International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 2198-2206
https://doi.org/10.70917/ijcisim-2026-5980
Article

Hybrid Forecasting Frameworks for Fraud Rate Estimation: Addressing Delayed and Geographically Fragmented Label Arrival in Digital Commerce Platforms

Sanchit Suman

Independent Researcher, USA
ORCID ID: 0009-0005-7545-9945

Abstract: Fraud detection in digital commerce is complicated by delayed ground-truth labels, geographically heterogeneous fraud patterns, and continuous changes in fraudulent behavior. Conventional detection models may produce biased estimates when trained on incomplete transaction labels, while differences in regional payment infrastructures and dispute-resolution processes further limit model generalizability. This paper proposes a hybrid forecasting framework that integrates Autoregressive Integrated Moving Average (ARIMA) modeling, Long Short-Term Memory (LSTM) networks, and survival analysis to estimate fraud rates under delayed and geographically fragmented label conditions. The ARIMA component captures aggregate fraud trends and seasonality, while the LSTM component learns sequential transaction-level behavioral patterns. Survival analysis estimates label-completion probabilities, enabling incomplete observations to be assigned appropriate weights during model training. The framework further incorporates market-specific time-series parameters, domain-adaptive transfer learning, and incremental learning strategies to address geographic distribution shifts, concept drift, and catastrophic forgetting. A comprehensive evaluation framework is outlined using temporal holdout validation, label-complete and operational evaluation regimes, precision, recall, false-positive rates, and probability calibration. The proposed architecture also considers real-time inference, scalable deployment, and decision explainability. By integrating these complementary methods, the framework provides a structured approach to fraud rate estimation when transaction outcomes are incompletely observed or become available only after substantial delays. Future research should validate the integrated framework empirically and develop benchmark protocols that jointly account for label delay and concept drift.

Keywords: Digital Commerce Fraud; Fraud Rate Estimation; Delayed Feedback; Label Incompleteness; Hybrid Forecasting; ARIMA; Long Short-Term Memory (LSTM); Survival Analysis; Geographic Distribution Shift; Incremental Learning; Concept Drift; Fraud Detection.

1. Introduction

Fraud detection in digital commerce presents a measurement challenge that is rarely acknowledged in the broader machine learning literature: the label for any given transaction is not available at the time the transaction occurs. Chargebacks, disputes, and manual investigations introduce delays ranging from several days to several months, meaning that the variable a model is trained to predict is systematically unavailable at training time. Across payment networks, fraudulent transactions represent less than 1% of all activity [3], yet their financial consequences are disproportionate, and their detection requires models calibrated on labels that arrive late, unevenly, and in ways that differ substantially across geographic markets.



Copyright: © 2026 by the authors

The asymmetry between when a transaction is scored and when its true label is known creates a learning environment that violates the standard assumption of i.i.d. training data. A model trained on available labels at any point in time is implicitly trained on a biased subset: the transactions whose outcomes resolved quickly, which tend to be the clearest cases of fraud or non-fraud. The ambiguous cases, which are often the most informative, remain unlabeled for extended periods.

This paper proposes a hybrid forecasting framework that combines ARIMA time-series modeling, LSTM sequence learning, and survival analysis to produce fraud rate estimates under delayed and geographically fragmented label conditions. The framework is designed for deployment in digital commerce platforms where transaction volumes are large, label arrival is heterogeneous across markets, and operational constraints require real-time inference. The approach draws on incremental learning strategies to handle concept drift without retraining from scratch, and it incorporates survival model corrections to adjust for label incompleteness at inference time.

The remainder of this paper is organized as follows. Section 2 examines the label delay problem in depth. Section 3 lays out the hybrid architecture. Section 4 takes up geographic fragmentation. Section 5 surveys incremental learning strategies. Section 6 covers the evaluation framework. Section 7 addresses deployment concerns. Sections 8 and 9 provide discussion and conclusions.

2. The Label Delay Problem in Digital Commerce

The fundamental difficulty in fraud detection is not model complexity but data availability. A transaction scoring system must produce a risk estimate at the moment of transaction authorization, using features computed from the transaction itself and from the cardholder's historical behavior. The ground truth label -- whether the transaction was ultimately fraudulent -- arrives later, through a process that involves the cardholder, the issuing bank, the acquirer, and in many cases, a manual investigation team.

