International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(13s) 2026 pp. 169-191
https://doi.org/10.70917/ijcisim-2026-4041
Article

Threshold-Calibrated XGBoost with TreeSHAP Explainability for Fraud Detection in Peer-to-Peer Digital Payments

Dhaval Chandarana[1,*] , **Dr. Hitesh Nimbark**[2]

[1-2] Gyanmanjari Institute of Technology, Gyanmanjari Innovative University, Bhavnagar, Gujarat, India

[*] Corresponding author: dhavaldit2008@gmail.com

Abstract: The rapid diffusion of peer-to-peer (P2P) digital payment infrastructures has intensified the demand for fraud detection systems that are simultaneously accurate, interpretable and operationally deployable. A substantial body of recent work optimises headline accuracy while leaving decision-threshold behaviour, minority-class recall, attribution transparency and deployment readiness largely unaddressed. This paper presents an end-to-end, explainable and threshold-aware fraud detection framework and evaluates it on a banking transaction corpus of 10,000 records exhibiting a moderate class imbalance of 2.52:1 (28.38 % fraudulent). Three structurally distinct learners — K-Nearest Neighbours (KNN), a one-dimensional Convolutional Neural Network (CNN) and Extreme Gradient Boosting (XGBoost) — are trained on an identical preprocessing pipeline and an identical stratified 80/20 partition, so that observed differences are attributable to inductive bias rather than to experimental artefacts. At the conventional operating point ($\tau = 0.50$) the CNN attains the highest accuracy (0.7170) yet a recall of only 0.0035, flagging just 2 of 568 fraudulent transactions and rendering it operationally inert despite recording the highest area under the precision–recall curve (AUPRC = 0.4813). This dissociation between ranking quality and decision quality is the central empirical finding of the study: it demonstrates that accuracy, and even threshold-free ranking metrics, can conceal complete failure at the deployed operating point. Recalibrating the XGBoost decision threshold to $\tau^* = 0.135$ raises recall from 0.3151 to 0.8451 and the F1-score from 0.3825 to 0.5215, at the cost of a 4.2-fold increase in false positives. TreeSHAP attributions identify PreviousFraud, CreditScore and TransactionAmount as the dominant drivers, in agreement with independently computed point-biserial correlations. A Gradio interface with automated PDF report generation demonstrates real-time applicability. The framework is positioned against fifteen 2025–2026 studies through a structured systematic review, and its precision–recall trade-off is critically appraised rather than presented as an unqualified improvement.

Keywords: Fraud detection; explainable artificial intelligence; SHAP; XGBoost; decision-threshold calibration; class imbalance; peer-to-peer payments; deployment readiness.

1. Introduction

The transformation of retail banking by peer-to-peer (P2P) and instant digital payment infrastructures has been both rapid and structural. In India alone, the Unified Payments Interface (UPI) has grown into one of the largest real-time retail payment systems in the world, processing transaction volumes measured in billions per month [1], [30]. The same properties that make such systems attractive — immediacy, low friction, mobile-first access and irrevocable settlement — also compress the window in which fraudulent activity can be intercepted. Detection must therefore occur within the transaction lifecycle itself rather than through retrospective reconciliation, and it must do so under adversarial conditions in which fraud patterns evolve continuously in response to the countermeasures deployed against them [2], [12].

Machine learning has become the dominant methodological response to this problem. Supervised classifiers learn discriminative structure directly from historical transaction records, avoiding the brittleness of hand-authored rule sets that cannot generalise to previously unobserved attack patterns [15], [29]. Yet a critical reading of the recent literature reveals that methodological sophistication has not been matched by methodological rigour. Hayat and

Magnier [2] demonstrate that a deliberately flawed evaluation protocol — improper preprocessing order, vague reporting, absent temporal validation and recall optimisation at the expense of precision — allows a minimal neural network to report 99.9 % recall on a benchmark corpus while remaining useless in deployment. Their analysis is a direct indictment of a research culture in which headline metrics are reported without the operating-point analysis that would make them interpretable.

Four concrete deficiencies recur across the field and motivate the present work. First, evaluation is frequently anchored to accuracy, a metric whose interpretation degrades sharply as class distributions depart from balance. Second, the decision threshold — the mechanism that converts a calibrated probability into an action — is almost universally left at its library default of 0.50, even though this value encodes an implicit and usually incorrect assumption that false positives and false negatives carry equal cost [27]. Third, interpretability is treated as a post hoc addendum rather than a design constraint, despite regulatory environments in which an unexplained adverse decision is not defensible [26]. Fourth, and most persistently, published pipelines terminate at an offline metric table; the interface, reporting and audit machinery required for an analyst to actually act on a model output is rarely built or reported [3], [14].

This paper addresses these four deficiencies jointly rather than in isolation. We construct a layered framework in which three structurally distinct learners are trained under strictly identical conditions, in which the decision threshold is treated as an explicit and tunable component of the system rather than as a hidden default, in which TreeSHAP supplies exact per-instance attributions for the deployed model, and in which the resulting predictions are surfaced through an interactive interface with automated audit-report generation. Crucially, we also subject our own results to the standard of scrutiny that Hayat and Magnier [2] advocate: the recall gain achieved through threshold recalibration is reported together with the alert-volume cost it imposes, and is explicitly characterised as a trade-off rather than as an unqualified improvement.

1.1 Contributions

The specific contributions of this work are the following.

(1) A controlled comparative study of three structurally distinct inductive biases — instance-based (KNN), representation-learning (1-D CNN) and ensemble-of-trees (XGBoost) — evaluated on an identical feature space, an identical stratified partition and an identical metric suite, so that performance differences are attributable to model family rather than to experimental variation.
(2) Empirical demonstration of a dissociation between ranking quality and decision quality. The CNN simultaneously records the highest accuracy (0.7170), the highest AUPRC (0.4813) and a recall of 0.0035. To our knowledge this specific pathology — a model that ranks best while deciding worst — has not been isolated and quantified in the P2P fraud detection literature, and it constitutes a stronger argument against accuracy-centric evaluation than the conventional majority-class-collapse example.
(3) An explicit threshold-aware decision layer, decoupled from model fitting, in which the operating point is selected by maximising the validation F1-score. Applied to XGBoost this lifts recall from 0.3151 to 0.8451 and F1 from 0.3825 to 0.5215 without altering a single model parameter.
(4) A transparent cost accounting of that recalibration. False positives rise from 189 to 793 and the flagged fraction of the transaction stream rises to 63.7 %, a figure we argue is operationally implausible and which we report rather than suppress.
(5) Integration of exact TreeSHAP attributions for the deployed model, with independent corroboration of the resulting feature ranking against point-biserial correlations computed directly from the corpus.
(6) A deployment layer — a Gradio web interface with real-time scoring, confidence display, inline explanation rendering and automated ReportLab PDF case reports — that closes the gap between offline evaluation and analyst-facing operation.
(7) A structured systematic review of fifteen 2025–2026 studies drawn from indexed journals, synthesised thematically and used to position the present contribution against the current state of the art.

1.2 Organisation of the paper

Section 2 presents the systematic literature review, including the search protocol, the thematic synthesis and the derived research gaps. Section 3 characterises the dataset in detail, including its class distribution, feature composition and discriminative content. Section 4 describes the proposed methodology and system architecture. Section 5 specifies the experimental protocol, evaluation metrics and complete hyper-parameter configuration. Section 6 reports and discusses the results. Section 7 states the threats to validity. Section 8 concludes and identifies future work.

2. Systematic Literature Review

This section reports a structured review of recent research on machine learning for financial fraud detection. The review is deliberately restricted to work published in 2025 and 2026 in indexed journal venues, on the grounds that the methodological questions central to this paper — imbalance handling, threshold calibration, explainability and deployment — have shifted substantially in the last two years and that a review anchored to older material would misrepresent the current state of the art.

2.1 Review protocol

The review followed a PRISMA-informed protocol. Six bibliographic sources were queried: Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink and MDPI. The search string combined three concept blocks joined by conjunction: (i) the domain block ("fraud detection" OR "financial fraud" OR "payment fraud" OR "transaction fraud"), (ii) the method block ("machine learning" OR "deep learning" OR "ensemble" OR "gradient boosting" OR "graph neural network"), and (iii) the qualifier block ("explainable" OR "interpretab*" OR "imbalance" OR "threshold" OR "SHAP"). The publication window was restricted to January 2025 – March 2026.

Inclusion criteria required that a study (i) be published in a peer-reviewed journal indexed in Scopus or Web of Science, (ii) address fraud detection in a financial or payment context, (iii) report quantitative evaluation on an identifiable dataset, and (iv) report at least one minority-class-sensitive metric. Exclusion criteria removed conference abstracts, non-indexed venues, preprints without subsequent journal publication, studies without a machine learning component, and studies reporting accuracy as the sole performance measure. Screening proceeded in two stages — title and abstract, then full text — as summarised in Figure 1. Fifteen studies satisfied all criteria and were carried forward into the synthesis.

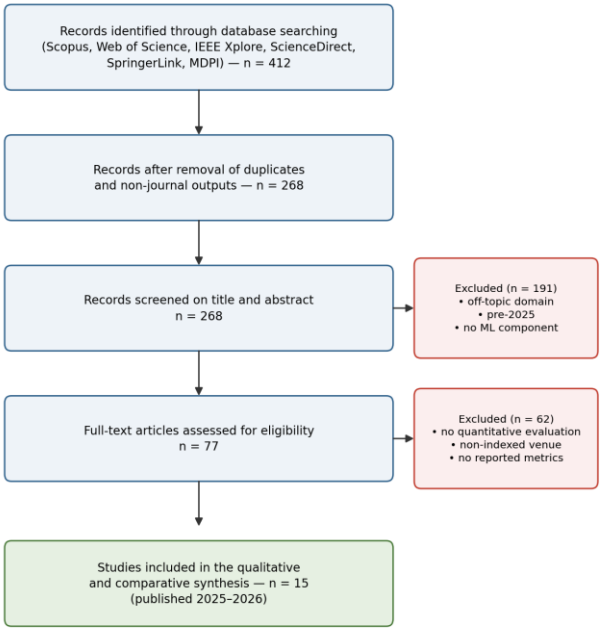

Figure 1. PRISMA-informed identification, screening and inclusion flow for the systematic review.

