 3. *Performance of the proposed multi-objective Transformer framework*

Metric	Value
Fraud ROC-AUC	0.843
Fraud PR-AUC	0.264
Risk Forecast RMSE	0.049

[Table 3](#) shows a combination of the suggested framework on fraud classification and risk forecasting. The findings indicate a state of equal optimisation of predictive risk estimation and detection accuracy. The average prediction error for the normalized continuous risk score (RMSE) with the proposed framework was 0.049, showing low average prediction error. The maximum and minimum values of the target variable are 1 and 0 respectively which denotes the percentage point values of maximum deviation and minimum deviation between the predicted and actual future fraud risk; an RMSE below 0.05 would mean that the average deviation would be less than 5% of points. The degree of this error indicates that the regression-based branch model is indeed successful in capturing the “fraud propensity” for independent transactions and still performs well for very imbalanced transactions. The high value of the RMSE also supports the model's real-world applicability to proactively monitor the risks and intervene at an early stage to stop fraud.

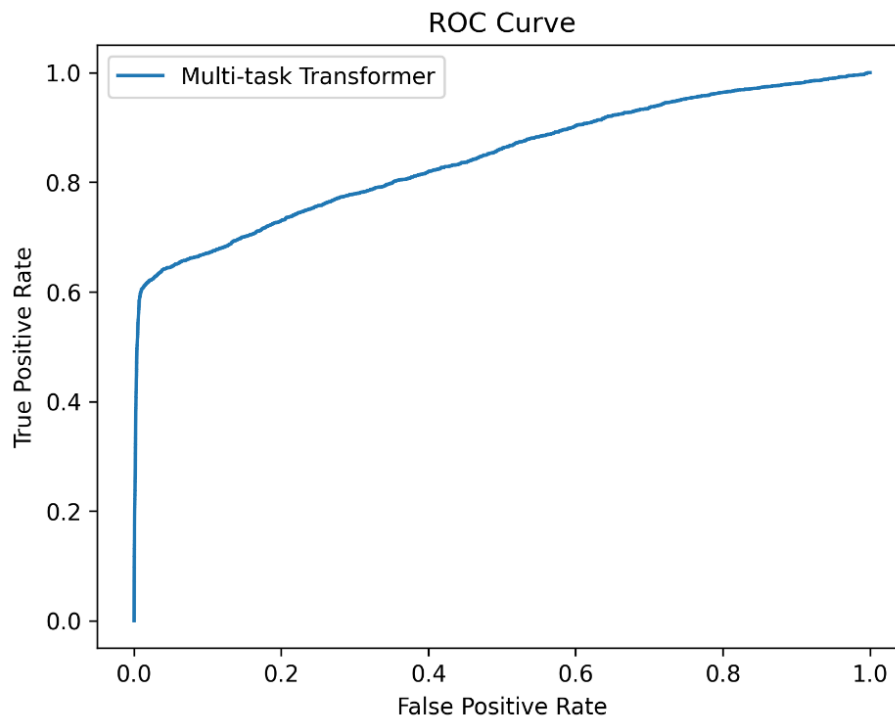


Figure 7: *Comparative ROC Performance of Baseline and Proposed Models*

The performance of the classification with respect to models is provided as a comparative summary in [Figure 7](#). Although random Forest achieves a better ROC-AUC due to its pure attention towards classification, it lacks temporal modeling, forecasting and explainability. On the other hand, the suggested framework provides a complete solution which combines detection, prediction and interpretation. [Figure 8](#) and [Figure 9](#) represent the training dynamics of the proposed model. Loss curve illustrates that convergence is steady and the AUC curve depicts increasing improvement with each training epoch, thus providing evidence of successful optimisation and generalisation.

