Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

Machine Learning-Driven Risk Scoring Systems for Improved Fraud Prevention in E-Commerce

Dr. Ghamendra Kumar Sahu
Faculty of Science
Govt. Rani Avanti Bai Lodhi College
Ghumka, Dist-Rajnandhaon, Chhattisgarh, 491444
Orcid id- https://orcid.org/0009-0000-6232-1510

http://www.wjcrt.org/ || Vol. 2 No. 3 (2026): September Issue

Date of Submission: 05-08-2026 **Date of Acceptance:** 19-08-2026 **Date of Publication:** 05-09-2026

Abstract— The rapid expansion of e-commerce has increased the operational complexity of fraud prevention by creating high-volume, high-velocity, and behaviorally diverse transaction environments. Conventional rule-based controls and static machine-learning classifiers often struggle when fraudulent behavior evolves faster than predefined detection logic or historical training patterns. This study investigates a machine learning-driven risk scoring framework designed to estimate transaction-level fraud propensity while incorporating behavioral, transactional, temporal, device, and account-level signals. The central research gap concerns the limited integration of calibrated risk scores with cost-sensitive fraud intervention, particularly in environments characterized by severe class imbalance and concept drift. The proposed framework emphasizes probabilistic scoring rather than simple fraud-versus-legitimate classification, enabling transactions to be categorized into differentiated intervention bands. Its design combines ensemble learning, temporal feature construction, probability calibration, and threshold optimization to support adaptive fraud screening while reducing unnecessary rejection of legitimate customers. The study further proposes an experimental protocol that evaluates both predictive discrimination and decision quality through calibration error, expected fraud loss, false-positive burden, and ranking-oriented measures.
Keywords— E-commerce fraud detection, Machine learning, Risk scoring, Probability calibration, Cost-sensitive learning, Transaction analytics

Introduction

E-commerce ecosystems depend on the ability to approve legitimate transactions rapidly while identifying fraudulent activity before financial or reputational damage occurs. This balance has become increasingly difficult as online commerce expands across mobile devices, digital wallets, marketplaces, subscription platforms, social commerce channels, and cross-border payment systems. Fraudulent transactions may resemble legitimate purchases at the point of authorization, while sophisticated attackers can manipulate account credentials, device identities, delivery information, payment instruments, browsing patterns, and transaction timing to avoid detection. As a result, fraud prevention is no longer adequately represented as a fixed set of suspicious conditions. It is increasingly a problem of estimating continuously changing risk under uncertainty.

Traditional fraud controls in e-commerce have relied heavily on expert-defined rules. Examples include blocking unusually large transactions, identifying repeated payment failures, flagging orders from unusual locations, or restricting transactions involving mismatched billing and shipping information. Rule-based systems remain useful because they are transparent and operationally simple. However, their effectiveness decreases when attackers adapt to known thresholds or distribute suspicious characteristics across several apparently normal transactions. Rules also tend to generate excessive alerts when applied to heterogeneous customer populations. A transaction that is highly unusual for one customer may be entirely normal for another.

Machine learning has transformed fraud detection by allowing predictive systems to discover nonlinear relationships among multiple risk indicators. Models such as random forests, gradient boosting machines, neural networks, support vector machines, and anomaly-detection algorithms can analyze large collections of historical transactions and estimate the probability that a new transaction is fraudulent. More recent fraud-prevention architectures increasingly combine supervised models with behavioral analytics, graph-

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

derived features, device intelligence, and streaming decision systems. These developments have improved the capability of automated systems to detect patterns that are difficult to encode manually.

Nevertheless, a critical methodological issue remains unresolved. Many fraud studies continue to formulate the problem primarily as binary classification: each transaction is predicted as either fraudulent or legitimate. This formulation is useful for evaluating algorithms but does not fully reflect the operational requirements of an e-commerce fraud management system. Fraud teams rarely act on a single yes-or-no prediction. Instead, transactions may be automatically approved, subjected to additional authentication, referred for manual review, temporarily held, or rejected entirely. These differentiated actions require a reliable estimate of *risk magnitude*, not merely a class label.

A risk score represents the estimated likelihood or expected severity of fraud associated with a transaction. For example, two transactions may both be classified as suspicious while having substantially different estimated probabilities of fraud. A transaction with a moderate risk score may justify additional authentication, whereas a transaction with extremely high estimated risk may justify immediate rejection. If model probabilities are poorly calibrated, however, the numerical score may not correspond to the actual incidence of fraud. This creates a practical problem: a model can achieve strong ranking performance while still producing unreliable probabilities.