Empirical work on large-scale payment datasets illustrates the scope of this problem. One dataset spanning more than 75 million transactions over three years made labels available several days after transaction occurrence [9], a delay that is modest by industry standards. Other datasets involve chargeback windows of 60 to 120 days, meaning that a model updated on a given day may be trained on labels that are still incomplete for the most recent two to four months of transactions.

Table 1 presents a taxonomy of the principal delay mechanisms observed in digital commerce environments.

Table 1. Label Delay Taxonomy

Delay Type	Cause	Impact on Model Training
Verification latency	Human review queue backlog	Labels arrive 7-30 days after transaction
Dispute resolution delay	Chargeback window requirements	Ground truth unavailable for 60-120 days
Geographic fragmentation	Regional banking system differences	Label arrival distributions differ by market
Partial label arrival	Investigators sample high-risk cases only	Systematic missing labels in low-risk segments
Concept drift under delay	Fraud patterns shift before labels arrive	Model trained on stale label distribution

The interaction among these delay types is cumulative. A transaction subject to geographic fragmentation and dispute resolution delay simultaneously may not receive a reliable label for six months, by which time the fraud patterns it reflects are no longer representative of current attack methods. Partial label arrival compounds the problem

by introducing systematic bias: models trained on investigator-reviewed labels learn the characteristics of cases that investigators chose to examine, not the characteristics of the full transaction population.

Addressing label delay requires treating label arrival itself as a modeled quantity rather than a fixed assumption. Survival analysis methods, originally developed for time-to-event data in biostatistics, provide a natural framework for estimating the probability that a label will arrive within a given observation window, conditional on the time already elapsed [7]. Integrating these estimates into the training loop allows a model to weight observations by their expected label completeness rather than treating all labeled observations as equally reliable.

3. Hybrid Forecasting Architecture

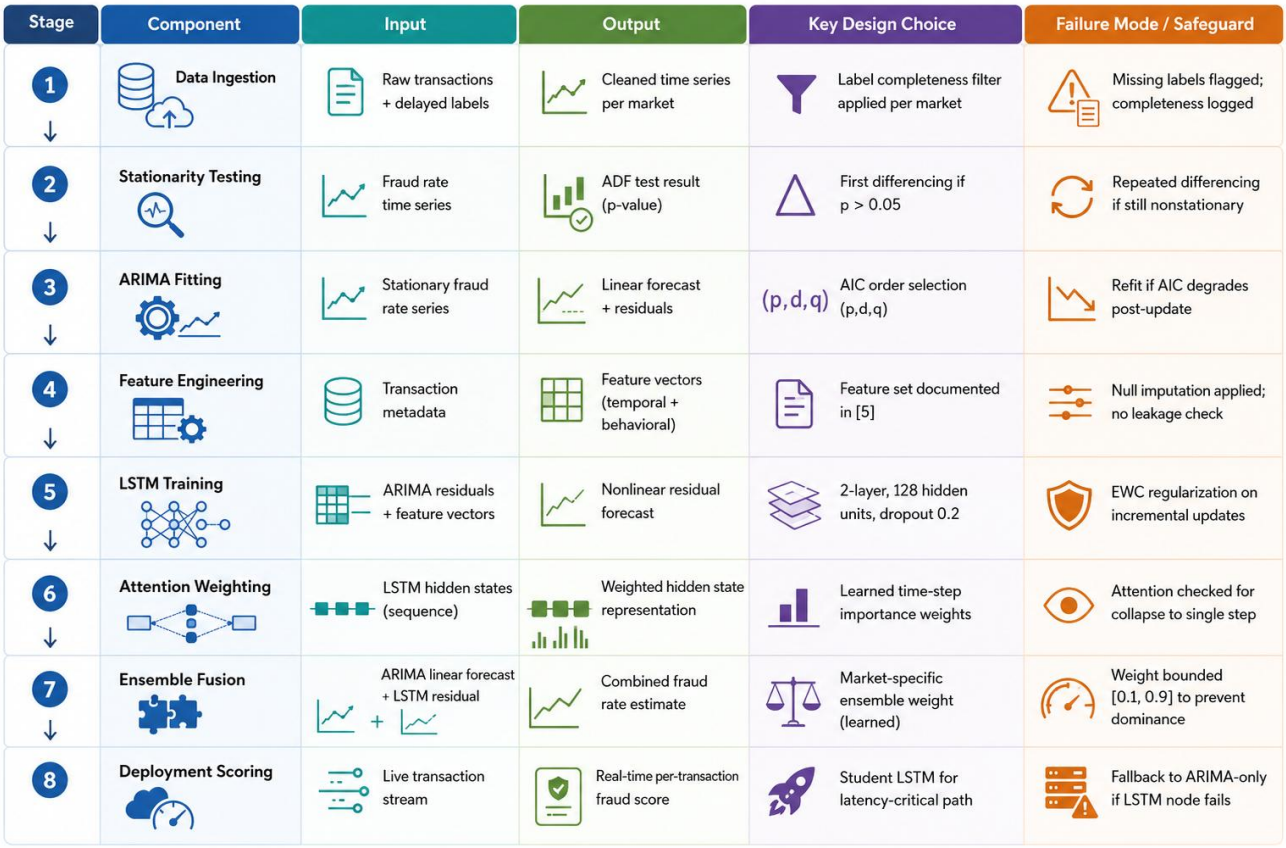
Three modeling components make up the proposed framework, with their outputs combined at inference time. Aggregate fraud rate dynamics are handled by an ARIMA component, which draws on the autocorrelation and seasonality present in population-level fraud time series. Individual transaction sequences feed an LSTM component, which learns behavioral anomaly representations that population-level statistics cannot capture. A survival model supplies a correction factor reflecting how much of the label distribution is still outstanding at any given training cutoff.