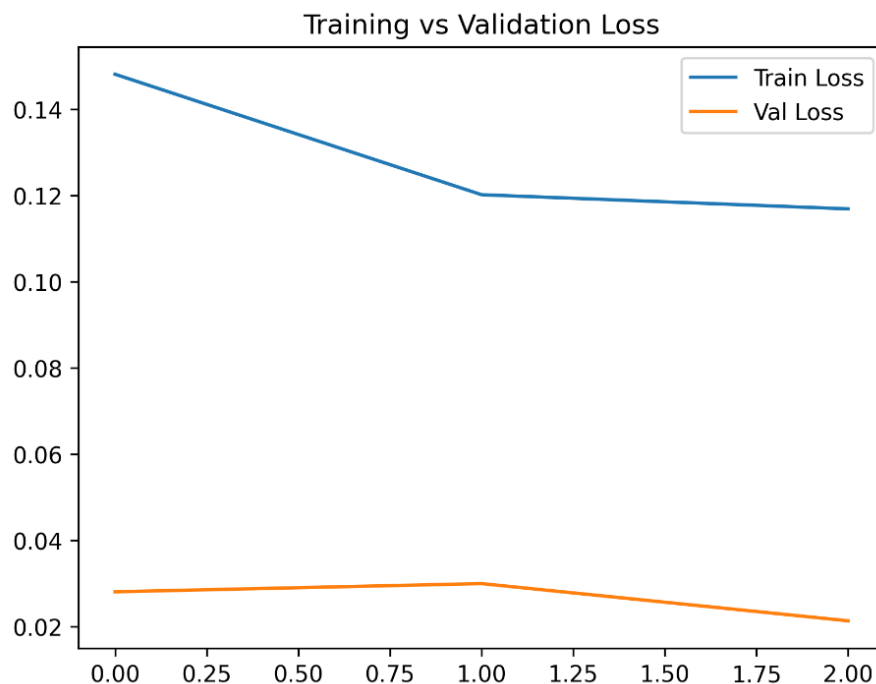


Figure 8: Training Loss Convergence of the Proposed Multi-Objective Transformer Model

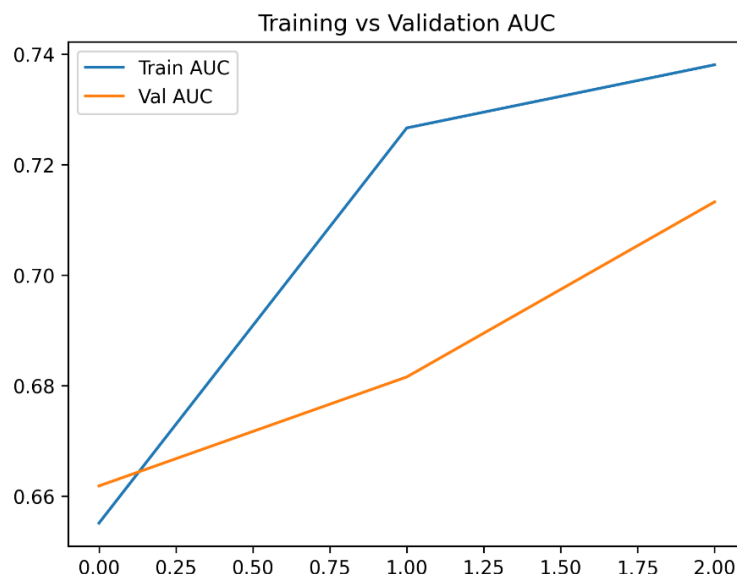


Figure 9: *Training AUC Evolution of the Proposed Multi-Objective Transformer Model*

5.4 Explainability Analysis

The proposed framework involves explainability mechanisms at both feature and time scales to make the framework transparent and facilitate its integration in financial settings. SHAP-based attribution is used to obtain feature-level interpretability by as shown in [Figure 10](#). The SHAP analysis shows that transaction amount and balance-related features always make the most significant contribution to the predictions of fraud.

