Journal of Machine Learning and Deep Learning (JMLDL)
Volume 3, Issue 1, Date: 30- June – 2026
ISSN: 2978-4336, DOI: 10.64820/AEPJMLDL.31.47.55.62026

LSTM Autoencoder vs. Isolation Forest vs. One-Class SVM for Financial Transaction Fraud Detection

Nguyen Trung Hieu, MBA Candidate
Executive Business Centre, University of Greenwich, London, England, United Kingdom · Alliance with FPT University, Ho Chi Minh City, Viet Nam
Email: nh7471m@gre.ac.uk
*Corresponding Author

***Abstract*—Financial fraud detection under extreme class imbalance (fraud rate < 1%) motivates unsupervised or semi-supervised anomaly detection approaches. This study compares three paradigms—LSTM Autoencoder (LSTM-AE), Isolation Forest (IF), and One-Class SVM (OCSVM)—on 100,000 synthetic transactions with a 0.35% fraud rate and 30 features (28 PCA-transformed components plus scaled Amount and Time). Using a stratified 60/20/20 train-validation-test split, evaluation is primarily via AUC-PR, which is more informative than AUC-ROC under extreme imbalance. IF achieves the highest AUC-PR (0.4169), vastly outperforming both OCSVM (0.0616) and LSTM-AE (0.0171). AUC-ROC rankings are IF (0.9784) > OCSVM (0.9100) > LSTM-AE (0.8148). McNemar tests confirm all pairwise differences are statistically significant ($p < 0.001$). Both LSTM-AE and OCSVM exhibit near-random precision on the minority class, suggesting that only IF effectively exploits the isolation structure of PCA-transformed fraud patterns. The LSTM-AE's failure is attributed to PCA destroying sequential dependencies that recurrent architectures require.**

Keywords—fraud detection; anomaly detection; autoencoder; isolation forest; one-class support vector machine; class imbalance; precision-recall

I. INTRODUCTION

Financial fraud causes billions of dollars in losses each year. When fraud rates fall below one percent, supervised classification struggles with extreme class imbalance [1]. Anomaly detection methods offer an alternative by learning the patterns of legitimate behavior and flagging deviations [2]. Deep learning autoencoders have attracted growing interest for this task [3]. Long short-term memory networks [4], and long short-term memory autoencoders in particular, can model sequential dependencies in transaction streams [5]. When features are transformed by principal component analysis, which is common in privacy-sensitive financial data, the sequential structure of the original features is removed, which may undermine the advantage of a recurrent model.

This study compares three anomaly detection paradigms on synthetic credit card data with features transformed by principal component analysis. The first is a long short-term memory autoencoder, a deep learning model based on reconstruction. The second is an isolation forest, a tree-based model based on isolation [6]. The third is a one-class support vector machine, a kernel-based model based on a decision boundary [7]. The primary evaluation metric is the area under the precision-recall curve, which is more informative than the area under the receiver operating characteristic curve under extreme imbalance [8].

The operational setting makes the problem harder than a balanced benchmark suggests. A production fraud filter scores a continuous stream of transactions, must keep the false alarm rate low enough that human review stays affordable, and must adapt as fraud tactics change over time. Labels are scarce, delayed, and noisy, since a transaction is confirmed as fraud only after a dispute. These constraints favor methods that learn the shape of legitimate behavior without heavy supervision, and raise the value of an honest, imbalance-aware evaluation that does not overstate performance.

The cost structure of fraud detection is sharply asymmetric and shapes the choice of metric. A missed fraud transfers a direct loss, while a false alert consumes scarce human-review capacity and, in aggregate, produces alert fatigue that erodes trust in the system [1]. Because legitimate transactions outnumber fraud by more than two hundred to one, even a small false-positive rate floods reviewers, so a detector is useful only when it concentrates true fraud among its top-ranked alerts. This is precisely what the area under the precision-recall curve captures, and it is why a detector with a high area under the receiver operating characteristic curve can still be impractical in operation.

Three research questions are addressed. The first asks whether the autoencoder outperforms traditional anomaly detectors on features transformed by principal component analysis. The second asks which method best handles the precision-recall trade-off at a fraud rate of 0.35 percent. The third asks how the principal component transformation affects the comparison between models.

Received: 3-3- 2026
Revised: 27-6-2026
Published: 30-6-2026

The gap addressed here is the mismatch between model assumptions and feature structure. Many strong fraud detectors assume informative relational or sequential structure, yet production data are often released only after a privacy-preserving transformation that removes such structure. A controlled comparison on transformed features, with a primary metric suited to extreme imbalance, isolates how each detector responds when its structural assumptions no longer hold [8].

II. Related Work

A. Anomaly Detection and Fraud

Comprehensive surveys map the anomaly-detection techniques used for financial fraud and the recurring challenge of extreme class imbalance [1], [2]. Reviews and comparative studies of machine learning and deep learning for fraud detection catalogue dozens of methods and report that no single approach dominates across data sets and metrics [9]-[12]. Recent reviews of unsupervised anomaly detection reach the same conclusion and stress the difficulty of evaluation under imbalance [13], [14]. A persistent theme is that imbalance handling, through resampling or cost-sensitive training, often matters as much as the choice of classifier [15], [16].

B. Deep Learning for Fraud Detection

Reconstruction-based and sequential deep models have been widely applied to card fraud. Autoencoder and deep-classifier hybrids [3], [17], autoencoder and recurrent combinations [18], and sequence models based on long short-term memory and attention [5], [19]-[21] all report strong results on public card-fraud benchmarks. Neural-network ensembles with feature engineering or data resampling have also been proposed [22]-[25], alongside simpler neural baselines [26] and hybrid unsupervised and supervised pipelines [27]. Variational autoencoders and attention-based autoencoders extend this family to the fraud setting [28], [29]. This study contributes controlled evidence that, once features are transformed by principal component analysis, a tree-based isolation method can outperform such deep architectures.