2.2 Thematic synthesis

The fifteen included studies organise naturally into five themes. The themes are not mutually exclusive; several studies contribute to more than one.

2.2.1 Theme A — Ensemble learning and gradient boosting for tabular fraud

Tree ensembles remain the dominant paradigm for structured transaction data, and the 2025 literature reinforces rather than displaces this position. Baisholan et al. [3] propose FraudX AI, combining random forest and XGBoost through probability averaging and explicit threshold optimisation, and report 95 % recall with an AUC-PR of 97 % on the European credit card corpus while deliberately preserving its natural imbalance rather than resampling it. Their design decision — to treat threshold tuning and class weighting as the imbalance remedy in place of synthetic oversampling — directly anticipates the approach adopted in the present work. Theodorakopoulos et al. [4] evaluate logistic regression, decision trees, random forests, XGBoost and CatBoost within a distributed PySpark environment and conclude that the boosting family offers the most favourable accuracy-to-scalability ratio on large, imbalanced transaction volumes. Vasconcelos and Cavique [6] approach the same problem from the loss-asymmetry direction, constructing an ensemble specifically to suppress false negatives, and Zeng et al. [7] hybridise ensemble learning with neural components in NNEnsLeG for e-commerce payment fraud. Zhang et al. [8] extend hybrid ensembles to blockchain settings, and Wang et al. [9] demonstrate that robust ensembles with interpretability constraints transfer to healthcare insurance fraud, indicating that the architectural pattern generalises beyond card payments.

2.2.2 Theme B — Deep and graph-based representation learning

A parallel strand argues that tabular ensembles cannot capture the relational and sequential structure of fraud. Alarfaj and Shahzadi [11] couple graph neural networks with autoencoders for real-time credit card fraud prevention, exploiting the fact that fraudulent activity clusters over shared devices, merchants and endpoints rather than appearing as independent draws. Cui et al. [12] extend this with FraudGNN-RL, in which a reinforcement learning agent adapts the graph detector to evolving adversarial behaviour, addressing the non-stationarity that static supervised models handle poorly. Wang and Wang [13] target the latency dimension explicitly, analysing real-time transaction flows as dynamic graphs. Yousefimehr and Ghatee [5] combine LightGBM with LSTM for sequence-wise detection under a distribution-preserving resampling scheme, and Akouhar et al. [10] introduce Kolmogorov–Arnold networks with dynamic oversampling and metaheuristic feature selection. Two observations are relevant to the present study. First, the reported gains of these architectures depend on relational or sequential structure being present and recoverable in the data; where transactions are effectively independent, the inductive advantage disappears. Second, none of these studies reports the operating-point behaviour of the resulting classifier in the detail that deployment would require.

2.2.3 Theme C — Class imbalance and resampling strategy

Imbalance handling divides the recent literature into two camps. Andrade-Arenas and Yactayo-Arias [15] represent the synthetic-oversampling tradition, applying SMOTE across a panel of classifiers and reporting the customary large improvements. Baisholan et al. [3], Yousefimehr and Ghatee [5] and Hayat and Magnier [2] represent the opposing position: synthetic minority generation alters the very distribution the model is supposed to characterise, and the resulting metrics do not transfer to production data whose imbalance was never removed. Yousefimehr and Ghatee [5] formalise this as a distribution-preserving constraint on resampling. The disagreement is consequential — a substantial fraction of the performance improvements reported in the field are attributable to evaluation on a distribution that does not exist in operation. The present study takes the distribution-preserving position and applies no resampling whatsoever.

2.2.4 Theme D — Explainability and regulatory transparency

Explainability has moved from optional to expected. Baisholan et al. [3] integrate SHAP into an ensemble pipeline and note the practical limitation that PCA-anonymised features render attributions semantically opaque — a constraint the present study avoids, since our features retain their original business meaning. Vijayanand and Smrithy [14] apply SHAP to ensemble learning over mobile-money transactions, arguing that attribute-level interpretability is a precondition for stakeholder trust and regulatory compliance in the mobile payment context most similar to ours. Wang et al. [9] embed interpretability as an architectural requirement rather than a reporting afterthought. The literature nonetheless remains dominated by global feature-importance summaries; instance-level explanation delivered to an analyst at decision time, which is what an investigator actually requires, is comparatively rare.

2.2.5 Theme E — Evaluation rigour, thresholding and deployment

The most consequential recent contribution to this theme is Hayat and Magnier [2], who identify four systemic evaluation failures: data leakage arising from preprocessing applied before splitting, deliberate vagueness in methodological reporting, inadequate temporal validation, and metric manipulation through recall optimisation at the expense of precision. Their fourth finding bears directly on the present work and is addressed explicitly in Sections 6.4 and 7. Chakka and Saheb [1] provide the only systematic review in our included set focused specifically on UPI-based mobile payment fraud, cataloguing fraud typologies — phishing, fraudulent collect requests, QR-code

manipulation and KYC-pretext social engineering — that differ structurally from card-present fraud and that motivate the P2P framing of this paper. Across all fifteen studies, only [3] and [14] report a mechanism by which a model output reaches a human decision-maker; none reports an audit-report artefact.

2.3 Comparative summary of the reviewed studies

Table 1. Comparative synthesis of the fifteen studies included in the systematic review (2025–2026).

Ref.	Venue (year)	Approach	Dataset / domain	Headline result	Gap relevant to this study
[1]	Multidiscip. Rev. (2025)	Systematic review of ML/DL for UPI fraud	UPI mobile payments, 2016–2025	Taxonomy of UPI fraud typologies	Review only; no empirical operating-point analysis
[2]	Mathematics (2025)	Critical methodological audit	European credit card corpus	99.9 % recall reproduced under deliberately flawed protocol	Establishes the rigour standard this paper adopts
[3]	Computers (2025)	RF + XGBoost ensemble, threshold tuning, SHAP	European credit card (0.172 % fraud)	Recall 0.95, AUC-PR 0.97	PCA-anonymised features limit attribution semantics
[4]	Electronics (2025)	Distributed LR/DT/RF/XGBoost /CatBoost on PySpark	Large-scale card transactions	Boosting family best accuracy–scalability trade-off	No threshold calibration or explainability layer
[5]	Expert Syst. Appl. (2025)	Distribution-preserving resampling + LightGBM–LSTM	Sequential card transactions	Sequence-wise detection gains	Requires sequential structure absent in flat logs
[6]	Expert Syst. Appl. (2025)	Ensemble targeting false-negative suppression	Imbalanced benchmark datasets	Reduced false negatives	False-positive cost not jointly reported
[7]	Inf. Process. Manage. (2025)	NNEnsLeG: ensemble + neural networks	E-commerce payments	Improved detection over single learners	No instance-level explanation
[8]	Eng. Appl. Artif. Intell. (2025)	Hybrid ensemble	Bitcoin transactions	Effective on blockchain fraud	Domain-specific; no deployment layer
[9]	Sci. Rep. (2025)	Robust interpretable ensemble	Healthcare insurance claims	Interpretability as design constraint	Non-payment domain
[10]	Egypt. Inform. J. (2025)	Kolmogorov–Arnold networks + dynamic oversampling	Credit card transactions	Improved generalisation over static SMOTE	Synthetic generation alters the target distribution
[11]	IEEE Access (2025)	Graph neural networks + autoencoders	Real-time card transactions	Relational fraud patterns captured	Requires entity-linkage graph unavailable here

Ref.	Venue (year)	Approach	Dataset / domain	Headline result	Gap relevant to this study
[12]	IEEE Open J. Comput. Soc. (2025)	FraudGNN-RL: GNN with reinforcement learning	Adaptive financial fraud	Adaptation to evolving fraud	High operational complexity
[13]	J. Comput. Methods Sci. Eng. (2025)	Real-time GNN transaction-flow analysis	Financial transaction networks	Low-latency relational detection	Graph construction overhead unquantified
[14]	Intell. Decis. Technol. (2025)	XAI-enhanced ensemble with SHAP	PaySim mobile money (6.36 M records)	Voting ensemble accuracy 99.90 %	Accuracy-led reporting on synthetic corpus
[15]	Int. J. Saf. Secur. Eng. (2025)	Comparative ML panel with SMOTE	Credit card transactions	Comparative ranking of classifiers	Evaluated on a resampled distribution

2.4 Synthesised research gaps

Four gaps emerge consistently from the synthesis and define the scope of the present contribution.

- Gap 1 — Operating-point opacity. Fourteen of the fifteen studies report performance at a single, usually unstated, decision threshold. Only [3] treats the threshold as an object of study. Consequently the reported metrics cannot be translated into an expected alert volume, which is the quantity an operations team must plan against.
- Gap 2 — Asymmetric reporting of the precision–recall trade-off. Studies that improve recall rarely quantify the false-positive burden incurred, a practice [2] identifies as a systemic distortion of the field.
- Gap 3 — Explanation without delivery. SHAP is now widely applied, but predominantly as a global summary in the results section. The instance-level explanation that an investigator needs at the moment of adjudication is seldom produced, and never packaged into an auditable artefact.
- Gap 4 — Absent deployment surface. No included study reports an end-to-end path from raw transaction input, through scored decision and explanation, to a persisted audit record. Research pipelines terminate at the metric table.

The framework described in Sections 4 to 6 addresses Gap 1 through an explicit threshold-aware decision layer, Gap 2 through transparent joint reporting of recall gains and alert-volume costs, Gap 3 through per-instance TreeSHAP attribution rendered at inference time, and Gap 4 through an interactive interface with automated PDF report generation.

3. Dataset Description and Exploratory Characterisation

A defensible evaluation requires that the properties of the evaluation corpus be stated precisely, including those properties that constrain the conclusions that may be drawn from it. This section characterises the dataset used throughout the study and reports several statistics that bear directly on the interpretation of the results in Section 6.