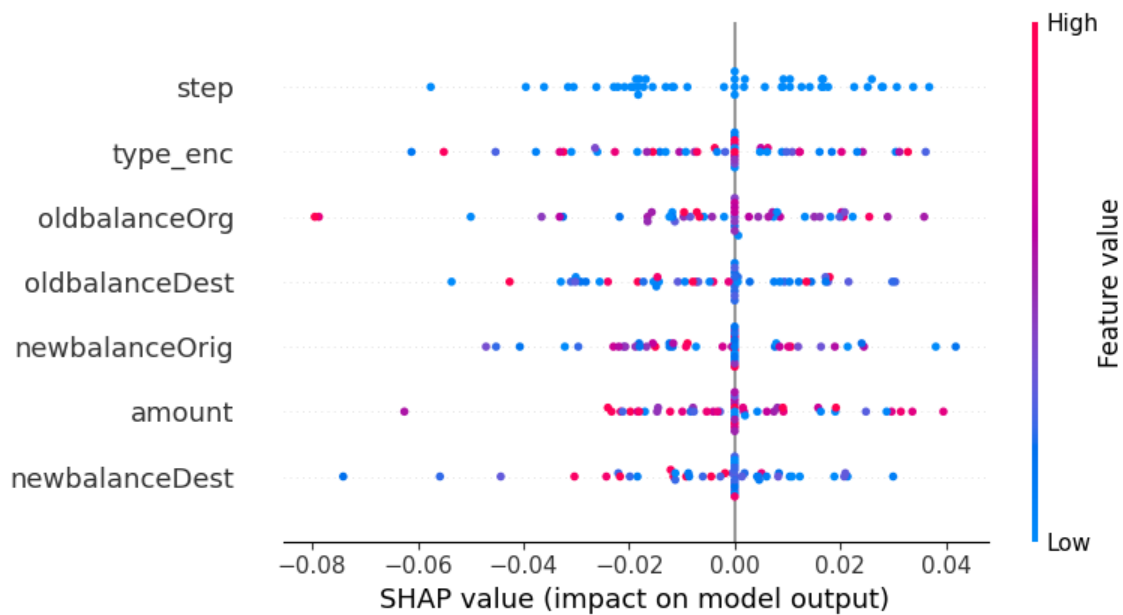


Figure 10: *Feature-Level Explainability of Fraud Predictions Using SHAP Values*

Specifically, original to destination balances variations and destination to original balances variations have high scores on importance, meaning that sudden or imprecise changes in balances are however the main drivers of anomalous behaviour. This observation is consistent with the knowledge in the domain, and during fraudulent transactions, unusual funds flows are frequent to quickly draw value or create money trails. In addition to needed levels of assurance in detecting known red flags of fraud, the SHAP results also give quantitative information on the impact of individual feature on model output. Positive values of SHAP imply that feature contributions are greater and hold an upward prediction on the probability of fraud, whereas negative ones are lesser and hold a downward prediction. The distribution of SHAP values of transactions of high amount is very broad which implies that large monetary transfers increase uncertainty and perception of risk in the model. Nevertheless, the fact that the contributions to legitimate and fraudulent features overlap proves that individual characteristics do not guarantee effective distinguishing in the first case, which is why multivariate and time modelling are essential.

