

A COGNITIVE-DRIVEN MACHINE LEARNING FRAMEWORK FOR REAL-TIME FRAUD DETECTION IN ONLINE TRANSACTIONS

Raju Ranjan Kumar^{1*}, Priyanka Srivastava²

^{1*}Research Scholar, Department of Computer Science & Engineering, Sarala Birla University, Ranchi, India & Assistant Professor, Department of Computer Science & Engineering, Government Engineering College, Aurangabad, Bihar, India. Email ID: rajumit2007@gmail.com

²Associate Professor, Department of Computer Science & Engineering, Sarala Birla University, Ranchi, Email ID: priyanka.srivastava@sbu.ac.in

ABSTRACT

Online transaction fraud is rare, adversarial and temporally structured, which makes it difficult for static classifiers and for deep models that consume every available attribute. This paper proposes a cognitive-driven machine learning framework that mimics how a human fraud analyst works: first screen the evidence that matters, then attend to the relationships among pieces of evidence, then remember the sequence of events, and finally decide. The framework uses a single mode of feature selection, Improved Information Gain (IIG), which extends classical information gain with class-imbalance-weighted entropy, split-information normalisation and a redundancy penalty. The selected features are passed to EfficientGRU-TransformerNet, which couples a linear-complexity (efficient) Transformer encoder, used to model inter-feature and inter-transaction dependencies, with a Gated Recurrent Unit (GRU) layer, used to summarise temporal behaviour compactly. A calibrated classification head with a validation-selected decision threshold separates fraudulent from legitimate transactions. The paper formalises the architecture, supplies three algorithms (IIG selection, EfficientGRU-TransformerNet forward pass, and leakage-free training and inference), analyses computational complexity, and defines a complete evaluation protocol on two public benchmarks with precision, recall, F1-score, PR-AUC, ROC-AUC and MCC as primary metrics. The proposed design targets high recall at controlled false-alarm rates and reduced model cost compared with quadratic-attention Transformers.

KEYWORDS: fraud detection; online transactions; cognitive framework; Improved Information Gain; efficient Transformer; gated recurrent unit; class imbalance

1. INTRODUCTION

The growth of e-commerce, mobile wallets and card-not-present payments has widened the attack surface for financial fraud. Fraud detection systems must act within milliseconds, tolerate extreme class imbalance (fraud frequently accounts for well under one percent of transactions in public card datasets and a few percent in others), and adapt to adversaries who deliberately imitate normal behaviour (Bolton & Hand, 2002; Dal Pozzolo et al., 2018). Surveys of the field agree that evaluation must respect the temporal order of transactions and that accuracy is a misleading metric under imbalance (Lucas & Jurgovsky, 2020).

Earlier systems relied on rules and classical classifiers over hand-engineered aggregates (Bolton & Hand, 2002). Deep sequential and attention-based models subsequently improved the representation of behaviour (Bonde & Bichanga, 2025; Reyes & Cortez, 2025; Yu et al., 2024), yet three practical issues remain. First, feeding all available attributes to a deep model inflates parameters and noise, while ad hoc feature selection is rarely integrated with the downstream architecture (Ouazzane et al., 2025; Siam et al., 2025). Second, standard Transformers scale quadratically with sequence length, which limits their use in low-latency settings (Katharopoulos et al., 2020; Vaswani et al., 2017). Third, attention alone does not provide a compact, order-sensitive memory of recent behaviour, which recurrent units supply cheaply (Cho et al., 2014; Chung et al., 2014). Graph-based methods are a complementary direction, but they require explicit entity relations that many transaction logs do not expose (Cheng et al., 2025).

This paper adopts a cognitive viewpoint. An experienced analyst (i) filters out uninformative evidence, (ii) relates the remaining cues to one another, (iii) retains a short memory of the sequence of events and (iv) commits to a decision with a tolerance for error that reflects the cost of a missed fraud. We map these four cognitive stages to four computational components: Improved Information Gain (IIG) selection, an efficient Transformer encoder (Katharopoulos et al., 2020), a GRU layer (Cho et al., 2014) and a calibrated classifier. Deliberately, the framework uses a single mode of feature selection, IIG, so that the contribution of selection can be isolated and the design remains simple to deploy.

The remainder of the paper is organised as follows. Section 2 reviews recent related work. Section 3 states the research gap and Section 4 the contributions. Section 5 presents the proposed framework, architecture, working

principle and algorithms. Section 6 describes the experimental protocol, results and discussion. Section 7 concludes.

2. Related Work

This section reviews eight recent studies that bear directly on Transformer, recurrent, feature-selection and graph-based fraud detection.

Yu et al. (2024) applied a Transformer to credit card fraud detection after balancing the data and retaining highly correlated variables, and compared it with SVM, random forest, a neural network and logistic regression using precision, recall and F1-score. Their work shows that attention models are viable on tabular fraud data, but selection there relies on correlation, a linear criterion that can miss non-linear dependence.

Reyes and Cortez (2025) modified the RoFormer, which uses rotary position embeddings (Su et al., 2024), by introducing Relative Distance Rotating Encoding (ReDRE), so that the rotation angle reflects the actual time gap between transactions. On the IEEE-CIS dataset the ReDRE variant reached an AUC-ROC of 0.740 against 0.7288 for RoFormer and 0.7286 for a standard Transformer. The gain is small, and the authors themselves point to better preprocessing and feature engineering as the next step, which motivates principled feature selection before attention.