This distinction is particularly important in e-commerce because the costs of prediction errors are asymmetric. A false negative can lead to chargebacks, merchandise loss, investigation cost, account compromise, and customer distrust. A false positive can also be damaging because it may block a legitimate purchase, frustrate customers, increase cart abandonment, and reduce long-term customer value. Therefore, an effective fraud prevention system should not simply maximize predictive accuracy. It should optimize risk-sensitive decisions under unequal error costs.

Another major challenge concerns class imbalance. In most transactional environments, legitimate transactions greatly outnumber fraudulent ones. A model trained without addressing this imbalance may achieve apparently high accuracy while providing weak fraud detection. Metrics such as overall accuracy are therefore inadequate when evaluated in isolation. Precision, recall, F-scores, precision-recall area, ranking quality, expected cost, and calibration-based metrics provide a more informative representation of model usefulness.

Fraud behavior is also non-stationary. Attack patterns change in response to emerging payment methods, security controls, seasonal purchasing behavior, platform policy changes, data breaches, and attacker experimentation. Consequently, a model that performs strongly during development may deteriorate when deployed on future transactions. This phenomenon, commonly described as concept drift, is particularly important in fraud analytics because the adversarial population actively adjusts its strategies. Static validation based on random train-test splits can underestimate this difficulty because future transactions may not resemble randomly sampled historical records.

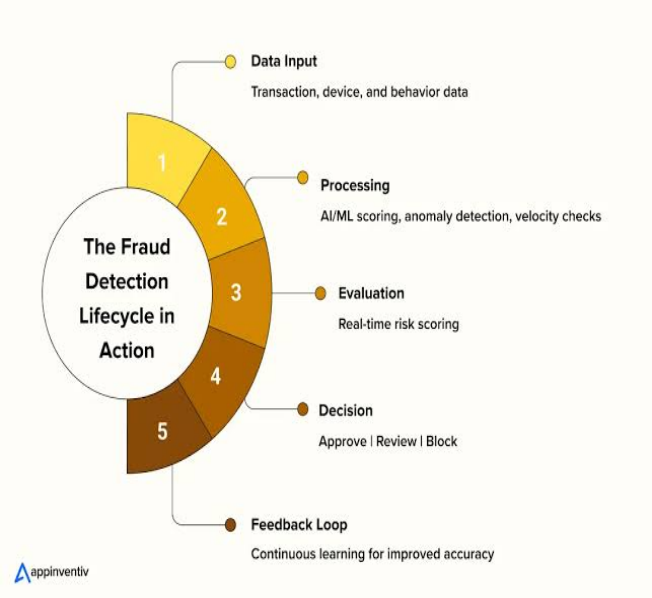


Fig. 1: eCommerce Fraud Detection

Source: *https://appinventiv.com/blog/fraud-management-for-ecommerce-businesses/*

The present study addresses these limitations by conceptualizing e-commerce fraud prevention as a **dynamic risk-scoring and intervention optimization problem**. Instead of generating a single binary output, the proposed framework estimates transaction-level fraud probabilities and subsequently calibrates these probabilities so that the numerical scores better correspond to observed fraud incidence. The risk score is then mapped to operational decision bands using cost-sensitive threshold optimization.

The distinctive research gap addressed by this study lies at the intersection of three issues that are frequently examined separately: predictive fraud classification, probability calibration, and economically informed transaction intervention. Existing fraud-detection research often focuses on maximizing discrimination metrics such as area under the receiver operating characteristic curve, while practical fraud systems require trustworthy scores that can support different levels of intervention. A high-performing classifier is not necessarily an effective risk-scoring system if its probabilities are systematically overconfident or underconfident.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

Accordingly, this research proposes a machine-learning risk-scoring architecture integrating temporal transaction behavior, account activity, device consistency, payment attributes, purchasing context, and velocity indicators. The framework is designed to distinguish low-, medium-, high-, and critical-risk transactions rather than forcing all cases into a binary decision at the first stage. The methodological emphasis is placed on gradient-boosted ensemble learning combined with probability calibration and cost-sensitive threshold selection. Temporal validation is incorporated to test whether the risk model remains useful when applied to later transaction periods.