C. Isolation and One-Class Methods

Isolation forests detect anomalies through the short path lengths needed to isolate rare points [6], and recent extensions improve the split geometry, add a deep variant, and add interpretability [30]-[32]. One-class support vector machines learn a boundary around the legitimate class [7], and comparative studies place one-class learners against binary classifiers on card fraud data [33], [34]. Beyond card fraud, one-class support vector machines and one-class autoencoders are widely used for intrusion and network anomaly detection, including lightweight and real-time designs [35]-[37], and one such study pairs principal-component analysis with a one-class detector, the same feature regime studied here [38].

D. Imbalance, Generative, and Graph Methods

Because fraud is the minority class, generative oversampling and graph-based modeling are active directions. Generative adversarial networks and conditional tabular generators synthesize minority samples [39]-[41], temporal transformers with conditional tabular generation and hybrid generative pipelines push this further [42], [43], graph neural networks model relations between accounts and transactions, including cost-sensitive and context-encoding variants and hybrid graph-attention autoencoders [44]-[47], and dedicated studies target the highly imbalanced card-fraud case directly, including improved gradient boosting and density-based clustering [48]-[50].

E. Graph and Transformer Architectures

Two model families have recently gained prominence for fraud detection. Graph neural networks represent accounts and transactions as nodes and edges, and surveys document rapid progress and open challenges in this direction [51]-[53]. Concrete designs add federated training across institutions and adaptive sampling that counters class imbalance [54], [55]. Transformer and attention architectures form the second family, with self-attention generative models, temporal transformers, and dedicated anomaly transformers reporting strong results on transaction streams [56]-[58]. Both families rely on informative relational or sequential structure, which the principal-component transformation used in the present data removes, so neither is expected to hold its advantage on these features.

Comparative and ensemble studies continue to refine card-fraud detection. Side-by-side evaluations of machine-learning techniques report data-dependent rankings rather than a single winner [59], [60], while variational-autoencoder and generative-adversarial hybrids address imbalance at the data level by reconstructing or synthesizing minority samples [61], [62].

F. Sequence and Hybrid Models

Long short-term memory networks remain a standard tool for sequence modeling [4], and recent hybrids combine the recurrent layer with language-model encoders, metaheuristic optimization, attention, and gradient-boosting back-ends for card fraud [63]-[66]. Density-based isolation hybrids extend the tree-based family with neighborhood information [67], and broad comparisons of resampling and learning methods on imbalanced card data place these approaches side by side [68]. Surveys of deep learning for time-series anomaly detection map the wider field [69].

G. Evaluation and Deployment Concerns

Evaluation under imbalance is itself a research topic, since the precision-recall plot is more informative than the receiver operating characteristic plot on imbalanced data [8], and pairwise classifier differences can be tested with the McNemar test [70]. Concept drift and online operation are important for deployment [71], [72], and privacy-preserving, real-time pipelines address the operational constraints of banking networks [73].

III. Method

A. Preliminaries

The comparison covers three anomaly-detection paradigms that differ in how each models normal behavior. The long short-term memory autoencoder is a reconstruction-based method that learns to rebuild legitimate transactions

and flags those it reconstructs poorly [4]. The isolation forest is a tree-based method that isolates points with short random partition paths, so rare points score as anomalies [6]. The one-class support vector machine is a boundary-based method that fits a frontier around the legitimate class in a kernel space [7]. The study asks which assumption, namely reconstruction, isolation, or boundary, best matches transaction features after a principal-component transformation, under a fraud rate below one percent.

B. Data

The study uses 100,000 synthetic financial transactions modeled after the structure of a public credit card fraud data set [10]. The data set includes 28 features transformed by principal component analysis, labeled V1 through V28, plus an amount and a time, both standardized, which yields 30 input features. The fraud rate is 0.35 percent, namely 350 fraudulent transactions among 99,650 legitimate ones, which gives an imbalance ratio of about one to 284. Table I summarizes the data set.

TABLE I. DATA SET CHARACTERISTICS

Property	Value
Total transactions	100,000
Fraudulent	350 (0.35%)
Legitimate	99,650 (99.65%)
Features	28 components + scaled amount + scaled time
Imbalance ratio	About 1 to 284
Split	60% train / 20% validation / 20% test (stratified)

The data are split by stratified sampling into 60 percent for training, namely 60,000 transactions, 20 percent for validation, namely 20,000, and 20 percent for testing, namely 20,000, which preserves the fraud ratio in each subset. The test set holds about 70 fraudulent transactions.

The synthetic data reproduce the salient statistical structure of the real data, namely features anonymized by principal component analysis, a sub-percent fraud prevalence, and a heavy-tailed transaction amount, while keeping the size tractable for autoencoder training. Timestamps are drawn uniformly over a two-day window. For legitimate transactions, the 28 components are sampled independently from a standard normal distribution, and the amount follows a log-normal distribution clipped to a realistic range. Fraudulent transactions share the same base distribution but with twelve components mean-shifted to create a separable but subtle signal that mirrors the directions reported for the real data, with the strongest shifts applied to components V14, V1, and V3, and fraudulent amounts follow a heavier log-normal distribution. Finally, the amount and the time are standardized to zero mean and unit variance, which yields the 30 model inputs. This controlled construction lets the comparison isolate how each detector responds to a known fraud signal under extreme imbalance.

C. Models

The long short-term memory autoencoder uses an encoder of two recurrent layers with 64 and 32 units and rectified linear activation, and a symmetric decoder that ends in a time-distributed dense layer of 30 units. Each transaction is reshaped into a single-step sequence. Training uses only legitimate transactions from the training set, which makes the method semi-supervised, with the Adam optimizer at a learning rate of 0.001, the mean squared error loss, a batch size of 256, and up to 50 epochs with early stopping at a patience of five. The anomaly score is the per-sample reconstruction error, given in Equation (1):

$$e_i = \frac{1}{d}\sum_{j=1}^{d}(x_{ij} - \hat{x}_{ij})^2 \tag{1}$$

where d, which equals 30, is the feature dimension. The detection threshold is set at the 95th percentile of the reconstruction errors on the normal training data.