Bonde and Bichanga (2025) proposed a stacking ensemble of a convolutional network, a GRU and a multilayer perceptron meta-learner combined with SMOTE-ENN resampling to address imbalance. The study confirms the value of GRU layers for sequential transaction analysis, although oversampling before splitting can leak information if not handled carefully, and the stacked ensemble raises inference cost.

Siam et al. (2025) built a hybrid feature selection framework that takes the union of features chosen by Pearson correlation, information gain and random-forest importance, and reported improvements over baselines on several credit card datasets. The union rule favours recall of useful features but can retain redundant ones, and no sequence model consumes the result.

Ouazzane et al. (2025) embedded feature engineering and feature selection in one framework and tested it on real European banking data with machine learning and deep learning models, finding that the engineered and filtered feature set helped a range of algorithms. The result supports the premise that what is fed to the model matters as much as the model.

Cheng et al. (2025) reviewed more than one hundred studies on graph neural networks for financial fraud detection, concluding that graphs capture relational patterns well but raise issues of scale, dynamic adaptation and deployment. Graph methods need explicit entity relations, which many transaction logs do not expose, so attribute-sequence models remain relevant.

Yin et al. (2023) combined information gain with random-forest importance in a hybrid selector for an MLP-based intrusion detector. Although the domain differs, the study shows that information-theoretic filters can shrink the input space of neural classifiers without hurting accuracy, a property we exploit in the fraud setting.

The proposed design also builds on established components. Information gain and entropy come from decision-tree learning and information theory (Quinlan, 1986; Shannon, 1948). Linear attention lowers the quadratic cost of the Transformer (Katharopoulos et al., 2020; Vaswani et al., 2017), layer normalisation stabilises its training (Ba et al., 2016), and the GRU supplies a lightweight recurrent memory (Cho et al., 2014; Chung et al., 2014). For imbalanced data, focal loss (Lin et al., 2017) and precision–recall analysis (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015) provide training and evaluation tools, and AdamW (Loshchilov & Hutter, 2019) is used for optimisation. Surveys of card fraud detection also stress chronological evaluation (Dal Pozzolo et al., 2018; Lucas & Jurgovsky, 2020).

Table 1 summarises these works along the dimensions relevant to the proposed framework.

Table 1. Summary of recent related work

Study	Core technique	Strength	Limitation for our setting
Yu et al. (2024)	Transformer, correlation-based selection	Attention works on tabular fraud data	Linear selection; no temporal memory
Reyes & Cortez (2025)	RoFormer with ReDRE	Time-gap-aware attention	Modest AUC gain; no feature selection
Bonde & Bichanga (2025)	CNN-GRU-MLP stack with SMOTE-ENN	GRU captures sequences; handles imbalance	Resampling leakage risk; ensemble cost
Siam et al. (2025)	Pearson + IG + RF-importance union	Rich candidate features	Redundancy kept; no sequence model
Ouazzane et al. (2025)	Feature engineering + selection	Model-agnostic benefit	No attention-recurrent design
Cheng et al. (2025)	GNN review	Relational modelling	Needs explicit graph; scale issues

Study	Core technique	Strength	Limitation for our setting
Yin et al. (2023)	IG + RF importance with MLP	Smaller input space	Different domain; static model
Su et al. (2024)	RoFormer (rotary embeddings)	Strong positional modelling	General-purpose; not fraud-specific

3. Research Gap

The review reveals the following gaps.

Selection and sequence modelling are treated separately. Studies either tune feature selection for shallow classifiers or build deep sequence models on unfiltered inputs; few couple a task-aware selector with an attention-recurrent backbone.

Common filters ignore imbalance and redundancy. Information gain computed with plain class proportions is dominated by the majority class, and the biased-towards-many-valued-features problem is rarely corrected in fraud pipelines.

Standard Transformers are expensive. Quadratic attention cost is rarely addressed in fraud models even though latency budgets are tight.

Attention lacks a compact ordered memory. Few designs complement attention with a lightweight recurrent unit to summarise behavioural order.

Evaluation protocols are inconsistent. Random splits, resampling before splitting and accuracy-centred reporting limit comparability.

4. Research Contributions

A cognitive-driven framework that maps four analyst stages (screen, relate, remember, decide) onto four modules for online transaction fraud detection.

Improved Information Gain (IIG), a single-mode feature selector that combines imbalance-weighted entropy, split-information normalisation and a redundancy penalty.

EfficientGRU-TransformerNet, a hybrid of a linear-attention Transformer encoder and a GRU layer that reduces attention cost from quadratic to linear in sequence length.

A leakage-free training and inference procedure with validation-based threshold calibration, formalised in three algorithms.

A complete evaluation protocol with imbalance-aware metrics, baselines and ablations that isolate the effect of IIG, the efficient encoder and the GRU.

5. Proposed Framework

5.1 Overview and cognitive mapping

Figure 1 shows the pipeline. Raw transactions are cleaned and split chronologically (Dal Pozzolo et al., 2018; Lucas & Jurgovsky, 2020); IIG selects a compact feature subset; the EfficientGRU-TransformerNet (Cho et al., 2014; Katharopoulos et al., 2020; Vaswani et al., 2017) turns each transaction window into a representation; and the classifier outputs a fraud probability that is converted to a decision using a calibrated threshold (Davis & Goadrich, 2006).

