Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU37583

A Benchmark of Unsupervised Anomaly Detectors for Credit Card Fraud under Two Training Protocols

Ruyu Guo

School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China
123090139@link.cuhk.edu.cn

Abstract. Credit card fraud detection is a critical problem in financial security, yet real-world deployment scenarios often lack labeled fraud samples at initial stages, making unsupervised anomaly detection an appealing approach. However, the comparative performance of different unsupervised detectors under varying training data compositions—whether trained exclusively on normal transactions or on mixed data containing unlabeled frauds—remains insufficiently studied. This paper presents a head-to-head benchmark of five unsupervised anomaly detectors—Mahalanobis distance, Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, and Elliptic Envelope—on the publicly available ULB Credit Card Fraud Detection dataset. Each detector is evaluated under two training protocols: novelty detection (trained solely on normal transactions) and outlier detection (trained on the full dataset containing both normal and fraudulent samples). Performance is measured using PR-AUC, Precision@k, and recall at low false positive rates. Results show that Mahalanobis distance and LOF achieve strong performance under the novelty detection protocol (PR-AUC > 0.73) but suffer catastrophic degradation when trained on mixed data, with PR-AUC dropping by 0.57 and 0.72 respectively. Isolation Forest gives solid and reliable results on both two tested protocols, and its PR-AUC score increases slightly from 0.3953 to 0.4432. After parameter adjustment (with parameters set as support_fraction=0.9, contamination=0.01), Elliptic Envelope gets moderate performance (PR-AUC≈0.37) and is almost insensible to protocol changes. The conclusions obtained in this part can provide actual instruction for choosing unsupervised detectors to be used in real fraud detection systems working under various conditions of label availability and data quality.

Keywords: unsupervised anomaly detection, credit card fraud detection, novelty detection, class imbalance, precision-recall AUC

1. Introduction

In the digital era, credit card fraud has developed into one of the most common types of financial crime. Since criminals continuously improve their attack methods to take advantage of existing prevention system loopholes, financial enterprises are under increasing pressure to apply effective detection frameworks which can adapt to newly emerged fraud patterns for themselves [1]. Though traditional rule-based fraud prevention systems play an important initial role in fighting fraud, they cannot give enough protection on their own; they must be helped by data detection systems capable

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU37583

of finding fraudulent transactions which are missed by simple rule-based detection [1]. In current many kinds of fraud detection techniques, anomaly detection—the task of finding patterns greatly different from usual—proves to be one of the effective methods. Detecting abnormal behavior is also an emerging research direction, which can detect fraud patterns that people have never seen before without prior knowledge of a certain type of fraud [1, 2].

Historically, supervised learning has been the main method used in fraud detection research, relying on labeled datasets to train classifiers that distinguish between legitimate and fraudulent transactions [1]. However, there are many fundamental problems when using supervised approaches to real-world fraud detection. First, it is very difficult to obtain accurately labeled fraud data: fraud is a rare event, and labels can only be confirmed after long investigation processes. Second, because fraud patterns change quickly, supervised models trained on historical data gradually lose their effectiveness as new fraud methods emerge. Third, extreme class imbalance is a natural feature of fraud datasets—fraudulent transactions usually account for less than 1% of all transactions in a dataset.a common problem in most transaction data sets is that the number of normal transactions is much larger than that of fraud transactions, which biases supervised classifiers toward the majority class and makes standard performance metrics like overall accuracy misleading [3]. These limitations have driven more and more researchers to use unsupervised anomaly detection methods, which do not need any labeled training data and can detect novel fraud patterns not encountered during training [2].