The isolation forest uses 200 estimators with a contamination parameter of 0.004 and the default maximum number of samples. Anomaly scores are the negated decision function values, where a higher value means more anomalous, as given in Equation (2):

$$s(\mathbf{x}) = 2^{-E[h(\mathbf{x})]/c(n)} \tag{2}$$

where h of x is the path length and c of n is the average length of an unsuccessful search. The one-class support vector machine uses a radial basis function kernel with a scaled gamma and a nu of 0.004. Because of computational limits, it is trained on a random subset of 10,000 training samples. Anomaly scores are the negated decision function values.

D. Evaluation Metrics

Evaluation under extreme imbalance must focus on the minority class. Let the counts of true positives, false positives, and false negatives be written as TP, FP, and FN. Precision and recall are defined in Equations (3) and (4), and the F1-score is the harmonic mean of precision and recall in Equation (5).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The primary metric is the area under the precision-recall curve, summarized by the average precision in Equation (6), where the precision and recall at the n-th threshold are written as the n-indexed precision and recall.

$$\text{AP} = \sum_{n}(R_n - R_{n-1})P_n \tag{6}$$

The average precision is far more informative than the area under the receiver operating characteristic curve under

extreme imbalance, because it ignores the large number of true negatives and rewards precision among the few flagged transactions [8]. The area under the receiver operating characteristic curve is still reported for comparison with prior work.

E. Evaluation Protocol

Models are evaluated on the held-out test set of 20,000 transactions, which holds about 70 fraudulent cases. The metrics are the area under the receiver operating characteristic curve and the area under the precision-recall curve, with the latter serving as the primary metric because of the extreme imbalance. Threshold-dependent F1, precision, and recall are also reported, using the native threshold of each model. Statistical comparisons use the McNemar test for pairwise classification differences [70] and bootstrap 95 percent confidence intervals for the area under the precision-recall curve, using 2,000 iterations. The full procedure is summarized in Algorithm 1.

Algorithm 1 Unsupervised Fraud Detection and Comparison Pipeline

Require: transactions split 60/20/20 (stratified); detectors {autoencoder, isolation forest, one-class SVM}
Ensure: per-model test metrics and significance tests
1: Fit the feature scaling on the training set only
2: Train the autoencoder on legitimate training transactions; set the threshold at the 95th percentile of training error
3: Fit the isolation forest and the one-class SVM on the training set
4: for each detector m do
Score every test transaction with m (higher score = more anomalous)
Compute AUC-ROC, AUC-PR, F1, precision, and recall
5: end for
6: Compare detectors pairwise with the McNemar test
7: Estimate 95% confidence intervals for AUC-PR by bootstrap (B = 2000)
8: return test metrics and test statistics

F. Reproducibility and Validation Design

The comparison is designed so that the only variable is the detector. The same 60/20/20 stratified split, the same scaling fitted on the training set only, and the same test set of about 70 fraudulent transactions are used for all three models. The autoencoder is trained only on legitimate transactions and its threshold is fixed at the 95th percentile of training reconstruction error, while the isolation forest and the one-class machine use fixed contamination and nu values of 0.004 rather than test-tuned thresholds. The primary metric, the area under the precision-recall curve, is threshold-free, so it does not depend on these native thresholds, whereas the reported F1, precision, and recall do. Significance is assessed with the McNemar test on paired test predictions and with bootstrap intervals over 2,000 iterations for the area under the precision-recall curve. Fixing the split, the scaling, and the thresholds in advance lets any performance gap be attributed to the detector rather than to tuning.

IV. Results and Discussion

A. Detection Performance

Table II presents the full test-set metrics. The isolation forest reaches the highest score on every metric. The gap on the area under the precision-recall curve is particularly revealing. The isolation forest reaches 0.4169, while both the one-class support vector machine, at 0.0616, and the autoencoder, at 0.0171, fall close to the baseline fraud rate, which indicates near-random precision on the minority class.

Table II. Anomaly Detection Performance (Test Set, 20,000 Transactions)

Metric	Autoencoder	Isolation forest	One-class SVM
AUC-ROC	0.8148	**0.9784**	0.9100
AUC-PR	0.0171	**0.4169**	0.0616
F1-score	0.0382	**0.3974**	0.1042
Precision	0.0203	**0.3605**	0.0606
Recall	0.3286	**0.4429**	0.3714

The results on the area under the receiver operating characteristic curve give a misleadingly optimistic picture for the one-class support vector machine and the autoencoder. Both reach a value above 0.81, which suggests reasonable discrimination. The area under the precision-recall curve, which focuses on the minority class, reveals that only the isolation forest achieves meaningful precision among flagged transactions. The value of 0.0171 for the autoencoder is only slightly above the random baseline of 0.0035, namely the fraud prevalence, while the value of 0.0616 for the one-class support vector machine, though better, remains impractical. Figure 1 shows that the gap on the precision-recall metric is far more dramatic than the gap on the receiver operating characteristic metric.

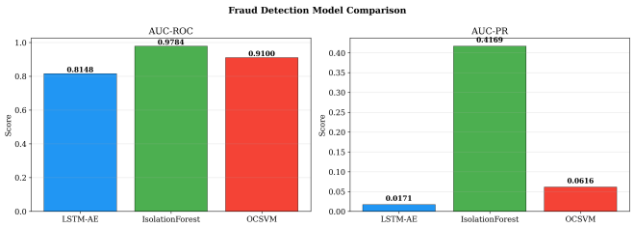

Fig. 1. Comparison on the area under the receiver operating characteristic curve and the area under the precision-recall curve. The isolation forest dominates on both metrics, and the gap on the precision-recall metric is especially large.

Figure 2 presents the receiver operating characteristic curves. The isolation forest and the one-class support vector machine both reach a high value, above 0.91, while the autoencoder lags at 0.815.

The threshold-dependent metrics confirm the advantage of the isolation forest, with an F1-score of 0.3974, a precision of 0.3605, and a recall of 0.4429. The one-class support vector machine reaches a higher recall, at 0.3714, than the autoencoder, at 0.3286, but at very low precision, at 0.0606, which generates many false positives. Figure 3 shows the precision-recall curves.