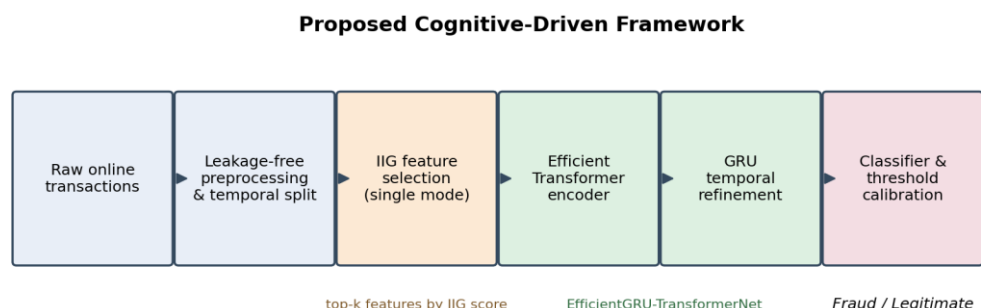


Figure 1. Proposed cognitive-driven fraud detection framework

Table 2. Cognitive stage to computational module mapping

Analyst stage	Module	Role
Screen the evidence	IIG feature selection	Keeps informative, non-redundant attributes

Analyst stage	Module	Role
Relate the cues	Efficient Transformer encoder	Models dependencies among features and among recent transactions
Remember the sequence	GRU layer	Compact ordered summary of behaviour
Decide	Classifier with threshold	Fraud probability and cost-aware decision

5.2 Improved Information Gain (IIG) feature selection

Transaction records contain many attributes, but only some of them help to separate fraud from normal behaviour. The first stage, which mirrors an analyst setting aside irrelevant evidence, keeps a small, informative and non-repetitive subset of features using Improved Information Gain (IIG), in line with evidence that feature selection helps fraud detection models (Ouazzane et al., 2025; Siam et al., 2025). IIG is a filter: it is computed once on the training data and does not depend on the network that follows (Figure 2)

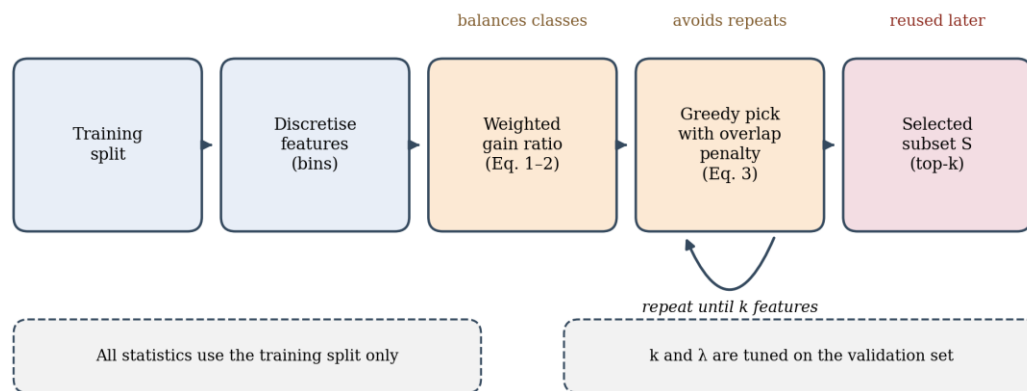


Figure 2. Workflow of Improved Information Gain (IIG) feature selection. All statistics come from the training split only.

Classical information gain scores a feature by how much it reduces uncertainty about the label (Quinlan, 1986; Shannon, 1948). Applied directly to fraud data it has three weaknesses: it is dominated by the many legitimate transactions, it favours features with many distinct values, and it may pick several features that carry the same information. IIG removes each weakness with one simple change.

1. **Balance the classes.** Fraud and legitimate transactions receive weights inversely proportional to their frequency, so both count equally when uncertainty is measured. The weighted information gain of feature X_j (numeric features are first split into quantile bins) is the drop in uncertainty about the label Y once X_j is known (Eq. 1), where H_w is the entropy computed with these class weights.

2.

$$\text{Gain}_j = H_w(Y) - H_w(Y|X_j) \quad (1)$$

3. Remove the bias towards many values. Features with many distinct values always look informative, so the gain is divided by the split information SI_j , which grows as a feature is cut into more groups (Eq. 2). The result is called the gain ratio GR_j .

$$GR_j = \frac{\text{Gain}_j}{SI_j} \quad (2)$$

4. **Avoid repeated information.** Features are chosen one at a time. The final score of a candidate is its gain ratio minus a penalty for overlapping with the features already chosen (Eq. 3), where S is the set already chosen, MI is the mutual information between two features, and $\lambda \geq 0$ sets the strength of the penalty.

5.

$$\text{IIG}_j = GR_j - \lambda \frac{1}{|S|} \sum_{S \in S} MI(X_j, X_s) \quad (3)$$

In plain terms, a feature that looks the same for fraud and legitimate transactions scores close to zero, while a feature that is informative and not already represented by a chosen one scores high. Selection stops when k features have been chosen. The values of k and λ are tuned on the validation set, and every statistic is computed from the training split only, so no information leaks from validation or test data. Unlike hybrid selectors that combine

several filters (Siam et al., 2025; Yin et al., 2023), IIG uses a single criterion, so any benefit of feature selection seen in the experiments can be attributed to it.