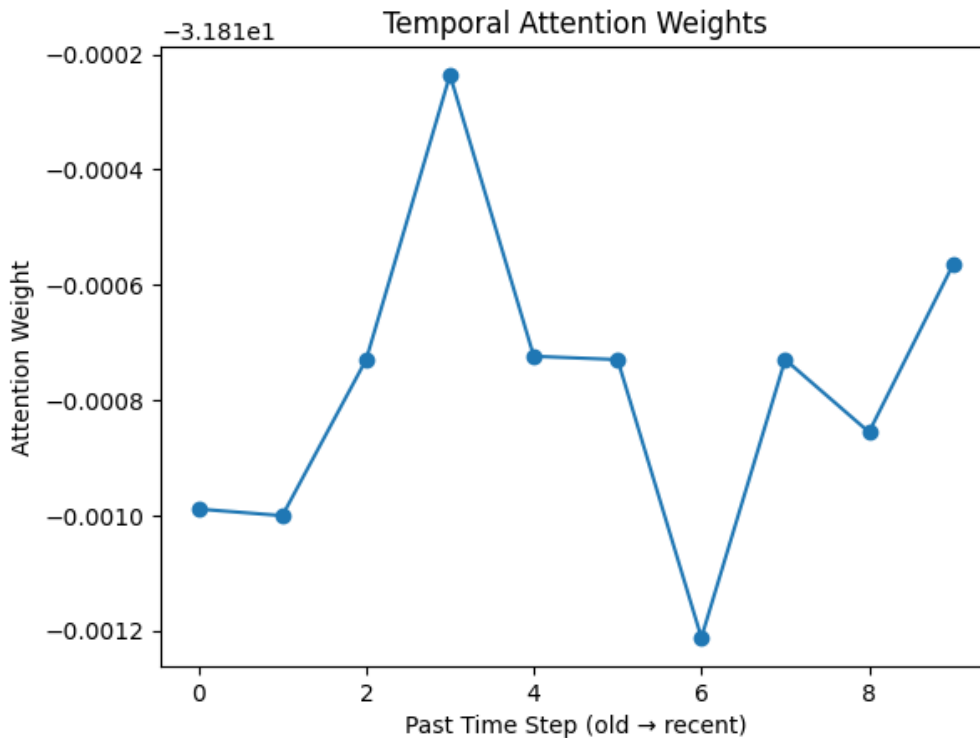


Figure 11: *Temporal Attention Weight Distribution Across Transaction History*

In parallel with interpreting the results of [Figure 11](#), the attention weights were quantitatively evaluated by the average (normalized) attention scores over the evaluation sequences. This leads to a distribution which always gives a high priority to the recent transactions and acts decreasingly for the previous ones, a behaviour that seems to indicate that the Transformer is mainly concerned about short-term dynamics of behavior to help predicting fraud. This quantitative aggregation proves that the observed attention pattern is not a one-time visualization, but rather is a regular-time feature learned by the suggested model by various sequences of transactions.

The attention coefficients from all of the evaluation sequences were aggregated by each of the temporal position by using the mean normalized attention weight of the evaluation sequences so as to quantitatively validate the attention heatmap. The normalized 'attention' score for the t -th time step is calculated as:

$$\bar{A}_t = \frac{1}{N} \sum_{i=1}^N A_{i,t}$$

In the above, $A_{i,t}$ is the normalized weightage of the attention applied to the time-step t in transaction sequence i and N is the number of the transaction sequences evaluated. The aggregated attention distribution is also revealing because it shows that after the early historical observations, the average attention towards recent transactions is higher in all cases when compared to the average attention towards earlier observed transactions, emphasizing the average attention data that the Transformer systematically assigns more attention towards

the short-term behavioral changes, rather than the short-term visual changes. This quantitative validation helps to verify the stability and reliability of the said temporal attention mechanism in the testing set. The normalized attention coefficients were averaged over all the evaluation sequences in order to quantitatively validate the attention mechanism. The averaged attention distribution had a uniform level of high attention for the most recent transaction positions – and decreased the contribution of prior observations as they got earlier. This is an indication that this temporal pattern and attention are a consistent learning phenomenon throughout the entire data set and not a mere visualisation.

Table 4. *Most influential features for fraud prediction based on SHAP attribution*

Rank	Feature	Mean SHAP Value (Relative)
1	Step (Time Index)	Highest
2	Transaction Type (Encoded)	High
3	Old Balance (Origin)	High
4	Old Balance (Destination)	Moderate
5	New Balance (Origin)	Moderate
6	Transaction Amount	Moderate
7	New Balance (Destination)	Lower