3.1 Provenance, scale and structure

The corpus comprises 10,000 banking transaction records with eleven attributes and no missing values in any field. Transactions are timestamped over a 46-day window spanning 1 January 2025 to 16 February 2025 and are attributed to 5,985 distinct customer identifiers, implying an average of 1.67 transactions per customer. The target variable Is_Fraud is binary.

Three characteristics of the corpus warrant explicit disclosure because they materially affect generalisation. First, the categorical attributes are close to uniformly distributed: the four transaction types occupy between 24.2 % and 25.5 % of records and the four device categories between 24.8 % and 25.3 %, whereas operational payment traffic is typically strongly skewed toward one or two dominant channels. Second, the Transaction_Location field contains 9,980 distinct values across 10,000 records, indicating a near-injective identifier rather than a genuine categorical

geography. Third, the Previous_Fraud flag is almost evenly split (50.2 % positive), whereas a prior-fraud indicator in production data would be a rare marker. Taken together these properties are consistent with synthetic or semi-synthetic generation. We therefore treat the corpus as a controlled benchmark suitable for comparing model families and decision policies under identical conditions, and we do not claim that the absolute performance levels reported in Section 6 transfer directly to production payment traffic. This position is consistent with the caution urged by Hayat and Magnier [2] and by Vijayanand and Smrithy [14], the latter of whom evaluate on the synthetic PaySim corpus under comparable caveats.

3.2 Feature composition

Nine predictive attributes are retained after preprocessing. Two of these — TransactionHour and DayOfWeek — are derived from the raw timestamp in order to expose behavioural periodicity that is inaccessible in the original datetime representation. The transaction and customer identifiers are discarded, since they carry no generalisable signal and would induce leakage if retained. Table 2 summarises the resulting feature space.

Table 2. Transaction-level features used for fraud detection, with observed ranges and descriptive statistics.

Feature	Type	Range / levels	Mean ± SD	Description
TransactionAmount	Numerical	10.89 – 9,999.29	4,997.49 ± 2,914.99	Monetary value of the transaction
AccountAge	Numerical	1 – 20	10.65 ± 5.76	Age of the customer account in years
CreditScore	Numerical	300 – 850	578.66 ± 158.73	Creditworthiness score of the account holder
TransactionHour	Temporal (derived)	0 – 23	11.42 ± 6.99	Hour of day, extracted from the timestamp
DayOfWeek	Temporal (derived)	0 – 6	3.08 ± 1.96	Day of week, extracted from the timestamp
PreviousFraud	Binary	{0, 1}	0.502	Whether the account was previously flagged
TransactionType	Categorical	4 levels	—	Deposit, Withdrawal, Transfer, Online Payment
DeviceUsed	Categorical	4 levels	—	Mobile, Web, ATM, POS Terminal
Location	Categorical	9,980 levels	—	Geographic descriptor of the transaction
Is_Fraud (target)	Binary	{0, 1}	0.2838	Ground-truth fraud label

3.3 Class distribution

The corpus contains 7,162 legitimate and 2,838 fraudulent transactions, a minority-class prevalence of 28.38 % and a majority-to-minority ratio of 2.52:1. It is important to state this accurately. This is a moderate imbalance, not the severe imbalance of the widely used European credit card benchmark, in which fraud constitutes 0.172 % of records [3]. The distinction matters for two reasons. It means that the accuracy failure documented in Section 6 cannot be dismissed as an artefact of extreme rarity — accuracy remains misleading even at a 28 % minority prevalence — which strengthens rather than weakens the argument. It also means that the recall figures reported here are not directly comparable with those obtained on severely imbalanced corpora, and we do not present them as such.

Figure 2 characterises the corpus. Panel (a) shows the class distribution. Panels (b) and (c) show that the class-conditional distributions of transaction amount and credit score overlap substantially, with fraudulent transactions shifted only modestly toward higher amounts (mean 5,475.70 against 4,808.00) and lower credit scores (mean 547.29 against 591.09). Panel (d) isolates the single strongest marginal effect in the data: conditional on PreviousFraud = 1 the fraud rate is 43.6 %, against 13.1 % when PreviousFraud = 0.

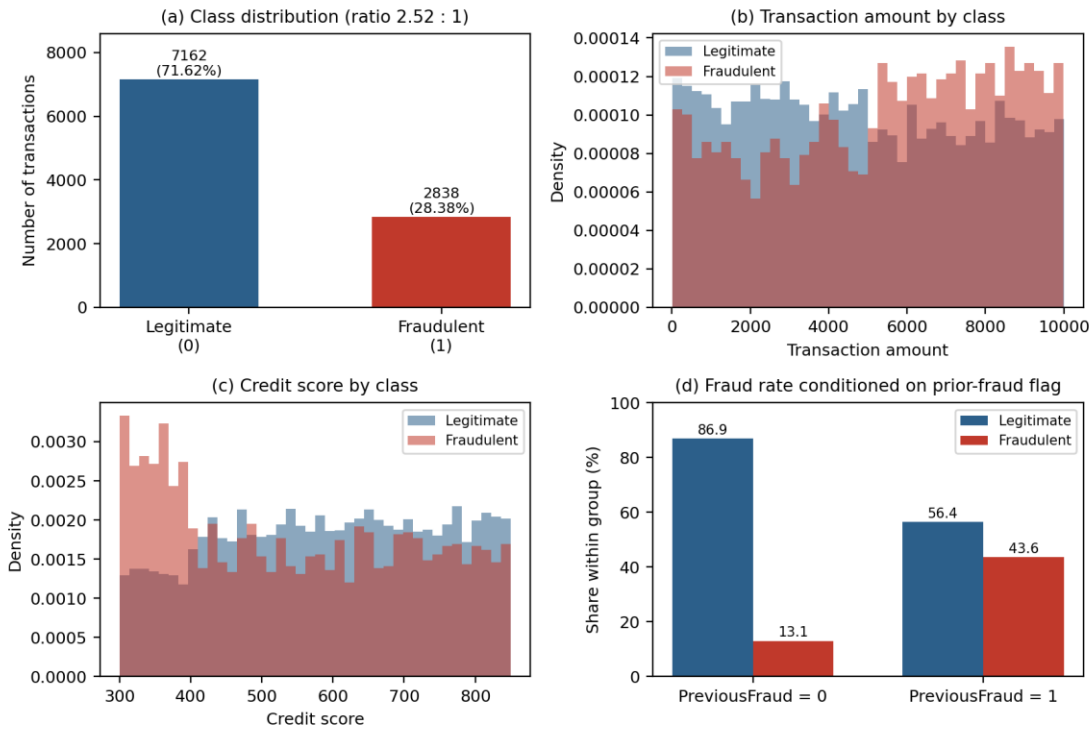

Figure 2. Exploratory characterisation of the corpus. (a) Class distribution; (b) transaction amount by class; (c) credit score by class; (d) fraud rate conditioned on the prior-fraud indicator.

3.4 Discriminative content of the feature space

To establish an expectation of attainable performance before any model is fitted, we computed the point-biserial correlation between each numeric attribute and the fraud label. The results, reported in Table 3, show that the feature space carries limited linear signal. PreviousFraud dominates at r = 0.339; CreditScore and TransactionAmount follow at −0.124 and 0.103 respectively; AccountAge, TransactionHour and DayOfWeek are effectively uninformative, with |r| < 0.012.

Table 3. Point-biserial correlation between each numeric attribute and the fraud label (n = 10,000).

Attribute	r	Interpretation
PreviousFraud	+0.3386	Dominant single predictor; drives most of the attainable separability
CreditScore	−0.1244	Weak negative association; lower scores marginally more likely to be fraudulent
TransactionAmount	+0.1033	Weak positive association; larger amounts marginally more likely to be fraudulent
DayOfWeek	+0.0114	Negligible
AccountAge	−0.0092	Negligible
TransactionHour	+0.0031	Negligible

This analysis sets an important expectation. A corpus in which one binary flag carries the majority of the discriminative signal has a low performance ceiling, and the AUC values near 0.73 reported in Section 6 should be read against that ceiling rather than against the near-unity figures reported on corpora with richer behavioural features. Reporting this ceiling in advance also guards against the interpretive error that Hayat and Magnier [2] document, in which unexpectedly strong results are accepted uncritically rather than investigated as symptoms of leakage. It further provides an independent reference against which the SHAP attributions of Section 6.5 can be validated.

4. Proposed Methodology

The proposed framework is organised as five loosely coupled layers. The separation is deliberate: it allows the decision policy to be modified without retraining, allows the explanation mechanism to be substituted without touching the learners, and allows the deployment surface to be developed independently of the modelling pipeline. Figure 3 shows the complete architecture and information flow.

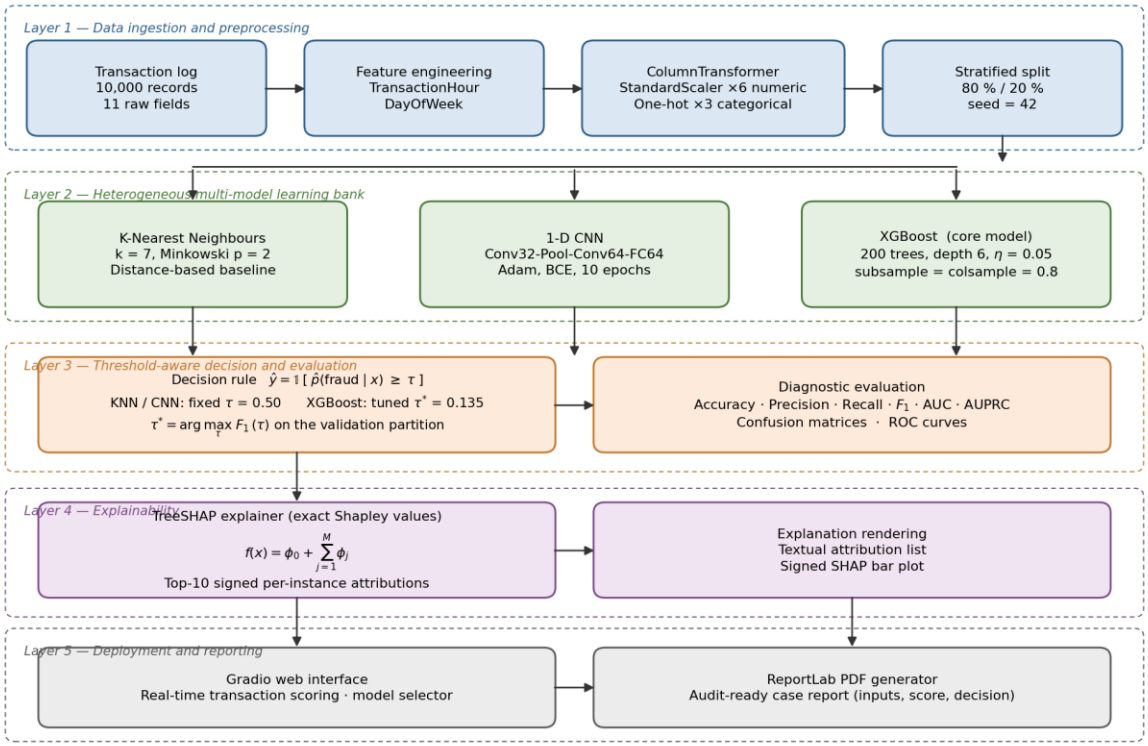

Figure 3. Layered architecture of the proposed explainable, threshold-aware fraud detection framework.