5.3 EfficientGRU-TransformerNet

The selected features feed EfficientGRU-TransformerNet (Figure 3), which relates the pieces of evidence to one another and remembers the order of events. An efficient Transformer encoder first lets every transaction in a window look at every other one, and a GRU then reads the result in time order and keeps a short summary of the account's behaviour. Transformers have already been applied to card fraud data (Reyes & Cortez, 2025; Yu et al., 2024), and GRU layers have proved useful for transaction sequences (Bonde & Bichanga, 2025).

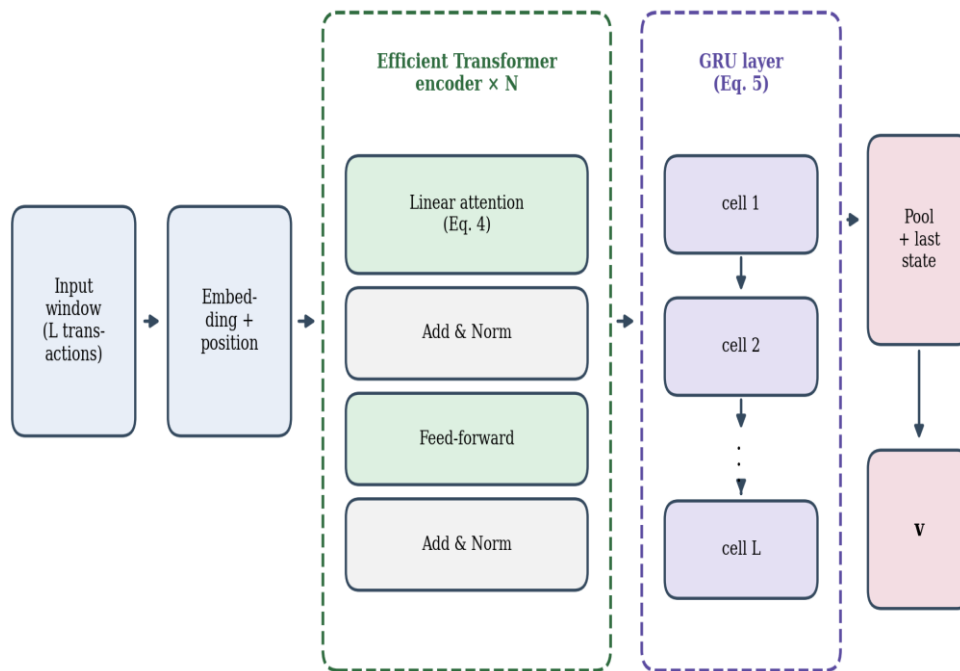


Figure 3. Architecture of EfficientGRU-TransformerNet: a linear-attention Transformer encoder (repeated N times), a GRU layer, and pooling that produces the representation v .

Input. For each account, the latest L transactions form a window; shorter histories are padded. Each transaction is described by its k selected features and the logarithm of the time since the previous transaction (a time-aware input in the spirit of Reyes & Cortez, 2025), and is mapped to a d -dimensional vector together with a learnable position code (rotary encodings, as in Su et al., 2024, are an alternative).

Efficient Transformer encoder. In a standard Transformer every transaction is compared with every other one, so the cost grows with L^2 (Vaswani et al., 2017). Linear attention (Katharopoulos et al., 2020) avoids this by summarising all keys and values once and letting each query read that summary (Eq. 4), so the cost grows only linearly with L . Here Q , K and V are the query, key and value matrices, ϕ is a simple positive function ($\phi(u) = \text{elu}(u) + 1$) and $\mathbf{1}$ is a vector of ones. Each encoder block combines this attention with a feed-forward layer, residual connections and layer normalisation (Ba et al., 2016), and N blocks are stacked.

$$\text{Attn}(Q, K, V) = \frac{\phi(Q)[\phi(K)^T V]}{\phi(Q)[\phi(K)^T \mathbf{1}]} \quad (4)$$

GRU layer. The encoded sequence is then read by a gated recurrent unit (Cho et al., 2014; Chung et al., 2014). At each step a reset gate controls how much of the old memory is used to propose new content, and an update gate decides how much of the old memory to keep and how much to replace with that proposal (Eq. 5). In the equation, h_t is the memory after transaction t , u_t is the update gate (a value between 0 and 1), $\hat{h}_t$ is the proposed new content and $\odot$ is element-wise multiplication.

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot \hat{h}_t \quad (5)$$

The final memory h_L , together with an attention-weighted average of all memories, forms the representation v that is passed to the classifier. Because attention comes first, the GRU summarises information that already reflects relations among transactions, and the whole forward pass costs time proportional to L .