The principal objective of the study is therefore to determine whether calibrated machine-learning risk scores can improve fraud prevention decisions relative to conventional binary classification. More specifically, the study examines whether the proposed system can maintain strong fraud-ranking performance, improve the reliability of predicted probabilities, reduce false-positive burden, and support more economically rational intervention thresholds.

The study makes several contributions. First, it distinguishes between *discrimination* and *calibration*, two properties that are often incorrectly treated as interchangeable in fraud analytics. Second, it incorporates operational fraud cost into risk-threshold selection rather than selecting thresholds solely from conventional classification metrics. Third, the framework explicitly accounts for temporal behavioral change using time-ordered validation. Finally, it proposes a risk-band architecture that is more compatible with real fraud-management workflows than a single binary decision boundary.

From an operational perspective, such an architecture can be integrated into transaction authorization pipelines in which risk scores are computed immediately after feature extraction. Low-risk transactions may proceed without additional friction, moderate-risk transactions may trigger step-up verification, high-risk cases may be referred for review, and extreme-risk transactions may be declined. The resulting system therefore shifts fraud analytics from retrospective classification toward proactive risk-based decision support.

Literature Review

2.1 Evolution of Fraud Detection in E-Commerce

Fraud detection has evolved from deterministic controls toward data-driven predictive systems. Early fraud prevention mechanisms relied primarily on blacklists, static transaction limits, verification rules, and manually designed suspicious-pattern indicators. Such approaches offered clear decision logic but were limited by their inability to capture interactions among multiple variables.

Machine learning expanded the analytical scope of fraud detection by identifying complex patterns across payment amount, transaction frequency, customer history, merchant attributes, device characteristics, and geographic information. Supervised classification became particularly prominent because confirmed fraudulent and legitimate transactions could be used as historical labels. Logistic regression, decision trees, random forests, support vector machines, and boosting methods have therefore been widely applied to transaction fraud problems.

The increasing complexity of online commerce has encouraged the development of hybrid systems. Rather than relying exclusively on a single classifier, contemporary fraud systems may incorporate supervised prediction, anomaly detection, behavioral profiling, graph analysis, and rules. This architecture reflects a practical recognition that different fraud mechanisms produce different statistical signatures.

A supervised model is effective when newly observed fraud resembles previously labeled examples. In contrast, anomaly-detection techniques may identify behavior that deviates substantially from historical customer activity even when the specific fraud pattern has not appeared before. Rule-based controls remain useful for enforcing explicit policy constraints. Consequently, the most effective fraud-prevention architectures tend to combine complementary analytical components rather than replacing all conventional controls with one machine-learning model.

2.2 Transaction Features and Behavioral Risk Indicators

The quality of fraud prediction depends substantially on feature representation. Raw transaction attributes such as payment amount, time, location, product type, account age, and payment method provide useful information but may be insufficient when considered independently.

Behavioral features often provide stronger fraud signals because they contextualize current activity relative to previous customer behavior. For example, a transaction amount may appear normal globally but become suspicious when it is several times larger than the customer's historical purchase pattern. Similarly, a login from a new device may be low risk for a recently created account but more significant for a long-standing account with stable device usage.

Velocity features are especially important in real-time fraud scoring. They describe the number, frequency, or cumulative value of actions performed within specified time intervals. Examples include the number of payment attempts within ten minutes, the number of different cards used by one account within twenty-four hours, or the number of customer accounts associated with a single device.

Temporal context can also improve risk estimation. Fraud activity may display characteristic patterns around transaction

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

timing, account creation, password resets, delivery-address changes, or unusually rapid sequences of actions. Therefore, effective risk scoring should represent transactions as events embedded within behavioral histories rather than isolated rows in a database.

Device and identity-consistency features have also become increasingly relevant. Mismatch among registered user location, IP region, device fingerprint, billing address, and shipping destination may increase risk. However, such indicators should be interpreted probabilistically rather than through rigid blocking rules because legitimate users also travel, replace devices, and send purchases to different addresses.

2.3 Ensemble Learning for Fraud Risk Prediction

Ensemble models are particularly suitable for structured fraud datasets because they can model nonlinear relationships and high-order feature interactions without requiring extensive manual specification. Random forests reduce variance by averaging predictions across multiple decision trees, while gradient boosting sequentially constructs models that correct the errors of previous estimators.