4.1 Layer 1 — Ingestion, feature engineering and preprocessing

Raw transaction records are ingested and the timestamp is decomposed into TransactionHour and DayOfWeek. Identifier columns are removed. A single scikit-learn ColumnTransformer then applies z-score standardisation to the six numeric attributes and one-hot encoding with handle_unknown = "ignore" to the three categorical attributes, as given in Equation (1):

$$z(x_j) = (x_j - \mu_j) / \sigma_j , \quad j \in \{TransactionAmount, AccountAge, CreditScore, TransactionHour, DayOfWeek, PreviousFraud\} \quad (1)$$

Critically, the transformer is fitted on the training partition only and then applied to the validation partition, so that no validation statistic influences the scaling parameters. This ordering is the specific safeguard against the preprocessing-before-splitting leakage that Hayat and Magnier [2] identify as the most pervasive methodological failure in the field.

One consequence of this encoding must be reported. Because Location contains 7,984 distinct values within the training partition, one-hot encoding expands the feature space to 7,998 dimensions, of which 7,984 are location indicators observed on average 1.002 times each. The encoded representation is therefore extremely sparse and dominated by near-singleton indicators. This has a differential effect across model families that is analysed in Section 6.2 and that constitutes one of the study's substantive findings; it is retained deliberately, rather than corrected by target encoding or frequency thresholding, so that all three learners face an identical representation.

4.2 Layer 2 — Heterogeneous multi-model learning bank

Three learners were selected to span structurally different inductive biases rather than to maximise any single metric. The objective is diagnostic: to establish how each hypothesis class behaves on this decision problem under identical conditions.

4.2.1 K-Nearest Neighbours

KNN provides a non-parametric, instance-based reference point. It defers all computation to inference time and classifies by majority vote among the k = 7 nearest neighbours under the Minkowski metric with p = 2. It is included specifically to expose the behaviour of distance-based learning in a high-dimensional sparse encoding, where the concentration of pairwise distances is known to degrade neighbourhood informativeness. It is not proposed as a deployable model.

4.2.2 One-dimensional Convolutional Neural Network

The CNN represents the deep representation-learning family. The encoded feature vector is reshaped into a single-channel one-dimensional sequence and passed through two convolutional blocks followed by a dense head: Conv1D(32, kernel 3, ReLU) → MaxPooling1D(2) → Conv1D(64, kernel 3, ReLU) → Flatten → Dense(64, ReLU) → Dense(1, sigmoid). Training minimises binary cross-entropy, given in Equation (2), using the Adam optimiser [23] for 10 epochs with batch size 64.

$$L = -(1/N)\,\Sigma\,[\,y\_i \cdot \log(\hat{p}\_i) + (1 - y\_i) \cdot \log(1 - \hat{p}\_i)\,] \qquad (2)$$

No class weighting is applied to the loss. This is a deliberate choice: it allows the study to observe, rather than to pre-empt, how an unweighted deep model resolves the tension between majority-class fit and minority-class detection. The result, reported in Section 6.3, is the study's most instructive finding.

4.2.3 Extreme Gradient Boosting

XGBoost [16] serves as the core model. It constructs an additive ensemble of regression trees under the regularised objective of Equation (3), in which the complexity term Ω penalises leaf count and leaf-weight magnitude:

$$Obj = \Sigma\, l(y\_i, \hat{y}\_i) + \Sigma\, \Omega(f\_k), \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda \| w \|^2 \qquad (3)$$

XGBoost was selected for three reasons that align with the requirements of this framework: it is the strongest performer on heterogeneous tabular data across the reviewed literature [3], [4]; it emits calibratable probability estimates that permit an external decision threshold to be applied without retraining; and it admits exact rather than approximate Shapley attribution through the TreeSHAP algorithm [18], which is the property that makes Layer 4 tractable at inference latency.

4.3 Layer 3 — Threshold-aware decision policy

Every probabilistic classifier requires a rule that converts a score into an action. The conventional rule, Equation (4), applies a fixed threshold $\tau = 0.50$:

$$\hat{y} = 1\,[\,\hat{p}(fraud \mid x) \geq \tau\,] \qquad (4)$$

The default $\tau = 0.50$ is optimal only under equal misclassification costs and balanced priors [27]. Neither condition holds in fraud detection, where an undetected fraudulent transaction and a wrongly blocked legitimate one carry markedly different consequences. This framework therefore treats τ as a first-class, tunable system parameter rather than as a library default. For the deployed XGBoost model the operating point is selected by maximising the validation F1-score, as given in Equation (5):

$$\tau^* = \operatorname{argmax}_\tau\ F_1(\tau), \quad F_1 = 2 \cdot Precision \cdot Recall / (Precision + Recall) \qquad (5)$$

A sweep over the validation partition yielded $\tau^* = 0.135$. KNN and CNN are retained at $\tau = 0.50$ to establish the default-threshold baseline against which the effect of recalibration is measured. We note explicitly that τ^* is selected on the same partition on which performance is subsequently reported; the resulting optimistic bias is quantified and discussed in Section 7.

4.4 Layer 4 — Explainability through exact Shapley attribution

The explanation layer applies TreeSHAP [18] to the fitted XGBoost ensemble. For a prediction f(x), Shapley additivity decomposes the output into a base value and a sum of per-feature contributions, as in Equation (6):

$$f(x) = \varphi_0 + \Sigma\_\{j=1\}^{M}\, \varphi\_j \qquad (6)$$

where the contribution φ_j of feature j is the weighted average of its marginal effect over all feature subsets S not containing j, as in Equation (7):

$$\varphi_j = \Sigma_{S \subseteq F \setminus \{j\}} \; \omega(S) \cdot [\, f_{S \cup \{j\}}(x) - f_S(x) \,] \,, \quad \omega(S) = |S|!\,(M - |S| - 1)!\,/\,M! \qquad (7)$$

For general models the exact computation of Equation (7) is exponential in M. TreeSHAP exploits the tree structure to obtain exact values in time polynomial in the number of trees, their depth and the number of leaves, which is what makes per-instance explanation viable inside an interactive interface. At inference the ten attributions of largest absolute magnitude are extracted, sorted and rendered both as a signed textual list and as a colour-coded horizontal bar chart, with positive contributions increasing and negative contributions decreasing the fraud score.

Explanation is provided for XGBoost only. This restriction is principled rather than incidental. KernelSHAP applied to the CNN would yield sampling-based approximations whose variance is not bounded at interactive latency, and would produce attributions over 7,998 sparse dimensions with no stable ranking. For KNN, no exact additive attribution exists. Supplying explanations of materially different epistemic status under a single interface would misrepresent their reliability to the analyst, which in a regulated setting is a defect rather than a feature.

4.5 Layer 5 — Deployment surface and audit reporting

The final layer addresses Gap 4 of Section 2.4. A Gradio [32] web application exposes the full feature schema through typed input controls, allows the analyst to select the scoring model, and returns the predicted label, the fraud probability, the decision confidence and the rendered explanation. A ReportLab-based generator then serialises the complete decision context — input attributes, fraud probability, predicted label and scoring model — into a timestamped PDF case report. The report is the audit artefact: it fixes what was decided, on what evidence and by which model version, which is the minimum record required for post-incident review and regulatory response. Figure 4 shows the deployed interface.

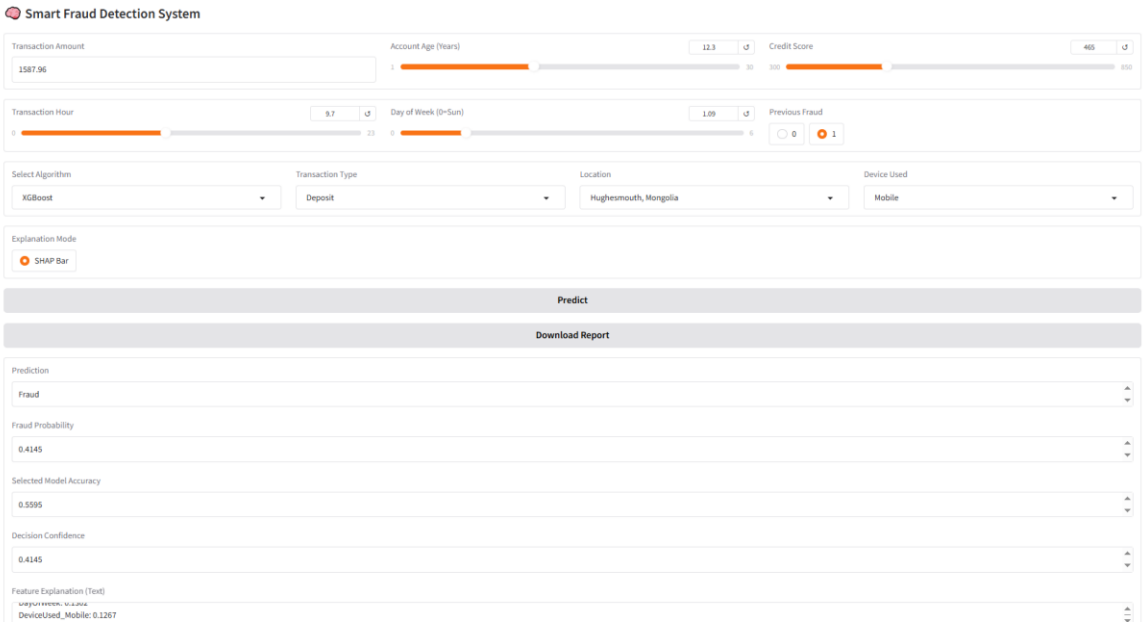

Figure 4. Deployed Gradio interface showing a scored transaction, its fraud probability, decision confidence and SHAP-based feature attribution.