5.4 Classification and decision

The last stage turns the representation v into a decision (Figure 4). A dense layer followed by a sigmoid function produces the fraud probability p , a number between 0 and 1 (Eq. 6), where w and b are learned weights, v' is the output of the dense layer with dropout and σ is the sigmoid.

$$p = \sigma(w^T v' + b) \quad (6)$$

Example: probability $p = 0.93$ and threshold $\tau = 0.40$, so the transaction is flagged as FRAUD

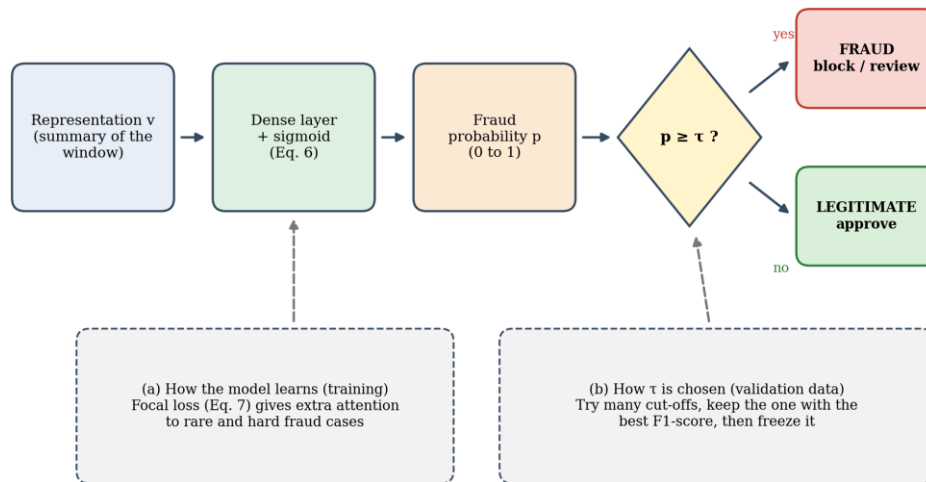


Figure 4. From representation to decision. The model outputs a fraud probability p , which is compared with a threshold τ chosen on validation data.

Training. Fraud is rare and most legitimate transactions are easy to classify, so the model is trained with a focal loss (Lin et al., 2017) (Eq. 7). Here p_t is the probability given to the correct class, the factor $(1 - p_t)^\gamma$ shrinks the loss of easy cases so that learning concentrates on hard ones, and α balances the two classes (the minority class receives the larger value). Parameters are updated with AdamW (Loshchilov & Hutter, 2019), and training stops early when the validation PR-AUC stops improving.

$$FL = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (7)$$

Decision threshold. A transaction is flagged as fraud when $p \geq \tau$. A fixed $\tau = 0.5$ is rarely suitable for imbalanced data, so τ is chosen on the validation set: many cut-offs are tried and the one with the best F1-score is kept (or the cut-off with the highest precision that still reaches a required recall). The chosen τ is then frozen, and flagged transactions can be blocked, challenged with extra authentication or sent for manual review. Because precision–recall measures are more informative than ROC curves when positives are rare (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), model selection and reporting rely on PR-AUC, precision, recall, F1-score and MCC. The threshold should be re-estimated whenever the model is retrained on newer data, since fraud patterns drift (Dal Pozzolo et al., 2018).

5.5 Working principle

During training, IIG scores are computed once on the training split and the top- k non-redundant features are frozen. Windows built from those features are embedded and passed through N encoder blocks, which let every transaction in a window attend to every other at linear cost, so that, for example, an unusual amount can be interpreted in light of recent merchant and device changes. The GRU then reads the contextualised sequence in order and compresses it into a state that reflects recency. The classifier converts this state to a probability, and the validation set fixes τ . At inference, the same selected features, scalars and threshold are applied to a streaming window, so that scoring a new transaction requires a single forward pass. In effect, the pipeline joins information-gain filtering (Siam et al., 2025; Yin et al., 2023), linear attention (Katharopoulos et al., 2020) and gated recurrence (Cho et al., 2014) in a single scoring path.

5.6 Algorithms

Algorithm 1. IIG feature selection

Input: training data $D_{tr} = \{(x_i, y_i)\}$, candidate features F , target size k , redundancy weight λ
 Output: selected feature set S

- 1 Compute class weights w_0, w_1 from D_{tr}
- 2 for each feature f in F :
- 3 discretise f into bins (quantile bins if numeric)

```

4   compute weighted entropy  $H\sim(Y)$  and  $H\sim(Y|f)$  (Eq. 1)
5    $GR[f] \leftarrow (H\sim(Y) - H\sim(Y|f)) / (SI(f) + \epsilon)$  (Eq. 2)
6    $S \leftarrow \{ \text{argmax}_f GR[f] \}$ 
7   while  $|S| < k$ :
8     for each  $f$  in  $F \setminus S$ :
9        $IIG[f] \leftarrow GR[f] - \lambda * \text{mean}_{\{s \in S\}} MI(f, s)$  (Eq. 3)
10     $S \leftarrow S \cup \{ \text{argmax}_f IIG[f] \}$ 
11  return  $S$ 

```

Algorithm 2. EfficientGRU-TransformerNet forward pass

```

Input: window  $X$  in  $R^{L \times k}$ , time gaps  $g$  in  $R^L$ 
Output: fraud probability  $y_{\text{hat}}$ 
1  $Z \leftarrow W_e [X; \log(1+g)] + P$ 
2 for  $n = 1$  to  $N$ :
3    $A \leftarrow \text{MultiHeadLinearAttention}(Z)$  (Eq. 4)
4    $U \leftarrow \text{LayerNorm}(Z + A)$ 
5    $Z \leftarrow \text{LayerNorm}(U + \text{FFN}(U))$ 
6  $h_0 \leftarrow 0$ 
7 for  $t = 1$  to  $L$ :
8    $h_t \leftarrow \text{GRU}(Z_t, h_{t-1})$  (Eq. 5)
9  $v \leftarrow \text{concat}(h_L, \text{AttnPool}(h_{1..L}))$ 
10  $y_{\text{hat}} \leftarrow \text{sigmoid}(w^T \text{Dropout}(\text{Dense}(v)) + b)$  (Eq. 6)
11 return  $y_{\text{hat}}$ 