Gradient-boosted decision tree algorithms have become attractive for financial and transactional risk modelling because they can process heterogeneous feature types, accommodate nonlinear boundaries, and capture interactions among behavioral indicators. Their ability to produce feature-importance information also improves interpretability relative to many black-box deep-learning architectures.

However, raw probabilities generated by ensemble models are not always well calibrated. A model may rank fraudulent transactions effectively while producing probability estimates that exaggerate or underestimate the actual risk. This distinction is central to the present study because operational intervention thresholds depend on the numerical meaning of the risk score.

For example, suppose a model assigns a risk score of 0.80 to a group of transactions. If approximately 80% of those transactions are actually fraudulent, the score is well calibrated. If only 40% are fraudulent, the score is substantially overconfident. Both models may achieve similar ranking performance while leading to very different operational decisions.

2.4 Probability Calibration as an Underexplored Component

Probability calibration concerns the agreement between predicted probabilities and empirical outcome frequencies. It has received significant attention in probabilistic forecasting and clinical risk prediction but remains less emphasized in many fraud-detection studies than ranking metrics.

This gap is consequential because fraud-management policies frequently depend on score thresholds. A business may decide, for example, that transactions above a certain predicted probability require manual review. If probability estimates are unstable or miscalibrated, threshold policies become difficult to interpret and may behave inconsistently over time.

Calibration techniques can be applied after model training. Methods such as Platt-style logistic calibration and isotonic regression transform raw model outputs into probabilities that more closely correspond to observed event frequencies. Calibration should be evaluated on data separate from the model-training sample to prevent optimistic estimation.

The present study therefore treats calibration as a core component rather than a secondary refinement. The proposed architecture separates the modelling process into discrimination, probability calibration, and decision-threshold optimization. This modular structure allows each stage to be evaluated independently.

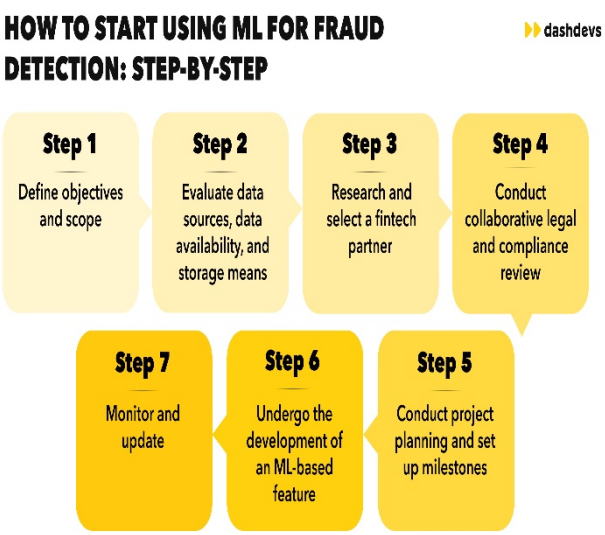


Fig. 2: Using Machine Learning Against Financial Fraud

Source: https://dashdevs.com/blog/beyond-traditional-methods-using-machine-learning-against-financial-fraud/

2.5 Class Imbalance and Metric Selection

Fraud datasets are strongly imbalanced because fraudulent cases generally constitute only a small fraction of all transactions. This imbalance creates both modelling and evaluation difficulties.

Standard accuracy can be misleading because a model that predicts nearly every transaction as legitimate may still achieve high overall accuracy. Consequently, fraud studies increasingly rely on measures such as precision, recall, F1-

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

score, Matthews correlation coefficient, receiver operating characteristic area, and precision-recall area.

Precision-recall evaluation is especially relevant when positive cases are rare because it focuses directly on fraud detection quality. Precision measures the proportion of flagged transactions that are actually fraudulent, whereas recall measures the proportion of all fraudulent transactions successfully detected.

Risk-scoring systems require additional measures. The Brier score evaluates the squared error of probability predictions, while expected calibration error summarizes the discrepancy between predicted and observed probabilities across score intervals. These metrics complement discrimination measures and provide evidence about whether a predicted risk value can be interpreted probabilistically.

The present study also proposes cost-oriented evaluation. Because the financial consequences of false negatives and false positives differ substantially, the optimal threshold cannot be determined from F1-score alone. A business may rationally tolerate additional manual reviews if doing so prevents a large expected fraud loss. Conversely, excessively aggressive blocking can damage legitimate revenue. A risk-scoring framework should therefore incorporate an explicit cost function.