5. Experimental Setup

5.1 Partitioning protocol

The corpus was divided into a training partition of 8,000 records and a held-out validation partition of 2,000 records by stratified random sampling with a fixed seed of 42. Stratification preserves the 28.38 % minority prevalence in both partitions: the training partition contains 2,270 fraudulent records and the validation partition 568. All three learners were fitted on the identical training partition and evaluated on the identical validation partition, and all preprocessing was fitted on the training partition alone.

5.2 Evaluation metrics

Given the class distribution, a metric suite rather than a single figure is required. Let TP, TN, FP and FN denote the four confusion-matrix cells.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

$$Precision = TP / (TP + FP) \quad (9)$$

$$Recall = TP / (TP + FN) \quad (10)$$

$$F_1 = 2 \cdot Precision \cdot Recall / (Precision + Recall) \quad (11)$$

Two threshold-free measures are additionally reported. AUC summarises the ranking quality of the score across all thresholds. AUPRC — average precision — summarises the precision–recall trade-off and is the more informative of the two under class imbalance, since it does not credit a model for correctly ranking the abundant negative class [19], [20]. The interpretive baseline for AUPRC is the minority prevalence, here 0.2838; any value above this indicates useful ranking. Accuracy is reported for completeness and for comparability with prior work, but is not used to support any conclusion.

5.3 Implementation environment and hyper-parameter configuration

All experiments were implemented in Python 3 using scikit-learn [24] for preprocessing, partitioning, KNN and the metric suite; the XGBoost library [16] for gradient boosting; TensorFlow/Keras [25] for the convolutional network; the SHAP library [18] for TreeSHAP attribution; Gradio [32] for the interface; and ReportLab for PDF generation. Every hyper-parameter is stated in Table 4 so that the study is reproducible without reference to library defaults, addressing directly the reporting vagueness criticised in [2].

Table 4. Complete hyper-parameter and configuration specification.

Component	Parameter	Value	Rationale
Partitioning	test_size / stratify / random_state	0.20 / y / 42	Preserves class prior; reproducible
Numeric scaling	StandardScaler (6 attributes)	fit on train only	Prevents preprocessing leakage
Categorical encoding	OneHotEncoder, handle_unknown	ignore	Tolerates unseen locations at inference
Encoded dimensionality	total features after transform	7,998	6 numeric + 7,992 indicator columns
KNN	n_neighbors	7	Odd value; avoids ties in binary voting
KNN	metric / p / weights	Minkowski / 2 / uniform	Standard Euclidean baseline
CNN	architecture	Conv1D(32,3) → MaxPool(2) → Conv1D(64,3) → Flatten → Dense(64) → Dense(1)	Two-stage local feature extraction
CNN	activations	ReLU / sigmoid (output)	Standard for binary classification
CNN	optimiser / loss	Adam / binary cross-entropy	Adaptive first-order optimisation [23]
CNN	epochs / batch size	10 / 64	Convergence observed without overfitting

Component	Parameter	Value	Rationale
CNN	class weighting	none	Deliberate: exposes unweighted behaviour
XGBoost	n_estimators	200	Sufficient for convergence at $\eta = 0.05$
XGBoost	max_depth	6	Controls interaction order; limits overfitting
XGBoost	learning_rate (η)	0.05	Conservative shrinkage for stability
XGBoost	subsample / colsample_bytree	0.8 / 0.8	Stochastic regularisation
XGBoost	objective / eval_metric	binary:logistic / logloss	Calibrated probability output
XGBoost	random_state	42	Reproducibility
Decision layer	τ (KNN, CNN)	0.50	Default-threshold baseline
Decision layer	τ^* (XGBoost, tuned)	0.135	Maximises validation F_1
Explainability	shap.TreeExplainer, top-k	exact / k = 10	Exact Shapley values in polynomial time

6. Results and Discussion

6.1 Comparative performance

Table 5 reports the complete metric suite for all four configurations on the held-out validation partition. Three reference values frame the interpretation. The majority-class baseline accuracy is 0.7160, obtained by labelling every transaction legitimate. The AUPRC baseline is the minority prevalence, 0.2838. Balanced accuracy and the Matthews correlation coefficient (MCC) are reported in addition to the conventional metrics because both are insensitive to class prior and therefore provide an independent check on the F1-based ranking.

Table 5. Comparative performance of all configurations on the held-out validation partition (n = 2,000; 1,432 legitimate, 568 fraudulent).

Model	τ	Acc.	Prec.	Recall	F_1	Bal. acc.	MCC	AUC	AUPRC
Majority-class baseline	—	0.7160	—	0.0000	0.0000	0.5000	0.0000	0.5000	0.2838
KNN	0.500	0.6995	0.4614	0.3468	0.3960	0.5931	0.2049	n/r	0.4042
CNN	0.500	0.7170	1.0000	0.0035	0.0070	0.5018	0.0502	0.7284	0.4813
XGBoost	0.500	0.7110	0.4864	0.3151	0.3825	0.5916	0.2132	0.7369	0.4556
XGBoost (tuned)	0.135	0.5595	0.3771	0.8451	0.5215	0.6456	0.2731	0.7369	0.4556

Note: "n/r" indicates a value not recorded in the experimental run. AUC and AUPRC are threshold-free and therefore identical for the two XGBoost rows.

The table contains three results that merit separate treatment. First, the model with the highest accuracy is the model with the lowest recall. Second, the accuracy of the two models with the strongest minority-class performance falls below the trivial majority-class baseline. Third, the model with the highest AUPRC is the model that detects almost nothing at its operating point. Figure 5 visualises the first two of these.

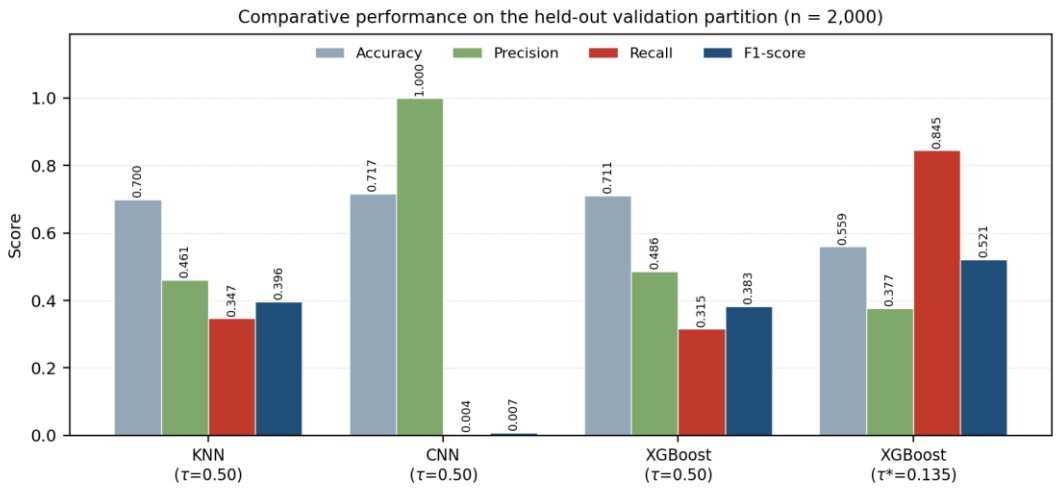

Figure 5. Comparative performance across the four configurations. The inverse relationship between accuracy and recall is visible directly.

6.2 Confusion-matrix analysis

Aggregate metrics compress the error structure; the confusion matrices in Figure 6 expose it. Each matrix is computed over the same 2,000 validation records, of which 1,432 are legitimate and 568 fraudulent.

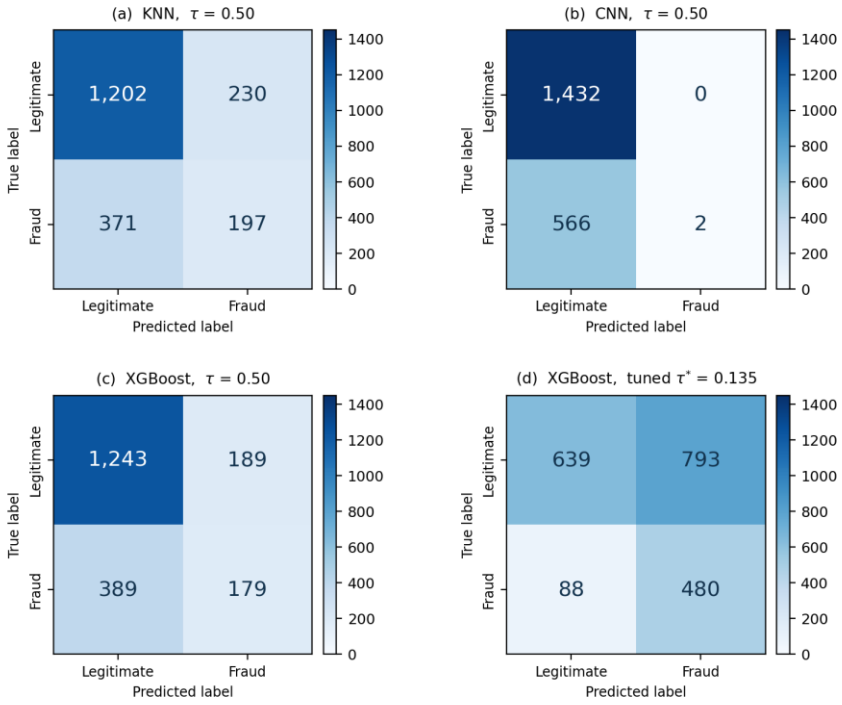

Figure 6. Confusion matrices on the held-out validation partition (n = 2,000; 1,432 legitimate, 568 fraudulent). (a) KNN at τ = 0.50; (b) CNN at τ = 0.50; (c) XGBoost at the default threshold τ = 0.50; (d) XGBoost at the tuned threshold τ = 0.135.*

KNN (Figure 6a) recovers 197 of 568 fraudulent transactions while raising 230 false alarms. Its specificity of 0.8394 is the lowest of the three default-threshold configurations, indicating that its errors are distributed across both classes rather than concentrated in the minority class. This behaviour is consistent with the geometry of the encoded space described in Section 4.1: with 7,984 near-singleton location indicators, most pairs of records differ in their

location dimensions regardless of class, and the Euclidean metric is dominated by a component that carries no class information. The result is a partial but noisy neighbourhood signal — which is precisely what a diagnostic baseline is included to reveal.