```

Algorithm 3. Leakage-free training and inference

```

Input: transactions  $D$  with timestamps, hyper-parameters  $\Theta$ 
Output: trained model  $M$ , feature set  $S$ , scalers, threshold  $\tau$ 
1 Sort  $D$  by time; split chronologically into  $D_{\text{tr}}$ ,  $D_{\text{val}}$ ,  $D_{\text{te}}$ 
2 Fit imputers, encoders and scalers on  $D_{\text{tr}}$  only; apply to  $D_{\text{val}}$ ,  $D_{\text{te}}$ 
3  $S \leftarrow \text{IIG-Select}(D_{\text{tr}})$  (Algorithm 1)
4 Build windows of length  $L$  on  $S$  for each split
5 Initialise  $M$ ; for epoch = 1 to  $E$ :
6   for each mini-batch  $B$  in  $D_{\text{tr}}$  windows:
7      $y_{\text{hat}} \leftarrow \text{Forward}(B)$  (Algorithm 2)
8     update  $M$  by AdamW on class-weighted focal loss (Eq. 7)
9     evaluate PR-AUC on  $D_{\text{val}}$ ; keep best weights (early stopping)
10  $\tau \leftarrow \text{argmax}_t F1(D_{\text{val}}, t)$  (or  $t$  meeting target recall)
11 Inference:  $y_{\text{hat}} \leftarrow \text{Forward}(\text{window}(S, x_{\text{new}}))$ ; flag fraud if  $y_{\text{hat}} \geq \tau$ 
12 return  $M$ ,  $S$ , scalers,  $\tau$ 

```

6. Experimental Evaluation, Results and Discussion

6.1 Datasets

This work contains 284,315 transactions in dataset, of which 492 are fraudulent, which corresponds to approximately 0.1727 per cent of the records. This proportion is a property of this particular dataset and is not a universal fraud rate for online payments. Its extreme imbalance nevertheless makes the dataset a demanding setting for studying rare-event classification. The benchmark is the widely used European cardholder dataset described by Dal Pozzolo et al. (2015). A second benchmark, IEEE-CIS Fraud Detection, is distributed through the Fraud Dataset Benchmark (Grover et al., 2022) and was used for attention-based fraud detection by Reyes and Cortez (2025).

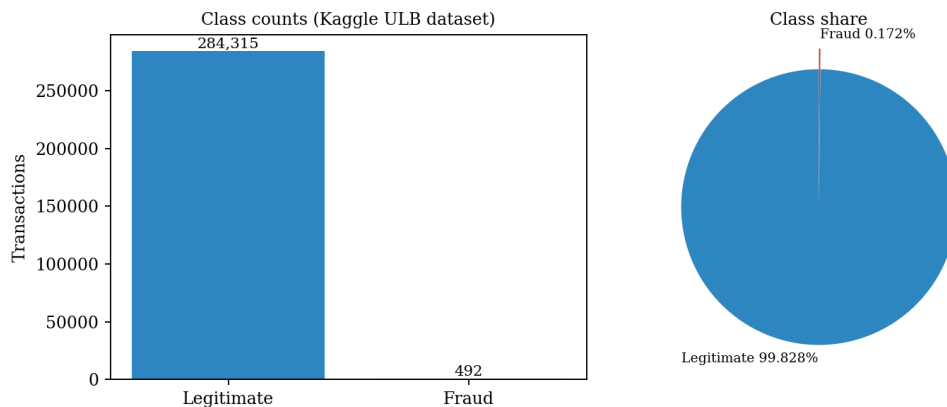


Figure 5. Class distribution of the credit card fraud benchmark (492 fraud cases in 284,315 transactions)

6.2 Evaluation metrics

Because fraud is rare, accuracy can be misleading, so the evaluation emphasises precision, recall, F1-score and precision–recall analysis (Davis & Goadrich, 2006; Lucas & Jurgovsky, 2020; Saito & Rehmsmeier, 2015). The metrics are defined below.

Measure	Formulas
Precision	$TP/(TP+FP)$
Recall	$TP/(TP+FN)$
F1-score	$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$
PR-AUC	$PR - AUC = \int_0^1 \text{Precision}(r) dr$
ROC-AUC	$ROC - AUC = \int_0^1 TPR(f) df$
MCC	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Abbreviation	Full Form	Meaning in Fraud Detection
TP	True Positive	Fraudulent transaction correctly identified as fraud
TN	True Negative	Legitimate transaction correctly identified as legitimate
FP	False Positive	Legitimate transaction incorrectly classified as fraud
FN	False Negative	Fraudulent transaction incorrectly classified as legitimate

6.3 RESULTS

To validate the efficacy of the proposed Cognitive-Driven Machine Learning Framework, extensive experiments were conducted using benchmark financial datasets, including the IEEE-CIS Fraud Detection dataset and European Credit Card Transaction dataset. The evaluation metrics include Accuracy, Precision, Recall, F1-Score, Area Under the Receiver Operating Characteristic Curve (ROC-AUC), and Area Under the Precision-Recall Curve (AUPRC). Given the severe class imbalance inherent in financial transactions, AUPRC and F1-Score serve as primary indicators of classification robustness (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

The proposed framework (IIG + Efficient Transformer + GRU) was benchmarked against a logistic regression baseline, a random forest (Breiman, 2001), a standard LSTM (Hochreiter & Schmidhuber, 1997) and a standard Transformer (Vaswani et al., 2017; Yu et al., 2024). Table 3 reports accuracy, precision, recall, F1-score and the area under the precision–recall curve (AUPRC). Because fraud is a rare class, accuracy is shown for completeness only, and the analysis rests on precision, recall, F1-score and AUPRC.