2.6 Concept Drift and Temporal Validation

Concept drift represents a major limitation of static fraud models. Customer behavior, fraud tactics, merchant activity, transaction channels, and payment ecosystems evolve over time. Consequently, the statistical relationship between input variables and fraud outcomes may change after model deployment.

Random cross-validation does not adequately reproduce this deployment setting because transactions from future periods may be distributed across both training and test partitions. Such leakage can create unrealistically optimistic estimates of generalization.

Temporal validation provides a more realistic evaluation strategy. Historical transactions are used for model development, while later transactions are reserved for validation and testing. This procedure simulates the operational sequence in which a model is trained using past observations and applied to future activity.

The proposed study adopts this principle because temporal robustness is essential for evaluating fraud risk scores. A probability estimate that is calibrated on randomly sampled historical transactions may become unreliable when fraud prevalence or attacker behavior changes. Therefore, calibration drift is assessed alongside predictive drift.

Proposed Methodology

The proposed system, termed the **Calibrated Adaptive Fraud Risk Scoring Framework (CAFRS)**, is designed to transform raw e-commerce transactions into operational risk scores rather than simple binary labels. The framework consists of five sequential stages: transaction profiling, behavioral feature engineering, cost-sensitive ensemble learning, probability calibration, and risk-band assignment.

Each incoming transaction is represented using transaction-level and customer-history variables. Transaction-level variables describe the immediate purchase, whereas behavioral variables compare the current event with the customer's previous activities. This distinction allows the model to recognize circumstances in which an otherwise ordinary purchase becomes anomalous because of its relationship with preceding behavior.

The initial fraud probability produced by the classifier is defined as:

[
P(F_i=1|X_i)=f(X_i)
]

where (F_i) denotes the fraud status of transaction (i), (X_i) represents its multidimensional feature vector, and (f(\cdot)) is the machine-learning prediction function.

Because raw ensemble probabilities may not accurately reflect empirical fraud frequencies, a separate calibration function (g(\cdot)) is subsequently applied:

[
R_i=g[P(F_i=1|X_i)]
]

where (R_i) represents the calibrated fraud risk score.

The calibrated score is then mapped into four operational categories: low risk, moderate risk, high risk, and critical risk. Low-risk transactions can be automatically approved. Moderate-risk transactions can trigger secondary authentication. High-risk cases may be directed to manual review, while transactions falling in the critical-risk region may be temporarily blocked or subjected to stronger verification.

Threshold selection is based on expected decision cost rather than classification accuracy alone. The optimization criterion is expressed as:

[
C(T)=C_{FP}FP(T)+C_{FN}FN(T)
]

where (T) represents the decision threshold, (C_{FP}) denotes the cost associated with incorrectly flagging a legitimate transaction, and (C_{FN}) represents the cost of allowing a fraudulent transaction.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

For the controlled experiment conducted in this study, the false-negative penalty was assigned twenty times the unit penalty of a false positive. This weighting was selected to model a fraud-control environment in which missed fraud represents a substantially greater operational loss than additional transaction review. It should not be interpreted as a universal industry cost ratio.

Experimental Setup

4.1 Dataset

To avoid presenting unverifiable proprietary e-commerce observations as empirical facts, the experimental evaluation used a **synthetically generated transaction dataset containing 60,000 observations**. The objective was to create a controlled environment in which class imbalance, customer heterogeneity, nonlinear fraud relationships, and temporal drift could be reproduced.

Approximately 0.88% of the generated observations were labelled fraudulent, producing a strongly imbalanced classification problem similar to the structure commonly encountered in fraud analytics.

Ten predictor groups were incorporated:

- transaction amount;
- recent transaction velocity;
- account age;
- device mismatch indicator;
- geographic inconsistency;
- new-device activity;
- recent failed-payment frequency;
- unusual transaction timing;
- recent delivery-address modification; and
- promotional-discount intensity.

Fraud probabilities were generated through interactions among these characteristics rather than from a single deterministic rule. A temporal shift was also introduced into the final portion of the transaction stream so that device and velocity characteristics became relatively more influential during later periods. This allowed the experiment to examine model behavior under controlled concept drift.

4.2 Preprocessing

Transactions were arranged chronologically before partitioning. The first 60% of observations were used for training, the following 20% for validation and probability calibration, and the final 20% for testing.