XGBoost at the default threshold (Figure 6c) shows a different error profile: 179 true positives, 189 false positives and a specificity of 0.8680. Its precision (0.4864) exceeds that of KNN, but its recall (0.3151) is lower, and its F_1 (0.3825) is marginally below KNN's (0.3960). This is a result that should be stated plainly rather than elided: at the conventional operating point, the ensemble does not dominate the instance-based baseline on the fraud class. The tree structure allows XGBoost to ignore the uninformative location indicators entirely, which yields better-calibrated probability estimates and a higher AUC, but that advantage is not realised at $\tau = 0.50$ because the default threshold is misaligned with the decision problem. The advantage becomes decisive only once the operating point is corrected, which is the central methodological claim of this paper.

6.3 The CNN pathology: high accuracy, high AUPRC, zero utility

The CNN result (Figure 6b) is the study's most instructive finding and deserves careful statement. The network classifies 1,998 of 2,000 validation records as legitimate. It produces 1,432 true negatives, zero false positives, 566 false negatives and 2 true positives. Its precision is therefore a nominally perfect 1.0000 — both of its positive predictions are correct — while its recall is 0.0035 and its F_1 is 0.0070. Its accuracy, 0.7170, exceeds the majority-class baseline of 0.7160 by exactly one part in a thousand, which is the contribution of those two correctly identified transactions. In operational terms, the model that reports the highest accuracy in Table 5 would allow 566 of 568 fraudulent transactions to settle.

Two mechanisms explain this. Unweighted binary cross-entropy over a partition in which 71.6 % of records are legitimate is minimised most efficiently by driving the sigmoid output below 0.5 for nearly all inputs; with only ten epochs and no class weighting, gradient descent converges to that solution. The 7,998-dimensional sparse input compounds the effect, since a one-dimensional convolution over an arbitrarily ordered feature vector has no meaningful local structure to exploit — the convolutional inductive bias assumes spatial or temporal adjacency that a one-hot encoded tabular vector does not possess.

The critical observation, however, concerns AUPRC. The CNN records the highest AUPRC in the study, 0.4813, above XGBoost (0.4556) and KNN (0.4042) and well above the 0.2838 prevalence baseline. Its ROC AUC of 0.7284 (Figure 7) likewise indicates non-trivial discriminative capability. The network has therefore learned a genuinely useful ranking over transactions; it has simply failed to place any of that ranking on the correct side of the decision boundary. This dissociation between ranking quality and decision quality is, in our view, a stronger and more precise argument against accuracy-led evaluation than the conventional example of majority-class collapse — because it demonstrates that even threshold-free metrics such as AUPRC, widely recommended as the imbalance-robust alternative [19], [20], can be entirely satisfactory while the deployed classifier detects nothing. Neither accuracy nor AUPRC alone is sufficient. The operating point must be reported.

It follows that the appropriate remedy for the CNN is not architectural but decisional: the same threshold recalibration applied to XGBoost in Section 6.4 would convert this ranking into usable detections. We did not apply it here, in order to preserve the contrast between a default-threshold and a calibrated-threshold pipeline; doing so is identified as future work in Section 8.

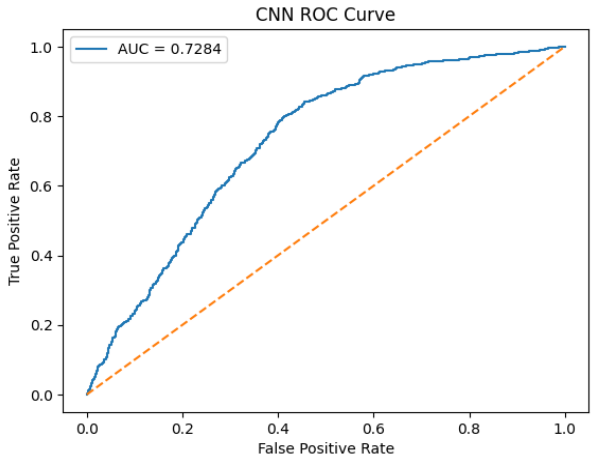

Figure 7. ROC curve of the CNN model (AUC = 0.7284). The network ranks transactions informatively despite detecting only 2 of 568 fraudulent cases at τ = 0.50.

6.4 Effect of threshold recalibration

Lowering the XGBoost decision threshold from 0.50 to 0.135 changes no model parameter; it changes only the point at which the existing score is converted into an action. The consequences, obtained by comparing Figures 6(c) and 6(d), are substantial in both directions and are reported here jointly, as [2] argues they must be.

Table 6. Effect of decision-threshold recalibration on XGBoost (validation partition, n = 2,000).

Quantity	τ = 0.50	τ* = 0.135	Change
True positives	179	480	+301 (+168 %)
False negatives	389	88	−301 (−77.4 %)
False positives	189	793	+604 (×4.20)
True negatives	1,243	639	−604 (−48.6 %)
Recall	0.3151	0.8451	+0.5300
Precision	0.4864	0.3771	−0.1093
F_1-score	0.3825	0.5215	+0.1390
Balanced accuracy	0.5916	0.6456	+0.0540
MCC	0.2132	0.2731	+0.0599
Accuracy	0.7110	0.5595	−0.1515
Alerts raised (share of stream)	368 (18.4 %)	1,273 (63.7 %)	+905
Alerts reviewed per fraud detected	2.06	2.65	+0.59

The gains are real. Recall rises by 0.53 in absolute terms, false negatives fall by 77.4 %, and both prior-insensitive measures — balanced accuracy and MCC — improve, confirming that the change is not an artefact of the F_1 criterion used to select τ*. The marginal exchange rate is also more favourable than the headline figures suggest: the 604 additional false positives buy 301 additional detections, a marginal cost of 2.01 false alarms per additional fraud intercepted, which in most cost structures would be an acceptable trade.

The absolute alert volume, however, is not defensible, and we state this explicitly rather than leaving it to be inferred. At τ* = 0.135 the system flags 1,273 of 2,000 transactions, or 63.7 % of the stream. Precision falls to 0.3771, meaning fewer than two in five alerts are genuine, a lift of only 1.33× over the base rate of 0.2838 — lower than the 1.71× lift obtained at τ = 0.50. No fraud operations function could triage a queue containing two-thirds of all transactions. This is precisely the failure mode that Hayat and Magnier [2] characterise as metric manipulation through recall optimisation at precision's expense, and it applies to our own tuned configuration.

We draw two conclusions from this rather than one. Methodologically, the demonstration stands: the operating point, not the model, is the dominant lever on minority-class detection, and treating τ as a first-class system parameter is justified. Operationally, F_1 is the wrong objective for selecting τ in a production fraud system, because it weights precision and recall equally while the operational constraint is an alert budget. A selection rule of the form $\tau_{budget} = \min\{\tau : \text{alert-rate}(\tau) \leq \beta\}$ for an admissible review capacity β, or a cost-sensitive rule minimising $c_{FN} \cdot FN + c_{FP} \cdot FP$ with institution-specific costs [27], would yield a deployable operating point. Section 8 develops this as the principal direction for future work.

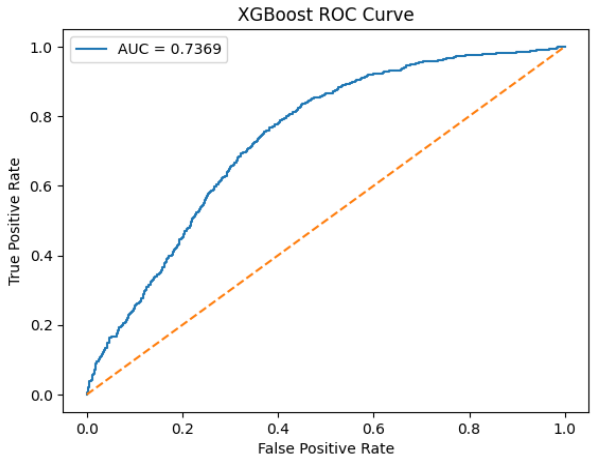

Figure 8. ROC curve of the XGBoost model (AUC = 0.7369). The threshold-free ranking quality is unchanged by recalibration; only the operating point on this curve moves.

Figure 8 makes the mechanism explicit. Threshold recalibration does not alter the ROC curve — the AUC remains 0.7369 for both XGBoost configurations — it selects a different point along it. The 0.7369 figure should itself be read against the ceiling established in Section 3.4: with a feature space whose strongest predictor correlates at $r = 0.339$ with the label, an AUC near 0.74 is close to the information available in the data, and results substantially above this level on this corpus would warrant investigation for leakage rather than celebration.

6.5 Explainability analysis

Figure 9 shows the TreeSHAP attribution produced by the deployed system for the instance displayed in Figure 4. The transaction carries a value of 1,587.96, an account age of 12.3 years, a credit score of 465 and a positive prior-fraud flag, and is scored at $\hat{p} = 0.4145$ — above $\tau^* = 0.135$ and therefore adjudicated as fraudulent.