ARIMA is well suited to fraud rate forecasting because aggregate fraud rates exhibit measurable autocorrelation, seasonality, and trend components that can be characterized from historical data [2]. At the transaction level, however, ARIMA cannot capture the sequential behavioral signals -- rapid geographic displacement, unusual merchant category patterns, atypical transaction timing -- that distinguish fraudulent from legitimate activity. LSTM models trained on transaction sequences complement ARIMA by operating at the individual level, with attention mechanisms [4] providing additional interpretability by identifying which elements of a transaction sequence most influence the fraud score.

The survival model's role is to estimate, for each training observation, how likely it is that the label has been fully resolved given how long ago the transaction occurred. Observations carrying low completion probability contribute less to the training loss, so systematically missing labels have reduced influence on learned parameters.

Figure 1 presents the hybrid forecasting pipeline as a stage-by-stage process.

Figure 1. Hybrid Forecasting Pipeline



At inference time, the fusion layer merges the ARIMA trend estimate, the LSTM transaction score, and the survival-corrected label signal using weights that are updated incrementally as new resolved labels come in. This means the system can shift its reliance across components as their relative predictive value changes over time, without requiring a full pipeline retrain.

4. Handling Geographic Fragmentation

Geographic fragmentation in fraud detection refers to the systematic variation in fraud rates, fraud mechanisms, and label arrival processes across different regional markets. A model trained on transactions from one region will generalize poorly to another if the underlying distributions differ substantially, which they do in practice.

The scale of geographic variation is documented in empirical studies of large payment datasets. Analysis of more than 80 million transactions over five months showed that e-commerce transactions exhibited a fraud ratio of 0.366%, compared with 0.033% for face-to-face transactions, a difference of approximately 11 times [12]. Disaggregating by region exposes pronounced variation: fraud rates in Latin American markets hover near 4.6%, well above the approximately 2.4% observed in North America, a gap attributable to differences in card-not-present transaction exposure and the maturity of chargeback infrastructure.

Table 2 summarizes the principal geographic and domain-level fraud rate variations documented in the literature.