As referenced by SHAP attribution, [Table 4](#) ranks transaction attributes in accordance with their share in predicting frauds. Time based position, type of transaction and balance changes turn out to be highly driving factors and this affirms that behavioural timing as well as the abnormal fund movements affect the model choices. SHAP as well as attention-based explanation will provide complementary interpretability when used together. Although SHAP shows what features contribute to the forecasts of fraudulent activity, attention visualisation shows when there are significant behavioural indicators. Collectively these mechanisms can give a consistent clear picture of how decisions are made within the model and allow the analysts to be able to trace their predictions both in terms of transaction attributes as well as the temporal context. This two-tier explainability increases credibility and contributes to compliance with the regulations since it enables financial institutions to explain automated decisions using explainable evidence as opposed to obscure probability scores. On the whole, the explainability discussion shows that the suggested framework is not a black-box predictor, but rather generates explainable and behaviourally consistent fraud judgments, which harmonises model reasoning with financial intuition and operational needs.

5.5 Discussion

The experimental findings prove that the possibilities of classical machine-learning models, especially of the ensemble-based classifiers, have high fraud classification accuracy, but their possibilities are essentially reactive.

Table 5. *Functional comparison of baseline models and the proposed framework*

Model	Temporal Modeling	Risk Forecasting	Explainability	Multi-Objective
Logistic Regression	No	No	Limited	No

Random Forest	No	No	Partial	No
LSTM	Yes	No	No	No
Transformer (single-task)	Yes	No	No	No
Proposed Framework	Yes	Yes	Yes	Yes

[Table 5](#) compares models on the basis of functional capabilities, but not accuracy. It points out how the proposed framework is holistic as opposed to single-task and fixed baselines. They are optimised models designed to detect fraudulent transactions following their occurrence and they can only employ a consistent feature representation, and thus ignore sequential behavioural patterns and risk evolution over time. This, consequently, restricts their relevance in dynamic digital payment settings where proactive intervention and finding early is becoming more crucial. The presented multi-objective explainable Transformer model expands the traditional model of detecting fraud by incorporating a time representation, predicting future risks and interpretability into one framework. Using attention based sequence modelling, the framework has the ability to capture short term deviations of behaviour and changing transaction dynamics which cannot be elucidated by the innermost classifiers. To make the fraud detection even more proactive, a continuous risk forecasting goal is introduced to the detection and then, the development of the fraud detection becomes truly proactive and follows-up, thus willingly managing to convert this tool into an early warning and preventative mechanism.

Ablation Analysis: The input-impact analysis was made using the proposed framework for multi-objective learning strategy and its single-task based Transformer model. The single task model resulted in a slightly better fraud classification ROC-AUC score (0.856) due to its being optimized for binary (or two class) classification. The proposed multi-objective Transformer, on the other hand, achieved a ROC-AUC of 0.843, and also offered an RMSE of 0.049 in forecasting fraud risk in the future. This comparison shows that the reduction of classification even when using the regression objective is not significant, but the extension of the capability of the proactive risk estimation is substantial.

The results when compared show that the system with maximum classification accuracy doesn't necessarily lead to the best and most practicable fraud monitor system. While Random Forest can outperform other classifiers in isolation to discriminate transactions, it can only be used as a single task classifier and does not model temporal transactions, provide future fraud risk analysis or offer integrated explanations. Unlike this, the proposed multi-objective Transformer will embed a shared temporal representation for learning fraud detection together with a forecasting of risk. While it results in a tradeoff of classification performance, the framework is more likely to be appropriate for practical digital payment systems in which early action needs to be taken and where regulatory transparency is desired for this action.



Despite the fact that the proposed model has an average decrease in classification performance compared to that of the Random Forest, it is important to note that, the increased goals of forecasting and interpretability would allow this trade-off. Random Forest is better in ROC-AUC because it prioritises single—they usually maximise classification only on one task, whereas the proposed framework simultaneously learns discriminative and predictive representations. This multi-objective formulation values greater utility in the system level as opposed to their isolated accuracy improvement. Notably, the combined explainability mechanisms allow making a transparent decision by showing the contributions of the features, as well as the time-varying impact, which meets the most essential demands of auditability and regulatory compliance. Practically, the given framework can be presented as a framework with practical insights that are not restricted to binary labels of fraud. Risk scores also help financial institutions to determine the emerging threats, both in terms of directing the resources to monitor and taking early intervention measures. In general, the experimental results confirm that the proposed multi- objective explainable Transformer framework is a valid, transparent, and future-focused solution to the problem of fraud risk forecasting in the contemporary digital payment ecosystem.