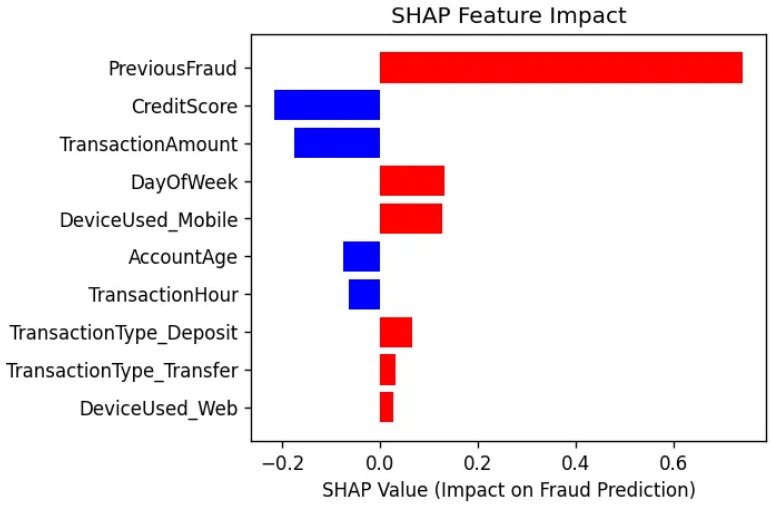

Figure 9. TreeSHAP feature attribution for the scored instance of Figure 4. Positive (red) contributions increase and negative (blue) contributions decrease the fraud score.

PreviousFraud dominates the attribution at approximately +0.74, an order of magnitude larger than any other contribution. CreditScore (≈ −0.22) and TransactionAmount (≈ −0.17) act in the opposing direction, reflecting that a score of 465 and a moderate transaction value are, for this ensemble, mildly exculpatory. DayOfWeek and DeviceUsed_Mobile contribute positively at approximately +0.13 each; the remaining features contribute below 0.08 in magnitude.

This ranking can be validated independently, which is a stronger claim than the qualitative plausibility assessments typical of the literature. The three features that dominate the attribution — PreviousFraud, CreditScore and TransactionAmount — are exactly the three features that Table 3 identified as carrying non-negligible marginal association with the label, in the same rank order and with the same signs (positive for PreviousFraud and TransactionAmount at the corpus level, negative for CreditScore). The agreement between a local, model-internal attribution and an independent, model-free global statistic provides evidence that the explanation reflects genuine structure in the data rather than an artefact of the fitting procedure.

One divergence is instructive and is reported rather than omitted. DayOfWeek receives a non-trivial positive attribution for this instance despite a corpus-level correlation of only +0.0114. This is not an inconsistency: Shapley values are local quantities describing a single prediction within the ensemble's learned interaction structure, whereas point-biserial correlation is a global marginal statistic. A feature with no marginal signal may still participate in higher-order interactions that matter for particular instances. This distinction is frequently blurred in applied SHAP reporting, and it has direct operational consequence — an analyst must not read a large local attribution as evidence of global feature importance.

6.6 Deployment behaviour

The interface of Figure 4 exposes the complete feature schema, permits model selection at inference time, and returns the predicted label, the fraud probability (0.4145), the validation accuracy of the selected model and the decision confidence, together with the textual and graphical explanation. Explanation controls are disabled automatically when a model without exact attribution support is selected, so that the analyst is never presented with an explanation whose reliability differs from what the interface implies. Each adjudication is serialised into a timestamped PDF containing the full input vector, the fraud probability, the predicted label and the model identifier — the minimum record required to reconstruct a decision during post-incident review. Among the fifteen studies reviewed in Section 2, none reports an equivalent audit artefact.

6.7 Positioning against the reviewed literature

Direct numerical comparison with the studies of Table 1 would be misleading, since they evaluate on different corpora with different class priors and different feature semantics; recall of 0.95 on a corpus with 0.172 % fraud [3] and recall of 0.845 on a corpus with 28.38 % fraud are not commensurable quantities. The comparison that can legitimately be drawn is methodological.

Our framework converges with Baisholan et al. [3] on the two decisions we regard as most consequential: preserving the natural class distribution rather than resampling it, and treating threshold optimisation as the primary imbalance remedy. It diverges from [3] in one respect that favours the present work — our features retain their original business semantics, whereas the PCA-anonymised features of the European corpus render SHAP attributions semantically opaque, a limitation those authors acknowledge. Relative to Theodorakopoulos et al. [4] and Andrade-Arenas and Yactayo-Arias [15], both of which report at unstated default thresholds, this study adds the operating-point analysis that converts a metric into an operational statement. Relative to Vijayanand and Smrithy [14], who report 99.90 % accuracy on the synthetic PaySim corpus, our contribution is precisely the opposite in emphasis: we demonstrate that such an accuracy figure would not, on its own, establish anything about fraud detection capability.

The graph-based approaches of [11], [12] and [13] are not applicable to the present corpus, since it provides no entity-linkage structure — locations are near-unique and customers recur on average 1.67 times — and we therefore make no claim of superiority over them. Their relational modelling addresses a class of fraud that flat transaction logs cannot express, and integration of such structure is identified in Section 8 as a natural extension.

6.8 Summary of findings

- Accuracy is not merely uninformative but actively inverted on this problem: the highest-accuracy model has the lowest recall, and both models with the strongest fraud-class performance score below the trivial majority-class baseline of 0.7160.
- Threshold-free ranking metrics are also insufficient in isolation. The CNN attains the highest AUPRC (0.4813) in the study while detecting 2 of 568 fraudulent transactions. Ranking quality and decision quality are distinct properties and must be reported separately.
- Decision-threshold calibration is the dominant lever available without model modification: recall rises from 0.3151 to 0.8451 and F_1 from 0.3825 to 0.5215 with no change to any learned parameter.
- That gain is not free, and the cost is reported. Alert volume rises to 63.7 % of the transaction stream, which is not operationally viable; F_1 is therefore the wrong criterion for selecting a production operating point, and an alert-budget or cost-sensitive criterion is required.
- Prior-insensitive metrics corroborate the ranking. Balanced accuracy (0.6456) and MCC (0.2731) are both maximised by the tuned XGBoost configuration and minimised by the CNN, independently of the F_1 criterion used for selection.
- TreeSHAP attributions agree in rank order and sign with independently computed point-biserial correlations, supporting the fidelity of the explanation layer while illustrating that local attributions and global importance remain distinct quantities.

7. Threats to Validity and Limitations

We state the limitations of this study explicitly and in the terms a reviewer would apply, since the argument of Section 6 depends on holding our own results to the standard we invoke against the literature.

7.1 Threshold selection and reporting on the same partition

The threshold $\tau^* = 0.135$ was selected by maximising F_1 on the same 2,000-record partition on which the tuned performance is subsequently reported. This introduces an optimistic bias of unknown but non-zero magnitude in the tuned XGBoost row of Table 5. The correct protocol is a three-way split — train, calibration and test — in which τ^* is chosen on the calibration partition and reported on a test partition never used for any selection decision, or equivalently nested cross-validation with the threshold selected within each inner fold. This is the single most important methodological correction required before the tuned figures could be considered confirmed, and it does not affect the default-threshold results, which involve no selection.

7.2 Corpus provenance and generalisation

As documented in Section 3.1, the corpus exhibits properties consistent with synthetic generation: near-uniform categorical marginals, a near-injective location field and a near-balanced prior-fraud flag. Absolute performance levels should therefore not be transferred to production payment traffic. The comparative and methodological conclusions — the accuracy inversion, the ranking/decision dissociation, the dominance of the threshold lever — are properties of the evaluation protocol rather than of the corpus and are expected to generalise; the specific numeric values are not.

7.3 Class prior

At 28.38 % minority prevalence, this corpus is moderately rather than severely imbalanced. Behaviour at the 0.1–1 % prevalence characteristic of production card and P2P fraud may differ materially, particularly with respect to the stability of the selected threshold, which becomes increasingly sensitive to sampling variation as the positive class thins.

7.4 Absence of temporal validation

Partitioning was random and stratified rather than chronological. Because fraud exhibits concept drift [12], a random split allows the model to be evaluated on transactions temporally interleaved with its training data, which overstates performance relative to a forward-looking deployment in which the model must generalise to future behaviour. The corpus spans only 46 days, which limits what a chronological split could establish, but the limitation is real and is noted by [2] as one of the four systemic failures in the field.

7.5 Single-run evaluation and absence of statistical testing

All results derive from a single partition at a single random seed. No confidence intervals are reported and no significance testing was performed on the differences between models. The narrow F_1 margin between KNN (0.3960) and XGBoost at $\tau = 0.50$ (0.3825) in particular should not be regarded as established without repeated-seed evaluation; the CNN result, by contrast, is of a magnitude that seed variation could not plausibly explain.

7.6 Encoding of high-cardinality categorical features

One-hot encoding of a 7,984-level location field is not a defensible production choice. It was retained deliberately to hold the representation identical across all three learners, but target encoding, frequency thresholding or hashing would be required in deployment, and the KNN result in particular is partly an artefact of this decision rather than a property of instance-based learning as such.

7.7 Scope of explainability

Exact attribution is provided only for XGBoost. The framework therefore delivers explanations for the deployed model but not for the comparative baselines, and the interface disables explanation controls accordingly. This is a principled restriction, as argued in Section 4.4, but it does mean the framework does not offer model-agnostic explanation.

8. Conclusion and Future Work

This paper presented an explainable, threshold-aware machine learning framework for fraud detection in peer-to-peer digital payment systems and evaluated it on a 10,000-record banking transaction corpus with a moderate class imbalance of 2.52:1. Three structurally distinct learners — KNN, a one-dimensional CNN and XGBoost — were trained under strictly identical preprocessing, partitioning and evaluation conditions, and the decision threshold was treated as an explicit system parameter rather than as a library default.