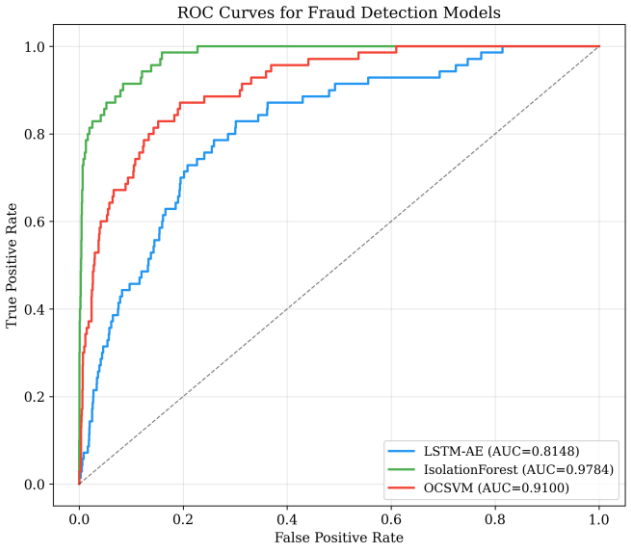

Fig. 2. Receiver operating characteristic curves for the three models. The isolation forest and the one-class support vector machine reach high values, while the autoencoder lags.

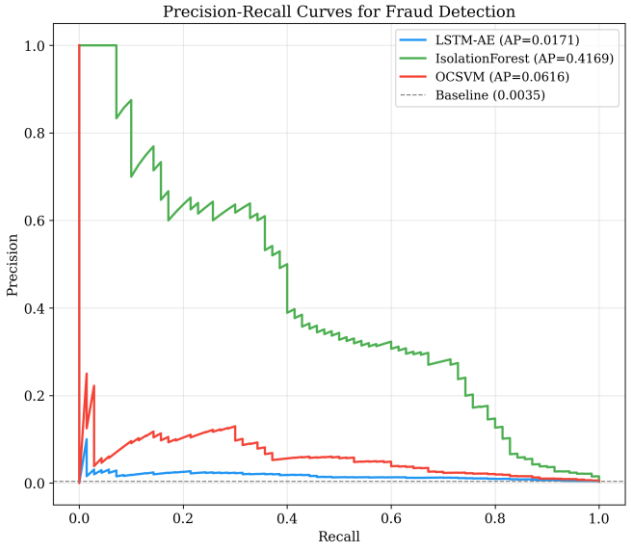

Fig. 3. Precision-recall curves. The isolation forest keeps non-trivial precision at moderate recall, while the one-class support vector machine and the autoencoder collapse to near-zero precision quickly.

B. *Statistical Tests*

Table III reports the McNemar test results. All pairwise tests reject the null hypothesis of equal classification performance, with every p-value below 0.001, which confirms that the ranking, with the isolation forest first, the one-class support vector machine second, and the autoencoder third, is statistically significant.

TABLE III. MCNEMAR TEST RESULTS (PAIRWISE COMPARISONS)

Comparison	McNemar chi-square	p-value
Isolation forest vs. autoencoder	936.51	<0.001
Isolation forest vs. one-class SVM	324.10	<0.001
One-class SVM vs. autoencoder	371.97	<0.001

Bootstrap 95 percent confidence intervals for the area under the precision-recall curve are from 0.2976 to 0.5420 for the isolation forest, from 0.0386 to 0.1086 for the one-class support vector machine, and from 0.0110 to 0.0334 for the autoencoder. All intervals are non-overlapping, which provides further evidence that the differences are robust to sampling variability. Figure 4 shows the reconstruction error distribution of the autoencoder.

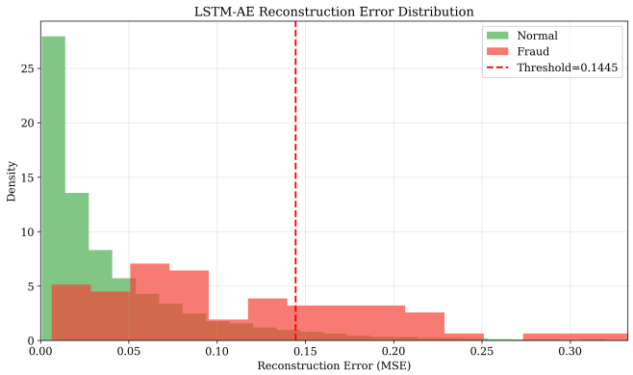

Fig. 4. Reconstruction error distribution of the autoencoder. The normal and fraud distributions overlap substantially, which explains the inability of the model to separate fraud reliably.

C. *Operating-Point Analysis*

Read at each model native threshold, the metrics in Table II expose where each detector places its boundary. The isolation forest, at a contamination of 0.004, reaches precision 0.3605 and recall 0.4429, so a large minority of flagged transactions are genuine fraud. The one-class support vector machine flags a similar fraction of fraud, with recall 0.3714, but at precision 0.0606, so most of its alerts are false. The autoencoder, at the 95th-percentile threshold, reaches recall 0.3286 at precision 0.0203. Because the area under the precision-recall curve is threshold-free, the dominance of the isolation forest, 0.4169 against 0.0616 and 0.0171, holds across all operating points rather than only at these thresholds. The practical consequence is alert volume: at equal recall the isolation forest raises far fewer false alerts for human review, which is the cost that decides whether a fraud filter is usable.

D. *Score Distribution and Error Analysis*

The error structure follows from the score distributions. Figure 4 shows that the reconstruction-error distributions of legitimate and fraudulent transactions overlap heavily, so no threshold on that score separates the classes well, which is why the autoencoder precision sits near the random baseline of 0.0035 set by the fraud prevalence. The isolation forest assigns systematically shorter path lengths to fraudulent points in the 30-dimensional transformed space, so its score ranks fraud above legitimate traffic and yields the only practical precision among the three. The one-class boundary captures part of this structure, which lifts its area under the receiver operating characteristic curve to 0.9100, but it admits many legitimate points near the margin, which keeps precision low. Recall is similar across the three methods, between 0.33 and 0.44, so the methods differ mainly in how many false alerts each raises to reach that recall.