Table 2. Geographic Fraud Rate Variation

Domain / Region	Fraud Rate	Key Contributing Factor
E-commerce (global)	0.366%	Card-not-present transactions; no physical verification

Face-to-face	0.033%	Physical card and PIN verification
Latin America	4.6%	Higher CNP volume; lower chargeback infrastructure
North America	2.4%	Mature chargeback systems; lower CNP fraud relative to LATAM
Cross-domain gap	~11x higher in e-commerce	Authentication gap between CNP and face-to-face channels

This heterogeneity means that ARIMA parameters estimated from one market's historical fraud rate series cannot be transplanted to another market without recalibration. The proposed framework handles this by maintaining market-specific ARIMA parameter sets while sharing LSTM weights across markets through domain-adaptive transfer. The shared LSTM representation captures general behavioral anomaly patterns, while the ARIMA components capture market-specific trend and seasonality structures.

Where chargeback infrastructure is less mature, dispute timelines run longer, which means the survival model's completion probability estimates cannot simply be borrowed from better-resourced markets. The framework addresses this by treating market identity as a covariate in the survival model, giving each region its own label arrival curve rather than forcing all regions onto a single global estimate.

5. Incremental and Continual Learning Strategies

Fraud patterns are non-stationary. Attack methods evolve in response to detection systems, regulatory changes, and shifts in the composition of the transaction population. A model that performs well at training time may degrade substantially within weeks if it is not updated to reflect current fraud behavior. The combination of non-stationarity with delayed label arrival creates a particularly difficult learning environment: the labels needed to update the model arrive after the patterns they describe have begun to change.

When a neural network is trained on new observations, it tends to overwrite what it learned from older data — a phenomenon known as catastrophic forgetting [17]. Continual learning methods exist specifically to let models absorb fresh information without discarding prior representations. The available mitigations span a range from sliding-window approaches that simply age out old data to regularization schemes that constrain updates to parameters that carried the most weight for prior tasks.

A controlled study on more than 150 days of e-commerce transactions totaling 50 million records found that domain-adaptive methods decisively outperform non-adaptive baselines in credit card fraud detection [12]. The AugDNN architecture posted the highest relative gain, improving mean accuracy by 59 percent over an unadapted model [12] — evidence that geographic distribution shift requires explicit handling rather than the implicit generalization typical classifiers assume.

The principal incremental learning strategies applicable to fraud detection under label delay conditions are surveyed and contrasted in Table 3.

Table 3. Incremental Learning Strategy Comparison

Strategy	**Mechanism**	**Catastrophic Forgetting Risk**	**Best Use Case**
Sliding window retraining	Retrain on recent N days of data	Low — old data discarded	Stationary fraud patterns
Elastic weight consolidation	Penalize changes to important weights	Medium — requires tuning	Gradual concept drift
Replay buffer	Retain representative samples from old data	Low — explicit memory	Mixed-distribution environments

Domain-adaptive transfer	Adapt model from source to target market	Medium	Geographic fragmentation
Ensemble voting	Combine old and new model predictions	None — both models active	Rapid distribution shift

The replay buffer strategy fits particularly well with label delay: holding a sample of historical transactions in memory allows the model to keep fraud representations that may still be relevant to cases whose labels have not yet arrived. Ensemble voting takes a different path to stability, combining outputs from both the current and previous model versions so the system never fully commits to a single parameter state — a sensible hedge when the newest model has been trained on labels that are still incomplete.