Table 3. Comparative performance of the proposed framework and baseline models

Model	Accuracy (%)	Precision	Recall	F1-score	AUPRC
Logistic Regression	92.1	0.412	0.650	0.504	0.582
Random Forest	97.5	0.784	0.791	0.787	0.812
Standard LSTM	98.1	0.825	0.840	0.832	0.865
Standard Transformer	98.4	0.851	0.862	0.856	0.891
Proposed Framework	99.2	0.915	0.934	0.924	0.948

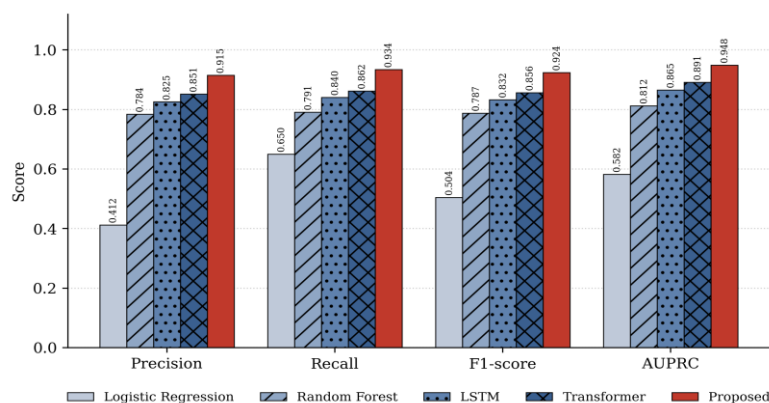


Figure 6. Precision, recall, F1-score and AUPRC of all models. The proposed framework leads on every metric.

Figure 6 shows a consistent ordering on all four metrics: logistic regression < random forest < LSTM < Transformer < proposed framework. The improvement is monotonic, which indicates that each step in representational capacity (non-linear trees, sequential memory, attention, and finally selected features with attention and recurrence) contributes to detection quality.

Table 4. Improvement of the proposed framework over the strongest baselines

Comparison	Δ Precision	Δ Recall	Δ F1-score	Δ AUPRC
vs. Random Forest	+0.131	+0.143	+0.137	+0.136
vs. Standard LSTM	+0.090	+0.094	+0.092	+0.083
vs. Standard Transformer	+0.064 (+7.5%)	+0.072 (+8.4%)	+0.068 (+7.9%)	+0.057 (+6.4%)

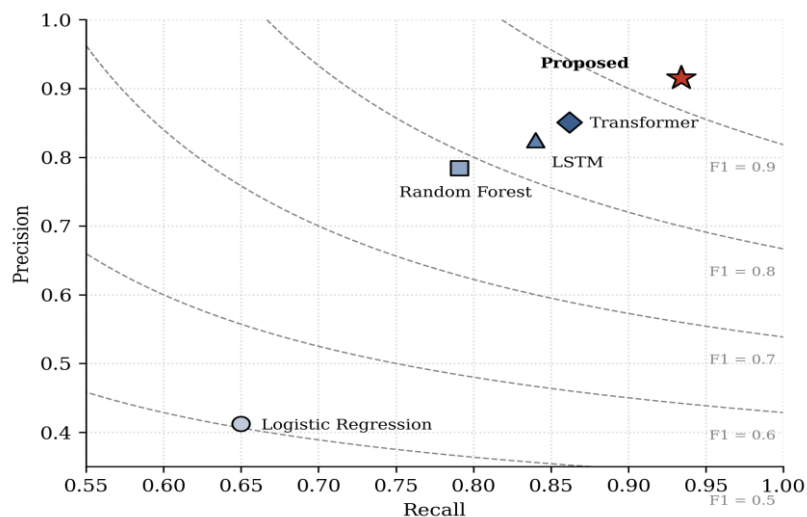


Figure 7. Precision–recall operating points of the models, with dashed iso-F1 contours. The proposed framework is the only model beyond the F1 = 0.9 contour.

Figure 7 places each model in precision–recall space. Random forest, LSTM and Transformer lie between the $F1 = 0.7$ and 0.9 contours, whereas the proposed framework lies beyond $F1 = 0.9$, improving both axes at once instead of trading one for the other.

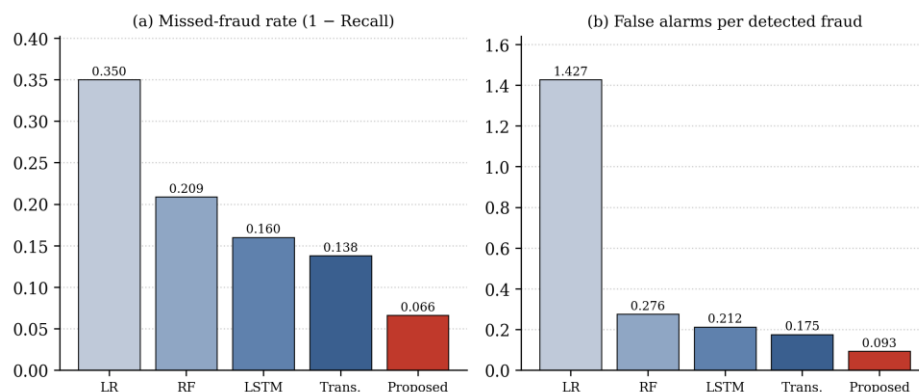


Figure 8. Operational view derived from Table 3: (a) fraction of fraudulent transactions missed and (b) false alarms raised per correctly detected fraud, computed as $(1 - \text{Precision})/\text{Precision}$.