E. Why Only the Isolation Forest Succeeds

The most striking finding is that only the isolation forest achieves practical fraud detection performance. Both the autoencoder and the one-class support vector machine produce near-random values on the area under the precision-recall curve, at 0.0171 and 0.0616. The path-based scoring of the isolation forest naturally exploits the sparse-region occupancy of fraudulent transactions in the 30-dimensional space transformed by principal component analysis, where anomalous points need shorter isolation paths [6], [31].

F. The Weak Result of the Autoencoder on Transformed Features

The weak result of the autoencoder, with a value of 0.0171, has a clear structural explanation. The principal component transformation produces orthogonal, statistically independent features that lack any sequential or temporal order. The recurrent connections of the long short-term memory are designed to capture dependencies across time steps, but with single-step sequences over independent components the recurrence adds only computational cost with no modeling benefit. The architecture effectively degenerates into an over-parameterized feed-forward autoencoder. The reconstruction error distributions of normal and fraud transactions overlap substantially, as Figure 4 shows, which confirms that the autoencoder cannot separate the two classes. This observation is consistent with reports that variational and attention-based autoencoders need structure-aware inputs to perform well [28], [29].

G. Limitations of the One-Class Support Vector Machine

The one-class support vector machine reaches a moderate value on the receiver operating characteristic curve, at 0.9100, but a poor value on the precision-recall curve, at 0.0616, which suggests that the decision boundary captures some fraud structure but produces excessive false positives. Its computational limits required training on only 10,000 of the 60,000 training samples, which likely limited the precision of the boundary in the 30-dimensional space [33], [34].

H. Practical Implications

These results carry two practical lessons. First, the area under the receiver operating characteristic curve can mislead under extreme imbalance, since the value of 0.815 for the autoencoder masks an essentially random performance on the minority class. Practitioners should prioritize the area under the precision-recall curve or precision-at-top-k metrics [8]. Second, model selection should be driven by data characteristics rather than architecture sophistication. Features transformed by principal component analysis favor methods that assess isolation or density in the feature space rather than sequential methods [48].

I. Generalizability and Robustness

Two factors support cautious generalization, and one limits it. First, the bootstrap intervals for the area under the precision-recall curve do not overlap across the three detectors, so the ordering is stable under resampling rather than a single-draw artifact. Second, the mechanism is general: the principal-component transformation removes the sequential and relational structure that the autoencoder and graph or transformer models rely on, while leaving the sparse-occupancy structure that isolation exploits, a setting common to privacy-protected financial features. The limiting factor is that the data are synthetic and drawn from one generation process, so the absolute scores should not be read as production performance, only the relative ordering under a controlled design. Comparative studies on real card data report the same qualitative difficulty for reconstruction- and boundary-based methods under extreme imbalance [68], which raises confidence in the transferable finding.

J. Comparison with Prior Work

The finding that an isolation method outperforms reconstruction- and boundary-based detectors is consistent with the wider literature, in which no single technique dominates across data sets and feature representations [59], [60]. Reports that graph and transformer models excel rely on relational or sequential structure that the principal-component transformation removes [51], [56], which explains why such structure-dependent methods are expected to struggle on the present features. Generative oversampling and autoencoder-adversarial hybrids address imbalance at the data level rather than the detector level [61], [62], and remain complementary to the isolation approach studied here.

K. Threats to Validity

Several factors bound these conclusions. Construct validity depends on the synthetic data, which mirror the public benchmark statistics but inject the fraud signal in a controlled way that may not match every real pattern. Internal validity is supported by a fixed split, native thresholds, and significance testing, yet the autoencoder and the one-class machine were not exhaustively tuned, and the latter trained on a subset for tractability. External validity is limited by the single data set and the static evaluation, since streaming operation with concept drift and incremental updating would better test transfer to production, where label delay and evolving tactics dominate [74], [75].

L. Limitations

Beyond the threats noted above, three narrower limitations remain. The principal-component transformation may not reflect all production environments, multi-step windowing for the autoencoder was not explored, and threshold optimization was not performed jointly across models. Future work should evaluate on real-world data sets and explore generative oversampling, graph-based methods, and concept-drift handling that have recently advanced card-fraud detection [39], [44], [71].

M. Alert Budget and Human Review

A further lesson concerns the alert budget. With about 70 fraudulent transactions in a test set of 20,000, a review team can act on only a limited number of daily alerts, so the precision at the operating point, not recall alone, decides whether the system is affordable. The isolation forest precision of 0.3605 means roughly one in three alerts is genuine fraud, which is workable, whereas the autoencoder precision of 0.0203 means about fifty false alerts for every true one, which is not. Choosing a detector therefore amounts to selecting the one that surfaces the most true fraud within a fixed review capacity, a criterion that the area under the

precision-recall curve approximates directly and that the area under the receiver operating characteristic curve does not.

N. Deployment Considerations

In deployment, the isolation forest offers practical advantages beyond its accuracy. It trains quickly, scores transactions in near-constant time, and exposes a single contamination parameter, which makes it easy to tune and monitor. The autoencoder needs careful threshold selection and more compute, yet returns near-random precision on these features, so its added cost is not justified here. The one-class support vector machine scales poorly with sample size, which forced training on a subset and limits its use on large transaction streams. Because fraud tactics drift over time, any deployed detector must be re-fitted on recent data and watched for a rising false-alarm rate [71]. Online and incremental designs that update the model as new labels arrive are a natural next step for keeping recall high as patterns change [72].