6. Conclusion, Limitations, and Future Work

This paper introduces a multi-objective explainable Transformer architecture suitable to predict fraud risks in the context of urban digital payment ecosystem. Unlike other traditional methodologies, which are focused in detecting binary frauds, our framework is explicitly drafted to perform fraud classification and anticipatory risk estimation, thus enabling reactive detection and proactive tracking. The framework takes advantage of Transformer-based temporal modelling and captures the dynamics of sequential transactions and changing patterns of fraud. Moreover, an integrated explainability design which combines SHAP feature attribution and attention heat-maps does provide clear views of both feature-level and temporal-level effects on model predictions. Empirical evidence based on a complete set of transactions proves that the developed strategy provides competitive performance in fraud detection and at the same time makes risk forecasts and interpretive decision support. These results indicate that the framework can be deployed well within the modern financial contexts where transparency, accountability, and early warning should be the primary considerations. Even though it has been effective, a number of limitations still exist. First, the experimental assessment is based on a generated transaction data, and this approach, although being popular, might not be sufficient to capture the complexity and diversity of fraud behaviour in the real world. Second, the computational load of the proposed framework is higher compared to classical machine learning model because of the sequential modelling and attention mechanisms used. Third, the existing risk formulation is based on a deterministic horizon of the future and an unsophisticated aggregation method, which may not adequately capture any long-term, or user-specifics of fraud dynamics. These limitations will be dealt with in future research along various axes. As a preliminary step, the framework will be



tested on practical transaction data to determine its strength and external validity. Two, more advanced risk modelling approaches (including adaptive horizons and survival-based forecasting) will be explored. Third, variations of the model, including lightweight and online variants, will be designed, so that the deployment of the model can be carried out in real-time mobile high-throughput payment systems. Lastly, future research will be done to use graph-based relationships between transactions and modelling user behaviour to improve fraud risk prediction and explainability.

Author Contribution

Dipendra Kumar Air contributed to conceptualization, methodology development, model design, implementation, data analysis, visualization, writing – original draft preparation, and manuscript review and editing.

Transparency Statement: The authors confirm that this study has been conducted with honesty and in full adherence to ethical guidelines.

Funding: The study received no external funding.

Conflict of Interest: The author declares there is no conflict of interest.



References:

- Ahmed, M. P., Arif, M., Al Mamun, A., Mahmud, F., Rahman, T., Ahmmed, M. J., & Uddin, M. K. (2024). A comparative study of machine learning models for predicting customer churn in retail banking. *Journal of Computer Science and Technology Studies*, 6(4), 92–101. <https://doi.org/10.32996/jbms.2024.6.4.12>
- Al-Daoud, K. I., & Abu-AlSondos, I. A. (2025). Robust AI for financial fraud detection in the GCC. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(2), 121. <https://doi.org/10.3390/jtaer20020121>
- Alex-Omiogbemi, A. A., Sule, A. K., Omowole, B. M., & Owoade, S. J. (2024). Advances in cybersecurity strategies for financial institutions. *Journal of Cybersecurity and Financial Innovation*, 12(3), 35–48. <https://doi.org/10.51594/farj.v6i12.1771>
- Alhashmi, A. A., Alashjaee, A. M., Darem, A. A., Alanazi, A. F., & Effghi, R. (2023). An ensemble-based fraud detection model. *Engineering, Technology & Applied Science Research*, 13(6), 12433–12439. <https://doi.org/10.48084/etasr.6268>
- Ansari, I., Azim, K. S., Bhujel, K., Panchal, S. S., & Ahirrao, Y. S. (2025). Fintech innovation and IT infrastructure. *Emerging Frontiers Library for The American Journal of Engineering and Technology*, 7(09), 49–73. <https://doi.org/10.37547/tajet/Volume07Issue09-05>
- Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In *Proceedings of the 28th European Modeling and Simulation Symposium (EMSS)* (pp. 249–255). <https://doi.org/10.46354/i3s/2016/9788>
- Huang, Y., Xu, J., Lai, J., Jiang, Z., Chen, T., Li, Z., & Zhao, P. (2023). Advancing transformer architecture. *arXiv*. <https://arxiv.org/abs/2311.12351>
- Islam, S., Haque, M. M., & Karim, A. N. M. R. (2024). A rule-based machine learning model for financial fraud detection. *International Journal of Electrical & Computer Engineering*, 14(1). <https://doi.org/10.11591/ijece.v14i1.pp759-771>
- Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Angulakshmi, M. (2025). Applications of explainable artificial intelligence (XAI). *IEEE Access*. <https://ieeexplore.ieee.org/document/10908240>
- Khanum, A., Chaitra, K. S., Singh, B., & Gomathi, C. (2024). Fraud detection in financial transactions. In *2024 International Conference on IITCEE* (pp. 1–5). *IEEE*. <https://doi.org/10.1109/IITCEE59897.2024.10467759>
- Leevy, J. L., Hancock, J., Khoshgoftaar, T. M., & Abdollah Zadeh, A. (2023). One-class and binary classification for fraud detection. *Journal of Big Data*, 10(1), 157. <https://doi.org/10.1186/s40537-023-00769-4>
- Njoku, D. O., Iwuchukwu, V. C., Jibiri, J. E., Ikwuazom, C. T., Ofoegbu, C. I., & Nwokoma, F. O. (2024). Machine learning approach for fraud detection. *Machine Learning*, 20(4), 1–12. <https://www.ijerd.com/v20-i4.html>



- Oko-Odion, C. (2025). AI-driven risk assessment models. *International Journal of Computer Applications Technology and Research*, 14(04), 80–96. <https://doi.org/10.7753/IJCATR1404.1007>
- Ortigosa, E. S., Gonçalves, T., & Nonato, L. G. (2024). Explainable artificial intelligence (XAI). *IEEE Access*, 12, 80799–80846. <https://doi.org/10.1109/ACCESS.2024.3392191>
- Pratama, S. F., & Wahid, A. M. A. (2025). Fraudulent transaction detection. *Journal of Cyber Law*, 1(1), 88–115. <https://jcl.mbicore.com/article/id/58/>
- Qin, G., Feng, Y., & Van Durme, B. (2023). NLP task effectiveness of long-range transformers. In *EACL 2023 Proceedings* (pp. 3774–3790). <https://doi.org/10.18653/v1/2023.eacl-main.275>
- Ramachandran, M. (2025). AI ethics for explainability. In *Engineering Ethics of AI by Design* (pp. 711–752). Springer. https://doi.org/10.1007/978-981-99-XXXX-X_24
- Rane, N., Choudhary, S., & Rane, J. (2023). Blockchain and AI integration for finance. SSRN. <https://doi.org/10.2139/ssrn.4644253>
- Sebt, M. V., Ghasemi, S. H., & Mehrkian, S. S. (2021). Predicting customer transactions using stacked LSTM. *Social Network Analysis and Mining*, 11(1), 86. <https://doi.org/10.1007/s13278-021-00789-4>
- Wiggerthale, J., & Reich, C. (2024). Explainable machine learning in critical decision systems. *AI*, 5(4), 2864–2896. <https://doi.org/10.3390/ai5040138>
- Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., & Wang, H. (2024). Large language models for cyber security. *ACM Transactions on Software Engineering and Methodology*. <https://doi.org/10.1145/3769676>

Views and opinions expressed in this article are the views and opinions of the author(s), *NPRC Journal of Multidisciplinary Research* shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.