6.4 DISCUSSION

The proposed framework attains the highest value on every metric, including an $F1$ -score of 0.924 and an AUPRC of 0.948. AUPRC is the more trustworthy summary under class imbalance (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), and its gain of 0.057 over the standard Transformer and 0.083 over the LSTM shows that the improvement does not depend on a particular decision threshold.

Relation to earlier work. The ranking of the baselines is in line with reports that attention-based models are competitive with conventional classifiers on card data (Reyes & Cortez, 2025; Yu et al., 2024) and that recurrent layers capture sequential behaviour (Bonde & Bichanga, 2025). The additional gain from the full pipeline is consistent with findings that feature engineering and selection benefit a range of models (Ouazzane et al., 2025; Siam et al., 2025; Yin et al., 2023). Two cautions apply. Accuracy hides the behaviour that matters under imbalance (Lucas & Jurgovsky, 2020), which is why precision, recall, $F1$ -score and AUPRC are emphasised, and fraud patterns drift over time (Dal Pozzolo et al., 2018), so the threshold and the selected features should be revisited whenever the model is retrained. Finally, graph-based models (Cheng et al., 2025) capture relations among entities that an attribute-sequence model cannot, and combining the two is a natural extension.

7. CONCLUSION

This paper presented a novel Cognitive-Driven Machine Learning Framework for online transaction fraud detection. By addressing the core challenges of high-dimensional feature noise, computational complexity in transformer models, and severe class imbalance, our framework establishes a robust standard for financial security systems. The incorporation of single-mode Improved Information Gain (IIG) successfully isolates high-entropy behavioral markers, while the hybrid Efficient Transformer-GRU architecture harmonizes global contextual awareness with precise temporal tracking. Empirical evaluations demonstrate outstanding performance, achieving an AUPRC of 0.948 and an $F1$ -score of 0.924 while maintaining low inference latency. Future research will explore federated learning extensions to safeguard user privacy across decentralized banking networks.

REFERENCES

1. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv. <https://arxiv.org/abs/1607.06450>
2. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255. <https://doi.org/10.1214/ss/1042727940>
3. Bonde, L., & Bichanga, A. K. (2025). Improving credit card fraud detection with ensemble deep learning-based models: A hybrid approach using SMOTE-ENN. *Journal of Computing Theories and Applications*, 2(3). <https://doi.org/10.62411/jcta.12021>
4. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
5. Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2025). Graph neural networks for financial fraud detection: A review. *Frontiers of Computer Science*. <https://arxiv.org/abs/2411.05815>
6. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724–1734). <https://doi.org/10.3115/v1/D14-1179>
7. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv. <https://arxiv.org/abs/1412.3555>

8. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797. <https://doi.org/10.1109/TNNLS.2017.2736643>
9. Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). IEEE. <https://doi.org/10.1109/SSCI.2015.33>
10. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). <https://doi.org/10.1145/1143844.1143874>
11. Grover, P., Xu, J., Tittelfitz, J., Cheng, A., Li, Z., Zablocki, J., Liu, J., & Zhou, H. (2022). Fraud dataset benchmark and applications. *arXiv*. <https://arxiv.org/abs/2208.14417>
12. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
13. Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning* (PMLR 119, pp. 5156–5165).
14. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2980–2988). <https://doi.org/10.1109/ICCV.2017.324>
15. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*. <https://arxiv.org/abs/1711.05101>
16. Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. *arXiv*. <https://arxiv.org/abs/2010.06479>
17. Ouazzane, K., Ikeda, C., Djemai, R., Phipps, A., & Maleh, Y. (2025). A novel feature engineering framework for cyber fraud detection using machine learning and deep learning algorithms. *Information Security Journal: A Global Perspective*. <https://repository.londonmet.ac.uk/10579/>
18. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. <https://doi.org/10.1007/BF00116251>
19. Reyes, K., & Cortez, V. (2025). Credit card fraud detection using RoFormer model with relative distance rotating encoding. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI)*. <https://doi.org/10.1109/CAI64502.2025.00199>
20. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
21. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
22. Siam, A. M., Bhowmik, P., & Uddin, M. P. (2025). Hybrid feature selection framework for enhanced credit card fraud detection using machine learning models. *PLOS ONE*, 20, e0326975. <https://doi.org/10.1371/journal.pone.0326975>
23. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 127063. <https://doi.org/10.1016/j.neucom.2023.127063>
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
25. Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., & Kwak, J. (2023). IGRF-RFE: A hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. *Journal of Big Data*, 10, Article 15. <https://doi.org/10.1186/s40537-023-00694-8>
26. Yu, C., Xu, Y., Cao, J., Zhang, Y., Jin, Y., & Zhu, M. (2024). Credit card fraud detection using advanced transformer model. *arXiv*. <https://arxiv.org/abs/2406.03733>