V. CONCLUSION

The isolation forest reaches an area under the precision-recall curve of 0.4169 and an area under the receiver operating characteristic curve of 0.9784 for fraud detection in transaction data transformed by principal component analysis, dramatically outperforming the autoencoder, at 0.0171, and the one-class support vector machine, at 0.0616, on the precision-recall metric. All pairwise differences are statistically significant, with every McNemar p-value below 0.001 and non-overlapping bootstrap intervals. Both the deep learning approach and the kernel-based approach show near-random precision on the minority class, while only the tree-based isolation method succeeds. These results show that the choice of architecture must align with feature characteristics, and that data transformed by principal component analysis specifically disadvantage sequential models while favoring isolation-based methods.

REFERENCES

[1] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert Systems with Applications*, vol. 193, pp. 116429, 2022. doi: 10.1016/j.eswa.2021.116429.

[2] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, "Credit card fraud detection in the era of disruptive technologies: A systematic review," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 1, pp. 145-174, 2023. doi: 10.1016/j.jksuci.2022.11.008.

[3] H. Fanai, and H. Abbasimehr, "A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection," *Expert Systems with Applications*, vol. 217, pp. 119562, 2023. doi: 10.1016/j.eswa.2023.119562.

[4] S. Hochreiter, and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[5] I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," *Journal of Big Data*, vol. 8, no. 1, art. 151, 2021. doi: 10.1186/s40537-021-00541-8.

[6] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation-Based Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, pp. 1-39, 2012. doi: 10.1145/2133360.2133363.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443-1471, 2001. doi: 10.1162/089976601750264965.

[8] 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, pp. e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[9] E. A. L. Marazqah Btoush, X. Zhou, R. Gururajan, K. C. Chan, R. Genrich, and P. Sankaran, "A systematic review of literature on credit card cyber fraud detection using machine and deep learning," *PeerJ Computer Science*, vol. 9, pp. e1278, 2023. doi: 10.7717/peerj-cs.1278.

[10] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700-39715, 2022. doi: 10.1109/access.2022.3166891.

[11] S. Motie and B. Raahemi, "Financial fraud detection using graph neural networks: A systematic review," *Expert Systems with Applications*, vol. 240, pp. 122156, 2024, doi: 10.1016/j.eswa.2023.122156.

[12] 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.

[13] S. Kumari, C. Prabha, A. Karim, M. M. Hassan, and S. Azam, "A Comprehensive Investigation of Anomaly Detection Methods in Deep Learning and Machine Learning: 2019–2023," *IET Information Security*, vol. 2024, no. 1, art. no. 8821891, 2024, doi: 10.1049/2024/8821891.

[14] A. Miguel-Diez, A. Campazas-Vega, C. Álvarez-Aparicio, G. Esteban-Costales, and Á. M. Guerrero-Higueras, "A systematic literature review of unsupervised learning algorithms for anomalous traffic detection based on flows," *Logic Journal of the IGPL*, vol. 34, no. 1, art. no. jzaf020, 2025, doi: 10.1093/jigpal/jzaf020.

[15] R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, "Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN," *IEEE Access*, vol. 12, pp. 54382-54396, 2024. doi: 10.1109/access.2024.3385781.

[16] T. Albalawi, and S. Dardouri, "Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation," *Frontiers in Artificial Intelligence*, vol. 8, art. 1643292, 2025. doi: 10.3389/frai.2025.1643292.

[17] H. Abbasimehr, and H. Fanai, "Credit card fraud detection using a Bayesian-optimized deep supervised autoencoder with blending ensemble learning," *Iran Journal of Computer Science*, vol. 9, no. 1, art. 10, 2026. doi: 10.1007/s42044-025-00374-1.

[18] D. Sehrawat and Y. Singh, "Auto-Encoder and LSTM-Based Credit Card Fraud Detection," *SN Computer Science*, vol. 4, no. 5, art. no. 557, 2023, doi: 10.1007/s42979-023-01977-w.

[19] J. Forough, and S. Momtazi, "Ensemble of deep sequential models for credit card fraud detection," *Applied Soft Computing*, vol. 99, pp. 106883, 2021. doi: 10.1016/j.asoc.2020.106883.

[20] K. EL Hachimi, G. Orhanou, and L. Ballihi, "Credit card fraud detection model based on optimized long short-term memory : Imbalanced class solution and overfitting reduction," *Expert Systems with Applications*, vol. 305, pp. 130871, 2026, doi: 10.1016/j.eswa.2025.130871.

[21] I. Akour, N. Mohamed, and S. Salloum, "Hybrid CNN-LSTM With Attention Mechanism for Robust Credit Card Fraud Detection," *IEEE Access*, vol. 13, pp. 114056-114068, 2025, doi: 10.1109/access.2025.3583253.

[22] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection," *IEEE Access*, vol. 10, pp. 16400-16407, 2022. doi: 10.1109/access.2022.3148298.

[23] I. D. Mienye, and Y. Sun, "A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection," *IEEE Access*, vol. 11, pp. 30628-30638, 2023. doi: 10.1109/access.2023.3262020.

[24] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An

Ensemble Machine Learning Approach," *Big Data and Cognitive Computing*, vol. 8, no. 1, pp. 6, 2024. doi: 10.3390/bdcc8010006.

[25] E. Ileberi and Y. Sun, "A Hybrid Deep Learning Ensemble Model for Credit Card Fraud Detection," *IEEE Access*, vol. 12, pp. 175829-175838, 2024, doi: 10.1109/access.2024.3502542.

[26] A. RB, and S. K. KR, "Credit card fraud detection using artificial neural network," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35-41, 2021. doi: 10.1016/j.gltp.2021.01.006.

[27] F. Carcillo, Y. A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Information Sciences*, vol. 557, pp. 317-331, 2021. doi: 10.1016/j.ins.2019.05.042.

[28] F. Alshameri and R. Xia, "An Evaluation of Variational Autoencoder in Credit Card Anomaly Detection," *Big Data Mining and Analytics*, vol. 7, no. 3, pp. 718-729, 2024, doi: 10.26599/bdma.2023.9020035.

[29] S. Shi, W. Luo, and G. Pau, "An attention-based balanced variational autoencoder method for credit card fraud detection," *Applied Soft Computing*, vol. 177, pp. 113190, 2025, doi: 10.1016/j.asoc.2025.113190.