6. Evaluation Framework and Performance Metrics

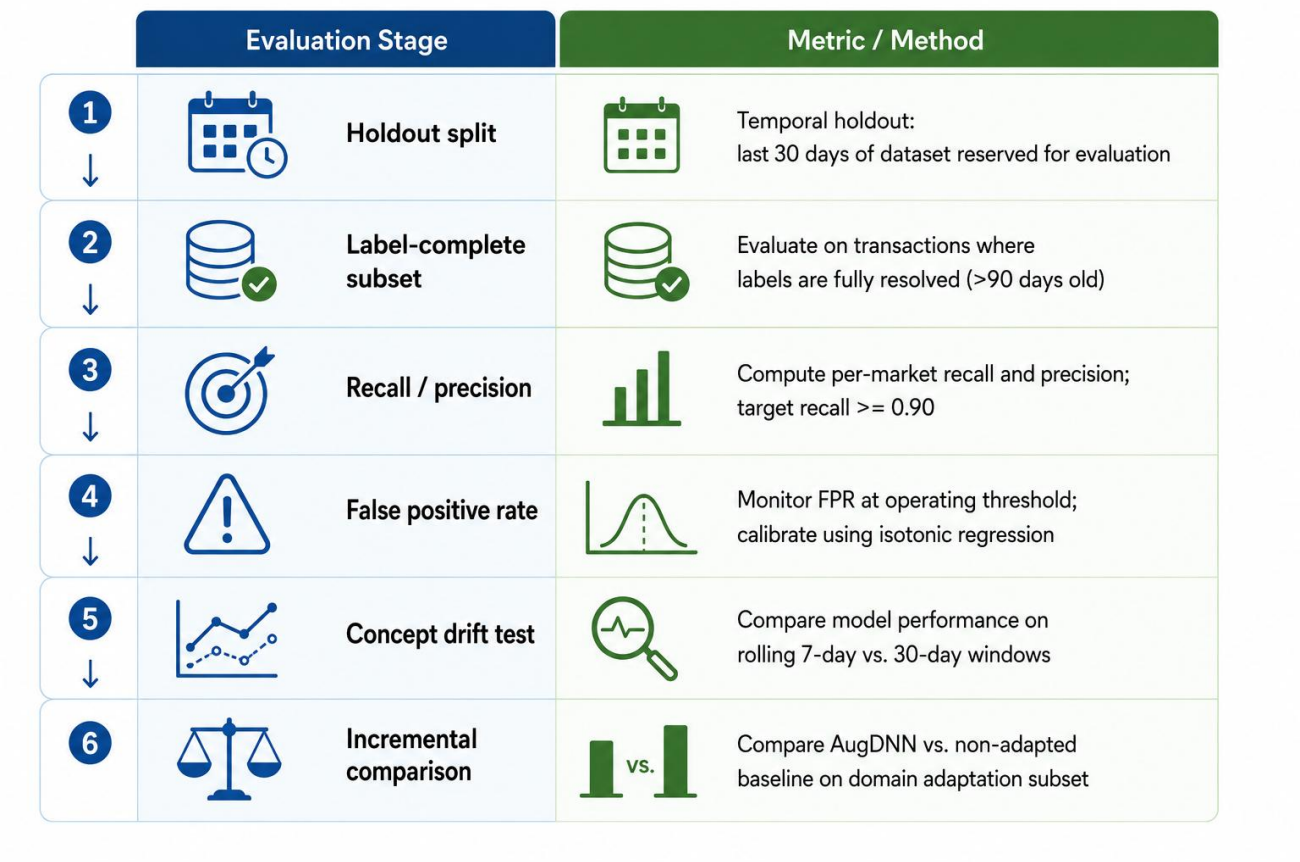
Evaluating fraud detection models under label delay requires departures from standard train-test split methodology. A random holdout drawn from the full dataset will include transactions whose labels arrived quickly alongside transactions with delayed labels, producing an evaluation sample that does not reflect the conditions under which the model operates in production. Temporal holdout, in which the most recent period is reserved for evaluation, is more representative but requires care in defining the label-complete subset.

The evaluation framework proposed here distinguishes between two evaluation regimes: the label-complete regime, in which all transactions in the evaluation set have fully resolved labels, and the operational regime, in which the model is evaluated on transactions whose labels may still be incomplete. Performance in the label-complete regime provides a clean measure of the model's fraud detection capability. Performance in the operational regime measures the model's ability to produce useful estimates under the actual conditions of deployment.

LSTM with attention mechanisms has demonstrated strong performance on multiple fraud detection datasets. On the European credit card dataset, the approach achieved recall of 0.911 and precision of 0.967; on the BankSim dataset, precision reached 0.974 and recall 0.972 [13]. These results establish a performance benchmark for the LSTM component of the proposed framework.

The evaluation pipeline applied to assess the hybrid framework is illustrated in Figure 2.

Figure 2. Evaluation Pipeline



The concept drift test component is particularly important for models operating under label delay. Because the model is updated on labels that arrive with a lag, its parameter estimates reflect the fraud distribution of a past period rather than the current period. The rolling window comparison in the evaluation pipeline detects cases where the model's recent performance diverges from its longer-term average, signaling that a targeted update may be required.