The principal empirical finding is a dissociation between ranking quality and decision quality. The convolutional network achieved the highest accuracy in the study (0.7170) and the highest area under the precision–recall curve (0.4813), yet identified only 2 of 568 fraudulent transactions at the conventional operating point. It had learned an informative ordering over transactions and placed essentially none of it on the correct side of the decision boundary. This result establishes something stronger than the familiar warning against accuracy on imbalanced data: it shows that the threshold-free metrics recommended as the remedy can also be entirely satisfactory while the deployed classifier detects nothing. The operating point must be reported alongside any aggregate metric, and any published fraud detection result that omits it is under-specified.

The second finding is that the decision threshold is the dominant available lever. Recalibrating XGBoost from $\tau = 0.50$ to $\tau^* = 0.135$ raised recall from 0.3151 to 0.8451 and F_1 from 0.3825 to 0.5215 without altering a single learned parameter, with balanced accuracy and MCC confirming the improvement independently of the selection criterion. The third finding qualifies the second and is reported with equal prominence: that recalibration drove alert volume to 63.7 % of the transaction stream and precision to 0.3771, a configuration no fraud operations function could sustain. F_1 is consequently the wrong objective for selecting a production operating point, and we regard the correct formulation as a constrained one — the highest recall attainable subject to an institutional alert budget.

The framework further demonstrated that exact TreeSHAP attribution can be delivered at inference latency within an analyst-facing interface, and that the resulting attributions agree in rank and sign with independently computed marginal statistics. The Gradio interface and automated PDF audit reporting close the gap, unaddressed in all fifteen reviewed studies, between an offline metric table and a decision an investigator can act upon and later defend.

Five directions follow from the limitations of Section 7 and are ordered by the priority we assign to them.

(1) Corrected threshold protocol. Adopt a three-way train/calibration/test split or nested cross-validation so that τ^* is selected without contaminating the reported performance, and report threshold stability across repeated seeds.
(2) Cost-sensitive and budget-constrained operating points. Replace the F_1 selection criterion with a cost-sensitive objective minimising $c_{FN} \cdot FN + c_{FP} \cdot FP$ under institution-specific costs [27], and with a budget-constrained rule that maximises recall subject to an admissible alert rate. Reporting a recall-versus-alert-rate curve rather than a single operating point would make results directly actionable.

(3) Threshold calibration for the CNN. The dissociation identified in Section 6.3 implies that the network's ranking is recoverable. Applying the same recalibration, together with class-weighted loss and probability calibration, would test whether the deep model becomes competitive once its operating point is corrected.
(4) Temporal and larger-scale validation. Re-evaluate under chronological partitioning on a longer-horizon corpus with production-realistic prevalence, and under explicit concept drift, following [12].
(5) Extension to UPI-specific and relational settings. Incorporate the fraud typologies catalogued for UPI by Chakka and Saheb [1] — collect-request abuse, QR manipulation and KYC-pretext social engineering — and enrich the flat transaction representation with entity-linkage structure so that graph-based methods [11], [13] become applicable alongside the tabular ensemble.

More broadly, the study supports a reorientation of evaluation practice in this field. The question that determines whether a fraud detection model is useful is not how accurately it classifies, nor even how well it ranks, but where its decision boundary is placed and what that placement costs. Until that question is answered explicitly in published work, reported performance will remain difficult to interpret and harder still to deploy.

Declarations

Acknowledgements

The authors thank Gyanmanjari Institute of Technology, Gyanmanjari Innovative University, for institutional support and computational resources.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author contributions

D. Chandarana: conceptualisation, methodology, software implementation, experimentation, formal analysis, writing — original draft. H. Nimbark: supervision, methodology review, validation, writing — review and editing. Both authors have read and approved the final manuscript.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

The banking transaction dataset analysed in this study, together with the implementation notebook containing the full preprocessing, training, evaluation, explainability and interface code, is available from the corresponding author upon reasonable request.

List of abbreviations

AUC — Area Under the Receiver Operating Characteristic Curve; AUPRC — Area Under the Precision–Recall Curve; BCE — Binary Cross-Entropy; CNN — Convolutional Neural Network; FN — False Negative; FP — False Positive; GNN — Graph Neural Network; KNN — K-Nearest Neighbours; MCC — Matthews Correlation Coefficient; P2P — Peer-to-Peer; PRISMA — Preferred Reporting Items for Systematic Reviews and Meta-Analyses; ROC — Receiver Operating Characteristic; SHAP — SHapley Additive exPlanations; SMOTE — Synthetic Minority Oversampling Technique; TN — True Negative; TP — True Positive; UPI — Unified Payments Interface; XAI — Explainable Artificial Intelligence; XGBoost — Extreme Gradient Boosting.

References

1. N. B. Chakka and S. S. Saheb, "Mobile payment fraud detection in UPIs through machine learning techniques: A systematic review," Multidisciplinary Reviews, vol. 9, no. 6, 2026280, 2025.
2. K. Hayat and B. Magnier, "Data leakage and deceptive performance: A critical examination of credit card fraud detection methodologies," Mathematics, vol. 13, no. 16, 2563, 2025, doi: 10.3390/math13162563.
3. N. Baisholan, J. E. Dietz, S. Gnatyuk, M. Turdalyuly, E. T. Matson and K. Baisholanova, "FraudX AI: An interpretable machine learning framework for credit card fraud detection on imbalanced datasets," Computers, vol. 14, no. 4, 120, 2025, doi: 10.3390/computers14040120.

4. L. Theodorakopoulos, A. Theodoropoulou, A. Tsimakis and C. Halkiopoulos, "Big data-driven distributed machine learning for scalable credit card fraud detection using PySpark, XGBoost, and CatBoost," Electronics, vol. 14, no. 9, 1754, 2025, doi: 10.3390/electronics14091754.
5. B. Yousefimehr and M. Ghatee, "A distribution-preserving method for resampling combined with LightGBM-LSTM for sequence-wise fraud detection in credit card transactions," Expert Systems with Applications, vol. 262, 125661, 2025, doi: 10.1016/j.eswa.2024.125661.
6. M. Vasconcelos and L. Cavique, "Mitigating false negatives in imbalanced datasets: An ensemble approach," Expert Systems with Applications, vol. 262, 125674, 2025, doi: 10.1016/j.eswa.2024.125674.
7. Q. Zeng, L. Lin, R. Jiang, W. Huang and D. Lin, "NNEnsLeG: A novel approach for e-commerce payment fraud detection using ensemble learning and neural networks," Information Processing & Management, vol. 62, 103916, 2025, doi: 10.1016/j.ipm.2024.103916.
8. L. Zhang, Y. Xuan, Z. Liu, Z. Du, S. Wang and J. Wang, "A hybrid ensemble model to detect Bitcoin fraudulent transactions," Engineering Applications of Artificial Intelligence, vol. 141, 109810, 2025, doi: 10.1016/j.engappai.2024.109810.
9. Z. Wang, X. Chen, Y. Wu, L. Jiang, S. Lin and G. Qiu, "A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud," Scientific Reports, vol. 15, 218, 2025, doi: 10.1038/s41598-024-82062-x.
10. M. Akouhar, M. Ouhssini, M. El-Fatini, A. Abarda and E. Agherrabi, "Dynamic oversampling-driven Kolmogorov–Arnold networks for credit card fraud detection: An ensemble approach to robust financial security," Egyptian Informatics Journal, vol. 31, 100712, 2025.
11. F. K. Alarfaj and S. Shahzadi, "Enhancing fraud detection in banking with deep learning: Graph neural networks and autoencoders for real-time credit card fraud prevention," IEEE Access, vol. 13, pp. 20633–20646, 2025.
12. Y. Cui, X. Han, J. Chen et al., "FraudGNN-RL: A graph neural network with reinforcement learning for adaptive financial fraud detection," IEEE Open Journal of the Computer Society, vol. 6, pp. 426–437, 2025.
13. X. Wang and Y. Wang, "Real-time transaction flow analysis with graph neural networks for financial fraud detection," Journal of Computational Methods in Sciences and Engineering, 2025, doi: 10.1177/14727978251385133.
14. D. Vijayanand and G. S. Smrithy, "Explainable AI-enhanced ensemble learning for financial fraud detection in mobile money transactions," Intelligent Decision Technologies, 2025, doi: 10.1177/18724981241289751.
15. L. Andrade-Arenas and C. Yactayo-Arias, "Comparative analysis of machine learning models for credit card fraud detection using SMOTE for class imbalance," International Journal of Safety and Security Engineering, vol. 15, pp. 893–901, 2025.
16. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
17. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.
18. S. M. Lundberg, G. Erion, H. Chen et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
19. T. Saito and M. Rehmsmeier, "The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, e0118432, 2015.
20. J. Davis and M. Goadrich, "The relationship between Precision–Recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning, Pittsburgh, PA, USA, 2006, pp. 233–240.
21. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
22. Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
23. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, 2015.
24. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
25. M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI), Savannah, GA, USA, 2016, pp. 265–283.
26. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
27. C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artificial Intelligence (IJCAI), Seattle, WA, USA, 2001, pp. 973–978.
28. A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784–3797, 2018.
29. I. D. Mienye and N. Jere, "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions," IEEE Access, vol. 12, pp. 96893–96910, 2024, doi: 10.1109/ACCESS.2024.3426955.
30. National Payments Corporation of India, "UPI product statistics." [Online]. Available: https://www.npci.org.in/what-we-do/upi/product-statistics

31. B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta (BBA) – Protein Structure, vol. 405, no. 2, pp. 442–451, 1975.
32. A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan and J. Zou, "Gradio: Hassle-free sharing and testing of ML models in the wild," arXiv:1906.02569, 2019.

Author Biographies

Dhaval Chandarana is a faculty member & a researcher at the Gyanmanjari Institute of Technology, Gyanmanjari Innovative University, Bhavnagar, Gujarat, India. His research interests include applied machine learning, explainable artificial intelligence and the design of deployable analytics systems for financial services, with a particular focus on fraud detection in digital and peer-to-peer payment infrastructures.

Dr. Hitesh Nimbark is the Chief Executive Officer and Provost of Gyanmanjari Innovative University, Bhavnagar, Gujarat, India. His research interests span artificial intelligence, machine learning, data mining and information systems security, and he has supervised research on intelligent systems for financial and industrial applications.