[30] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended Isolation Forest," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1479-1489, 2021, doi: 10.1109/tkde.2019.2947676.

[31] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep Isolation Forest for Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12591-12604, 2023, doi: 10.1109/tkde.2023.3270293.

[32] M. Carletti, M. Terzi, and G. A. Susto, "Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest," *Engineering Applications of Artificial Intelligence*, vol. 119, pp. 105730, 2023, doi: 10.1016/j.engappai.2022.105730.

[33] J. L. Leevy, J. Hancock, and T. M. Khoshgoftaar, "Comparative analysis of binary and one-class classification techniques for credit card fraud data," *Journal of Big Data*, vol. 10, no. 1, art. no. 118, 2023, doi: 10.1186/s40537-023-00794-5.

[34] J. L. Leevy, J. Hancock, T. M. Khoshgoftaar, and A. Abdollah Zadeh, "Investigating the effectiveness of one-class and binary classification for fraud detection," *Journal of Big Data*, vol. 10, no. 1, art. no. 157, 2023, doi: 10.1186/s40537-023-00825-1.

[35] C. Wang, Y. Sun, S. Lv, C. Wang, H. Liu, and B. Wang, "Intrusion Detection System Based on One-Class Support Vector Machine and Gaussian Mixture Model," *Electronics*, vol. 12, no. 4, pp. 930, 2023, doi: 10.3390/electronics12040930.

[36] W. Yao, L. Hu, Y. Hou, and X. Li, "A Lightweight Intelligent Network Intrusion Detection System Using One-Class Autoencoder and Ensemble Learning for IoT," *Sensors*, vol. 23, no. 8, pp. 4141, 2023, doi: 10.3390/s23084141.

[37] A. G. Ayad, M. M. El-Gayar, N. A. Hikal, and N. A. Sakr, "Efficient Real-Time Anomaly Detection in IoT Networks Using One-Class Autoencoder and Deep Neural Network," *Electronics*, vol. 14, no. 1, pp. 104, 2024, doi: 10.3390/electronics14010104.

[38] E. Gencturk, B. Ustubioglu, and G. Ulutas, "Handling Imbalanced IoMT Network Data for Intrusion Detection via PCA and One-Class SVM," *Applied Sciences*, vol. 16, no. 10, pp. 4701, 2026, doi: 10.3390/app16104701.

[39] F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem, and T. Al-Hadhrami, "Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection," *IEEE Access*, vol. 11, pp. 89694-89710, 2023, doi: 10.1109/access.2023.3306621.

[40] E. Strelcenia and S. Prakoonwit, "A Survey on GAN Techniques for Data Augmentation to Address the Imbalanced Data Issues in Credit Card Fraud Detection," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 304-329, 2023, doi: 10.3390/make5010019.

[41] X. Zhao and S. Guan, "CTCN: a novel credit card fraud detection method based on Conditional Tabular Generative Adversarial Networks and Temporal Convolutional Network," *PeerJ Computer Science*, vol. 9, pp. e1634, 2023, doi: 10.7717/peerj-cs.1634.

[42] J. Chen, Y. Liang, J. Liu, and M. Zhou, "Temporal Transformer with Conditional Tabular GAN for Credit Card Fraud Detection: A Sequential Deep Learning Approach," *Mathematics*, vol. 14, no. 7, pp. 1183, 2026, doi: 10.3390/math14071183.

[43] I. D. Mienye and T. G. Swart, "A Hybrid Deep Learning Approach with Generative Adversarial Network for Credit Card Fraud Detection," *Technologies*, vol. 12, no. 10, pp. 186, 2024, doi: 10.3390/technologies12100186.

[44] G. Tong and J. Shen, "Financial transaction fraud detector based on imbalance learning and graph neural network," *Applied Soft Computing*, vol. 149, pp. 110984, 2023, doi: 10.1016/j.asoc.2023.110984.

[45] C. Lou, Y. Wang, J. Li, Y. Qian, and X. Li, "Graph neural network for fraud detection via context encoding and adaptive aggregation," *Expert Systems with Applications*, vol. 261, pp. 125473, 2025, doi: 10.1016/j.eswa.2024.125473.

[46] X. Hu, H. Chen, H. Chen, S. Liu, X. Li, S. Zhang, Y. Wang, and X. Xue, "Cost-Sensitive GNN-Based Imbalanced Learning for Mobile Social Network Fraud Detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 2675-2690, 2024, doi: 10.1109/tcss.2023.3302651.

[47] I. D. Mienye, E. Esenogho, and C. Modisane, "Detecting Imbalanced Credit Card Fraud via Hybrid Graph Attention and Variational Autoencoder Ensembles," *AppliedMath*, vol. 5, no. 4, pp. 131, 2025, doi: 10.3390/appliedmath5040131.

[48] D. Breskuvienė and G. Dzemyda, "Enhancing credit card fraud detection: highly imbalanced data case," *Journal of Big Data*, vol. 11, no. 1, art. no. 182, 2024, doi: 10.1186/s40537-024-01059-5.

[49] X. Zhao, Y. Liu, and Q. Zhao, "Improved LightGBM for Extremely Imbalanced Data and Application to Credit Card Fraud Detection," *IEEE Access*, vol. 12, pp. 159316-159335, 2024, doi: 10.1109/access.2024.3487212.

[50] M. A. Ghalwash, S. M. Abdelrazek, N. H. Eladawi, and H. A. Ghalwash, "Enhancing credit card fraud detection using DBSCAN-augmented disjunctive voting ensemble," *Scientific Reports*, vol. 15, no. 1, art. no. 39754, 2025, doi: 10.1038/s41598-025-22960-w.

[51] D. Cheng, Y. Zou, S. Xiang, and C. Jiang, "Graph neural networks for financial fraud detection: a review," *Frontiers of Computer Science*, vol. 19, no. 9, art. no. 199609, 2025, doi: 10.1007/s11704-024-40474-y.