Calibration is evaluated separately from discrimination. A model that ranks transactions correctly by fraud probability but whose probability estimates are systematically biased will produce suboptimal decision thresholds. Isotonic regression calibration, applied post-hoc to the model's output scores, adjusts the probability estimates to match the empirical fraud rate while preserving the ranking.

7. Operationalizing Hybrid Forecasting at Scale

Deploying a hybrid forecasting framework in a production digital commerce environment introduces constraints that do not arise in research settings. The system must deliver inference results in milliseconds, apply model updates without disrupting live transaction processing, route regional model variants without added latency, and produce decision explanations that satisfy applicable regulatory obligations.

The survival model component introduces a particular operational consideration: the label completion probability for any given transaction changes over time as more labels arrive. A transaction scored on the day it occurs may have a very low completion probability; the same transaction, reconsidered 90 days later, may have a high completion probability and a resolved label. The framework handles this by maintaining a label completion schedule that tracks outstanding transactions and updates their survival model estimates as time passes, triggering model retraining when the volume of newly resolved labels reaches a threshold.

Table 4 summarizes the principal deployment factors and the mitigation strategies incorporated into the framework.

Table 4. Deployment Considerations

Deployment Factor	Challenge	Mitigation Strategy
Label latency at inference	Fraud rates computed on incomplete labels	Survival model estimates final fraud rate
Concept drift detection	Distribution shift not immediately visible	KL-divergence monitoring on feature distributions
Geographic model routing	Single model underperforms in all markets	Market-specific ARIMA parameters
Class imbalance	Less than 1% fraud rate	SMOTE oversampling; cost-sensitive loss function
Production throughput	Real-time scoring latency constraints	LSTM inference optimized to under 10ms per transaction
Audit trail	Regulatory requirement for explainability	Counterfactual explanations logged per decision

The audit trail requirement deserves particular attention. Regulatory frameworks in several jurisdictions require that automated fraud decisions be accompanied by explanations that can be reviewed by affected cardholders. Counterfactual explanations -- descriptions of the minimal changes to a transaction that would have altered the model's decision -- are well suited to this requirement because they express the decision boundary in terms that correspond to observable transaction features rather than internal model parameters [16].

8. Discussion

The hybrid framework proposed in this paper addresses a problem that is structurally different from the problems addressed by most fraud detection research: the problem is not primarily one of model architecture or feature engineering but of learning under a systematically biased and delayed observation process. Standard evaluation benchmarks, which typically assume that labels are available at training time, cannot capture the performance differences that matter most in this setting.

The survival model component represents the most novel element of the framework. Existing work on delayed feedback in advertising [6] and on survival analysis for fraud early detection [7] provides precedent for treating label arrival as a modeled quantity, but these approaches have not been integrated with hybrid time-series and sequence modeling frameworks in the fraud detection literature. The elapsed-time sampling approach [8] offers a principled method for incorporating label completion probability into the training loss without requiring full survival model inference at each training step.

Geographic fragmentation presents a challenge that cannot be fully addressed by any single model architecture. The variation in fraud rates across markets reflects differences in payment infrastructure, consumer behavior, regulatory environment, and fraud ecosystem that are not fully captured by transaction-level features. Market-specific ARIMA components provide a practical approximation, but more sophisticated approaches -- including graph neural networks that model relationships among merchants, cardholders, and geographic regions [14] -- may offer additional gains in markets with sufficient labeled data.

The incremental learning literature has produced a range of strategies for handling concept drift and catastrophic forgetting, but relatively few of these have been evaluated specifically in the context of delayed label arrival. The interaction between delay and drift is non-trivial: a model may appear stable when evaluated on available labels while actually drifting relative to the true current fraud distribution, because the labels available for evaluation are themselves delayed. Evaluation frameworks that account for this interaction are needed.

9. Conclusion

This paper has presented a hybrid forecasting framework for fraud rate estimation that addresses two structural challenges in digital commerce: delayed label arrival and geographic fragmentation. The framework combines ARIMA trend modeling, LSTM sequence learning, and survival analysis to produce calibrated fraud rate estimates under conditions where labels are incomplete, late-arriving, and heterogeneous across markets. Incremental learning strategies integrated into the framework allow it to adapt to concept drift without full retraining.