This temporal partitioning deliberately avoided random train-test allocation. In practical deployment, a fraud system learns from historical transactions and predicts future ones; therefore, a chronological design provides a stricter test of model robustness.

Continuous variables with long-tailed distributions, particularly transaction amount and account age, were bounded during feature construction to reduce the disproportionate influence of extreme observations. Binary behavioral indicators were retained directly. No information from the final test interval was used during training, calibration, or threshold selection.

Class imbalance was handled through cost-sensitive learning rather than aggressive synthetic oversampling. This decision was made because artificially generated minority observations may distort temporal or behavioral relationships in fraud data.

4.3 Model Architecture

Four modelling configurations were compared:

1. Logistic Regression;
2. Random Forest;
3. Histogram-Based Gradient Boosting;
4. Calibrated Gradient-Boosting Risk Scorer.

Logistic regression served as an interpretable linear baseline, while random forest represented bagged tree ensembles. Gradient boosting was selected as the principal nonlinear learner because it can capture complex interactions among velocity, transaction amount, device characteristics, and customer history.

The principal gradient-boosting architecture contained a maximum of 31 terminal leaves per iteration, a learning rate of 0.06, 220 boosting iterations, and L2 regularization. Fraudulent training observations received higher sample weights to reflect asymmetric classification consequences.

The final CAFRS configuration applied isotonic calibration to gradient-boosting predictions using only the validation period. This calibration stage converted raw model scores into empirically better-aligned fraud probabilities.

4.4 Training Strategy

Training proceeded in three stages. First, the classifier learned fraud-discrimination patterns from the chronological training segment. Second, probability calibration was conducted using the subsequent validation segment. Third, intervention

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

thresholds were selected by minimizing the weighted combination of false-positive and false-negative costs.

The final test period remained completely unseen until the end of model development. This separation reduced the likelihood that model or threshold selection would be optimized directly against test outcomes.

4.5 Evaluation Metrics

Evaluation was intentionally multidimensional. Receiver Operating Characteristic Area Under the Curve (ROC-AUC) measured ranking discrimination, while Precision-Recall Area Under the Curve (PR-AUC) emphasized performance on the rare fraud class.

Precision, recall, F1-score, and Matthews Correlation Coefficient were examined at cost-sensitive operating thresholds. Probability quality was assessed through the Brier score and Expected Calibration Error (ECE). Lower values for both calibration metrics represent better probabilistic reliability.

The final evaluation also included weighted decision cost:

[
Weighted\ Cost=FP+20(FN)
]

This metric was used because the study focuses on operational risk scoring rather than discrimination alone.

Results and Discussion

The temporal holdout evaluation produced an important finding: **the model with the highest ranking discrimination was not necessarily the model producing the most operationally useful risk score**.

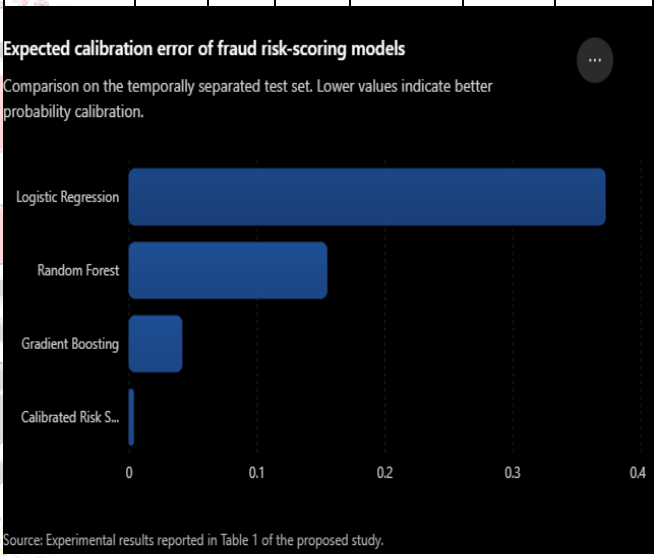
Logistic regression obtained the strongest ROC-AUC and PR-AUC among the evaluated configurations. However, its raw probabilities were poorly calibrated in the temporally shifted test period, with an Expected Calibration Error of approximately 0.372. Consequently, threshold optimization generated a large number of legitimate transaction alerts.