[52] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu, "A Comprehensive Survey on Graph Anomaly Detection With Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12012-12038, 2023, doi: 10.1109/tkde.2021.3118815.

[53] H. Qiao, H. Tong, B. An, I. King, C. Aggarwal, and G. Pang, "Deep Graph Anomaly Detection: A Survey and New Perspectives," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 9, pp. 5106-5126, 2025, doi: 10.1109/tkde.2025.3581578.

[54] Y. Tang and Y. Liang, "Credit card fraud detection based on federated graph learning," *Expert Systems with Applications*, vol. 256, pp. 124979, 2024, doi: 10.1016/j.eswa.2024.124979.

[55] Y. Tian, G. Liu, J. Wang, and M. Zhou, "ASA-GNN: Adaptive Sampling and Aggregation-Based Graph Neural Network for Transaction Fraud Detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 3536-3549, 2024, doi: 10.1109/tcss.2023.3335485.

[56] C. Zhao, X. Sun, M. Wu, and L. Kang, "Advancing financial fraud detection: Self-attention generative adversarial networks for precise and effective identification," *Finance Research Letters*, vol. 60, pp. 104843, 2024, doi: 10.1016/j.frl.2023.104843.

[57] [P. Dubey, P. Dubey, and P. N. Bokoro, "A Unified Transformer–BDI Architecture for Financial Fraud Detection: Distributed Knowledge Transfer Across Diverse Datasets," *Forecasting*, vol. 7, no. 2, pp. 31, 2025, doi: 10.3390/forecast7020031.

[58] S. Tuli, G. Casale, and N. R. Jennings, "TranAD," *Proceedings of the VLDB Endowment*, vol. 15, no. 6, pp. 1201-1214, 2022, doi: 10.14778/3514061.3514067.

[59] V. Sinap, "Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets," *Turkish Journal of Engineering*, vol. 8, no. 2, pp. 196-208, 2024, doi: 10.31127/tuje.1386127.

[60] K. H. Ahmed, S. Axelsson, Y. Li, and A. M. Sagheer, "A credit card fraud detection approach based on ensemble machine learning classifier with hybrid data sampling," *Machine Learning with Applications*, vol. 20, pp. 100675, 2025, doi: 10.1016/j.mlwa.2025.100675.

[61] Y. Ding, W. Kang, J. Feng, B. Peng, and A. Yang, "Credit Card Fraud Detection Based on Improved Variational Autoencoder Generative Adversarial Network," *IEEE Access*, vol. 11, pp. 83680-83691, 2023, doi: 10.1109/access.2023.3302339.

[62] M. RezvaniNejad and A. Sabzali Yameqani, "WDAE-GAN: A hybrid dual autoencoder and generative adversarial framework with wavelet denoising for credit card fraud detection," *Expert Systems with Applications*, vol. 299, pp. 130078, 2026, doi: 10.1016/j.eswa.2025.130078.

[63] O. Ndama, S. Ndama, I. Bensassi, and E. M. En-Naimi, "A novel BERT-long short-term memory hybrid model for effective credit card fraud detection," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 15, no. 1, pp. 788, 2026, doi: 10.11591/ijai.v15.i1.pp788-797.

[64] S. E. V. S. Pillai, G. S. Nadella, K. Meduri, N. A. Priyadharsini, A. Bhuvanesh, and D. Kumar, "A walrus optimization-enhanced long short-term memory model for credit fraud detection in banking," *International Journal of Information Technology*, 2025, doi: 10.1007/s41870-025-02574-1.

[65] S. Jiang, R. Dong, J. Wang, and M. Xia, "Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network," *Systems*, vol. 11, no. 6, pp. 305, 2023, doi: 10.3390/systems11060305.

[66] L. Ding, L. Liu, Y. Wang, P. Shi, and J. Yu, "An AutoEncoder enhanced light gradient boosting machine method for credit card fraud detection," *PeerJ Computer Science*, vol. 10, pp. e2323, 2024, doi: 10.7717/peerj-cs.2323.

[67] M. Nalini, B. Yamini, C. Ambhika, and R. Siva Subramanian, "Enhancing early attack detection: novel hybrid density-based isolation forest for improved anomaly detection," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 5-6, pp. 3429-3447, 2024, doi: 10.1007/s13042-024-02460-5.

[68] E. Btoush, T. Kobbaey, H. Tamimi, and X. Zhou, "Machine Learning-Based Cyber Fraud Detection: A Comparative Study of Resampling Methods for Imbalanced Credit Card Data," *Applied Sciences*, vol. 16, no. 2, pp. 850, 2026, doi: 10.3390/app16020850.

[69] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep Learning for Time Series Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1-42, 2024, doi: 10.1145/3691338.

[70] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153-157, 1947. doi: 10.1007/bf02295996.

[71] S. Arora, R. Rani, and N. Saxena, "A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 4, art. no. e1536, 2024, doi: 10.1002/widm.1536.

[72] G. Charizanos, H. Demirhan, and D. İçen, "An online fuzzy fraud detection framework for credit card transactions," *Expert Systems with Applications*, vol. 252, pp. 124127, 2024, doi: 10.1016/j.eswa.2024.124127.

[73] H. Abbassi, S. El Mendili, and Y. Gahi, "Adaptive, Privacy-Enhanced Real-Time Fraud Detection in Banking Networks Through Federated Learning and VAE-QLSTM Fusion," *Big Data and Cognitive Computing*, vol. 9, no. 7, pp. 185, 2025, doi: 10.3390/bdcc9070185.

[74] B. Lebichot, G. M. Paldino, W. Siblini, L. He-Guelton, F. Oblé, and G. Bontempi, "Incremental learning strategies for credit cards fraud detection," *International Journal of Data Science and Analytics*, vol. 12, no. 2, pp. 165-174, 2021, doi: 10.1007/s41060-021-00258-0.

[75] E. Gadimov and E. Birihanu, "Real-time suspicious detection framework for financial data streams," *International Journal of Information Technology*, 2025, doi: 10.1007/s41870-025-02529-6.