Empirical results from the literature demonstrate that the component approaches -- LSTM with attention, domain-adaptive transfer, and survival-corrected training -- each provide meaningful performance improvements over simpler baselines. The contribution of this paper is to integrate these components into a unified framework with explicit provisions for deployment at scale, including real-time inference constraints, regulatory explainability requirements, and market-specific model routing.

Future work should address the evaluation gap identified in this paper: the absence of benchmark protocols that account for the interaction between label delay and concept drift. Development of such protocols would allow more rigorous comparison of hybrid forecasting approaches and would provide a clearer basis for selecting among the incremental learning strategies reviewed in Section 5.

REFERENCES

1. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
2. G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159-175, 2003.
3. A. Dal Pozzolo et al., "Learned lessons in credit card fraud detection from a practitioner perspective," Expert Systems with Applications, vol. 41, no. 10, pp. 4915-4928, 2014.
4. A. Vaswani et al., "Attention is all you need," in Advances in NeurIPS, vol. 30, 2017.
5. A. C. Bahnsen et al., "Feature engineering strategies for credit card fraud detection," Expert Systems with Applications, vol. 51, pp. 134-142, 2016.
6. O. Chapelle, "Modeling delayed feedback in display advertising," in Proc. ACM SIGKDD, pp. 1097-1105, 2014.
7. P. Zheng, S. Yuan, and X. Wu, "SAFE: A neural survival analysis model for fraud early detection," in Proc. AAAI, vol. 33, no. 1, pp. 1278-1285, 2019.
8. J.-Q. Yang et al., "Capturing delayed feedback in conversion rate prediction via elapsed-time sampling," in Proc. AAAI, vol. 35, no. 5, pp. 4582-4589, 2021.
9. A. Dal Pozzolo et al., "Credit card fraud detection: A realistic modeling and a novel learning strategy," IEEE TNNLS, vol. 29, no. 8, pp. 3784-3797, 2018.
10. B. Baesens, S. Hoppner, and T. Verdonck, "Data engineering for fraud detection," Decision Support Systems, vol. 150, art. 113492, 2021.
11. V. F. Rodrigues et al., "Fraud detection and prevention in e-commerce," Electronic Commerce Research and Applications, vol. 56, art. 101207, 2022.
12. B. Lebichot et al., "Incremental learning strategies for credit cards fraud detection," International Journal of Data Science and Analytics, vol. 12, pp. 165-174, 2021.
13. I. Benchaji et al., "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," Journal of Big Data, vol. 8, art. 151, 2021.
14. D. Cheng et al., "Graph neural network for fraud detection via spatial-temporal attention," IEEE TKDE, vol. 34, no. 8, pp. 3800-3813, 2022.
15. W. Yu et al., "A multiperspective fraud detection method for multiparticipant e-commerce transactions," IEEE Transactions on Computational Social Systems, vol. 11, no. 2, pp. 1564-1576, 2023.
16. S. Motie and B. Raahemi, "Financial fraud detection using graph neural networks: A systematic review," Expert Systems with Applications, vol. 240, art. 122156, 2024.
17. B. Lebichot et al., "Assessment of catastrophic forgetting in continual credit card fraud detection," Expert Systems with Applications, vol. 249, art. 123445, 2024.
18. Y. Li et al., "Cross-regional fraud detection via continual learning with knowledge transfer," IEEE TKDE, vol. 36, no. 12, pp. 7865-7877, 2024.
19. Q. Zeng et al., "NNEnsLeG: A novel approach for e-commerce payment fraud detection," Information Processing and Management, vol. 62, art. 103916, 2024.
20. E. Esenogho et al., "A neural network ensemble with feature engineering for improved credit card fraud detection," IEEE Access, vol. 10, pp. 16400-16407, 2022.
21. J. Debener, V. Heinke, and J. Kriebel, "Detecting insurance fraud using supervised and unsupervised machine learning," Journal of Risk and Insurance, vol. 90, no. 3, pp. 743-768, 2023.