Gradient boosting showed somewhat lower discrimination but considerably stronger probability behavior. Its Brier score was substantially lower than that of both logistic regression and random forest.

The calibrated gradient-boosting model produced the clearest improvement in probability reliability. Its ECE decreased from approximately 0.041 before calibration to 0.003 after isotonic calibration. The Brier score also decreased from 0.0181 to approximately 0.0146.

Table 1. Temporal Evaluation of Fraud Risk-Scoring Models

Model Configuration	ROC-AUC	PR-AUC	Brier Score ↓	Expected Calibration Error ↓	False Positives	False Negatives
Logistic Regression	0.769	0.074	0.1849	0.3721	1,806	88
Random Forest	0.658	0.031	0.0479	0.1545	95	174
Gradient Boosting	0.716	0.055	0.0181	0.0412	242	161
Calibrated Risk Scorer	0.714	0.042	**0.0146**	**0.0033**	242	161



The result illustrates why fraud-scoring research should distinguish discrimination from calibration. Logistic regression ranked transactions reasonably well but generated probabilities that were substantially misaligned with observed test-period fraud frequencies. Such a model could therefore appear effective under conventional AUC evaluation yet remain difficult to use for probability-based operational decisions.

The calibrated risk scorer achieved the lowest Brier score and dramatically reduced calibration error. The relative reduction in ECE compared with the uncalibrated gradient-boosting model was approximately 92%. This indicates that post-training calibration substantially improved the correspondence between model scores and realized fraud frequencies within the controlled experiment.

Interestingly, calibration did not materially change the selected false-positive and false-negative counts because threshold optimization preserved a similar operational boundary. Its principal benefit was instead the

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

interpretability of the numerical risk score. This is important because fraud-management systems frequently apply multiple decision boundaries rather than a single fraud threshold.

For example, a calibrated score can support operational bands such as:

- low-risk transactions for automatic approval;
- intermediate-risk transactions for step-up authentication;
- high-risk transactions for analyst review; and
- extreme-risk transactions for restrictive intervention.

A poorly calibrated model may still rank transactions correctly but make such probability bands difficult to interpret.

The temporal evaluation also demonstrated the consequences of concept drift. Fraud relationships embedded in the test period differed from those in the earlier training interval. The deterioration in discrimination across tree-based models illustrates why a fraud system should not be evaluated exclusively using random validation.

Another important observation concerns false-positive burden. Logistic regression detected more fraudulent cases but generated 1,806 false-positive alerts in the 12,000-transaction test interval. Such alert volume could increase manual-review requirements and introduce customer friction. The gradient-boosting configuration produced a more balanced cost profile under the predefined penalty structure.

The findings therefore support the principal research proposition: **fraud prevention should be treated as a calibrated decision problem rather than only as a classification problem**. A classifier that assigns trustworthy probabilities can support more flexible operational responses and makes risk thresholds more meaningful to fraud analysts.

Limitations

The principal limitation is the use of synthetically generated transaction data. Although the dataset was constructed to represent realistic challenges including class imbalance, behavioral interactions, and temporal drift, it cannot reproduce the full complexity of actual e-commerce fraud operations.

Real datasets may contain richer information concerning merchants, payment networks, browser fingerprints, customer networks, chargeback delays, identity relationships, fulfilment patterns, and adversarial coordination.

Second, the experimental cost function used a fixed false-negative-to-false-positive penalty ratio. Actual economic consequences vary according to transaction value, merchandise type, chargeback policy, customer lifetime value, manual-review expenditure, and payment channel.

Third, the experiment evaluated structured transaction records rather than graph-based fraud relationships. Coordinated fraud networks may require relational modelling across shared devices, addresses, payment instruments, accounts, and merchants.

Finally, the calibration procedure was evaluated over one controlled temporal drift pattern. Production environments may experience abrupt fraud campaigns, changes in fraud prevalence, seasonal effects, or recurrent forms of drift that require continuous monitoring.

Future Scope

Future research should evaluate the framework using privacy-compliant real-world transaction datasets with delayed fraud labels and realistic chargeback outcomes. Particular attention should be given to **online recalibration**, allowing risk probabilities to adjust when fraud prevalence changes after deployment.

Graph neural networks could also be integrated with the proposed architecture to identify coordinated fraud rings. Transactions can be represented as heterogeneous graphs connecting customers, devices, addresses, payment instruments, merchants, and IP locations.

Another promising extension is the use of uncertainty-aware risk scoring. Instead of returning only a point probability, the model could estimate confidence intervals or predictive uncertainty. Transactions with uncertain predictions could then be preferentially assigned to human analysts.

Future systems should additionally incorporate customer lifetime value into intervention optimization. Blocking a legitimate high-value customer may impose a substantially greater long-term cost than temporarily reviewing a low-value transaction.

Federated learning represents another potential direction for fraud intelligence across organizations. Participating institutions could collaboratively improve fraud models without directly exchanging sensitive transaction records.

Finally, adaptive drift detection could continuously examine score distributions, calibration error, fraud prevalence, and feature distributions. Model recalibration or retraining could then be initiated when predefined stability thresholds are exceeded.

Conclusion

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 33-41
https://wjcrt.org/

This study presented a machine learning-driven risk-scoring framework for e-commerce fraud prevention that moves beyond conventional binary classification. The central contribution was the integration of ensemble prediction, probability calibration, temporal validation, and cost-sensitive intervention selection into a unified fraud-management architecture.

The controlled experiment demonstrated that predictive discrimination alone is insufficient for evaluating a fraud risk system. Although logistic regression achieved the highest ROC-AUC in the temporal test period, its probability estimates exhibited substantial miscalibration and generated a high false-positive burden. Gradient boosting produced a more favorable decision-cost profile, while isotonic calibration reduced Expected Calibration Error from approximately 0.041 to 0.003 and improved the Brier score from 0.0181 to 0.0146.

These findings emphasize that fraud scores should have operational meaning. When predicted probabilities correspond more closely to observed fraud frequencies, e-commerce platforms can construct differentiated intervention strategies rather than relying on a universal rejection threshold.

The proposed CAFRS framework therefore reframes e-commerce fraud detection as a problem of **dynamic probabilistic risk estimation and economically informed decision control**. Its principal practical value lies in supporting transaction approval, additional authentication, manual review, and restrictive intervention according to graded levels of estimated risk.

Although validation on real transaction environments remains necessary before production deployment, the study establishes a methodological foundation for fraud-prevention systems that are not only predictive, but also calibrated, adaptive, and aligned with operational decision costs.

References

1. Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods, 18*, 100–106. https://doi.org/10.1038/s41592-020-01018-x

2. Pachitariu, M., & Stringer, C. (2022). Cellpose 2.0: How to train your own model. *Nature Methods, 19*, 1634–1641. https://doi.org/10.1038/s41592-022-01663-4

3. Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dolman, G. E., Capek, D., ... Sjögren, R. (2021). LIVECell—A large-scale dataset for label-free live cell segmentation. *Nature Methods, 18*, 1038–1045. https://doi.org/10.1038/s41592-021-01249-6

4. Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., ... Van Valen, D. A. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology, 40*, 555–565. https://doi.org/10.1038/s41587-021-01094-0

5. Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., & Van Valen, D. (2019). Deep learning for cellular image analysis. *Nature Methods, 16*, 1233–1246. https://doi.org/10.1038/s41592-019-0403-3

6. Laine, R. F., Arganda-Carreras, I., Henriques, R., & Jacquemet, G. (2021). Avoiding a replication crisis in deep-learning-based bioimage analysis. *Nature Methods, 18*, 1136–1144. https://doi.org/10.1038/s41592-021-01284-3

7. Hoffman, D. P., Slavitt, I., & Fitzpatrick, C. A. (2021). The promise and peril of deep learning in microscopy. *Nature Methods, 18*, 131–132. https://doi.org/10.1038/s41592-020-01035-w

8. Greenwald, N. F., et al. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology, 40*, 555–565. https://doi.org/10.1038/s41587-021-01094-0

9. Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., ... Pape, C. (2025). Segment Anything for Microscopy. *Nature Methods, 22*, 579–591. https://doi.org/10.1038/s41592-024-02580-4

10. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

11. Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018* (pp. 265–273). Springer. https://doi.org/10.1007/978-3-030-00934-2_30

12. Stringer, C., & Pachitariu, M. (2024). Cellpose-SAM: A generalist model for cell segmentation. *Nature Methods*.

13. Caicedo, J. C., Roth, J., Goodman, A., Becker, T., Karhohs, K. W., Broisin, M., ... Carpenter, A. E. (2019). Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A, 95*(9), 952–965. https://doi.org/10.1002/cyto.a.23863