Unsupervised anomaly detectors identify anomalies by first modeling the underlying structure of normal data, then flagging instances that deviate from this learned structure. There are many approaches proposed in existing research. Distance-based methods, such as Mahalanobis distance, calculate how far a sample is from a multivariate normal distribution using the average value and covariance matrix of training data. Tree-based methods, like Isolation Forest, separate anomalies through random feature cutting, taking advantage of the fact that anomalies can be divided more easily because they are very different from normal points [4]. Density-based methods, such as Local Outlier Factor (LOF), compare the local average density around a given point with that of its neighbors, and mark points with significant density differences as outliers wIt has significantly lower density than its surrounding neighbors [5]. Boundary-based methods, like one-class support vector machine, learn a decision surface enclosing all normal data, treating any points outside this boundary as anomalies. Covariance-based methods, such as Elliptic Envelope, assume normal data follows a Gaussian distribution and use robust covariance estimation to identify outliers. Each of these methods embeds a different inductive bias concerning the definition of an anomaly, and their relative performance can vary greatly depending on dataset characteristics and the specific detection scenario [2].

Despite the increasing number of unsupervised anomaly detection techniques proposed, there still exist some critical issues that have not been discussed in previous work. First, most existing studies evaluate each detector separately or test them on synthetic data, which makes it difficult to compare their relative strengths and weaknesses under controlled conditions on real data. Second, the influence of training data composition—specifically, whether the detector is trained solely on normal samples (novelty detection) or on mixed data containing unlabeled anomalies (outlier detection)—has received little systematic investigation. This distinction is highly relevant in practice: in many real-world deployment scenarios, historical fraud labels are unavailable when the system is initialized, so novelty detection is often used. However, there is no thorough study about the difference between these two methods.n is needed at this stage; once some fraud cases are found and labeled, the system can be retrained under the outlier detection setting. Knowing how different

DOI: 10.54254/2753-8818/2026.PU37583

detectors react to this change is important for practical system design, but such comparative studies are rare.

To address these research gaps, this paper conducts systematic and direct comparative experiments on five unsupervised anomaly detectors using the same real-world credit card transaction dataset. This dataset was jointly released by the Machine Learning Group of the Université Libre de Bruxelles (ULB) and Worldline [6], which contains 284,807 transactions, with only 492 (0.172%) being fraudulent transactions, forming a real and challenging imbalanced scenario similar to the actual situation. Each detector is tested with two different training approaches: Protocol A (Novelty Detection), where the model is trained using only legitimate transactions; and Protocol B (Outlier Detection), where the model is trained using the entire population including all samples (both legitimate and fraudulent). Three different methods are adopted for performance evaluation metrics specifically designed for imbalanced classification tasks: the area under the precision-recall curve (PR-AUC), Precision@k, and recall at low false positive rates [3]. This paper makes three main contributions. First, a comprehensive comparison of five unsupervised anomaly detectors is conducted under the same experimental conditions, providing empirical insights into their relative effectiveness for credit card fraud detection. Second, the sensitivity of each detector to the choice of training protocol is quantified, revealing which methods are robust to training data contamination and which are critically dependent on clean normal training data. Third, a set of practical suggestions is proposed for practitioners who have to choose an unsupervised anomaly detector to be used in a fraud detection system under different situations of the availability of labels.

2. Methodology

2.1. Data source and sampling

This study utilizes the publicly available Credit Card Fraud Detection dataset published by the Machine Learning Group of Université Libre de Bruxelles (ULB) in collaboration with Worldline [6, 7]. The dataset comprises 284,807 real credit card transactions made by European cardholders in September 2013, of which only 492 (0.172%) are fraudulent, making it a highly imbalanced dataset. Each transaction is represented by 30 numerical features: Time (seconds elapsed since the first transaction), V1 through V28 (anonymized principal components obtained via principal component analysis (PCA)), and Amount (transaction amount). The label Class indicates whether a transaction is fraudulent (1) or legitimate (0).

Given that certain detectors, such as Local Outlier Factor (LOF) and Elliptic Envelope, incur high computational costs on the full dataset ($O(n^2)$ or covariance inversion), a sub-sampling strategy is adopted. All 492 fraudulent transactions are retained, while 49,508 legitimate transactions are randomly sampled from the 284,315 normal transactions, resulting in a working set of approximately 50,000 transactions. No synthetic sample generation techniques (e.g., the Synthetic Minority Over-sampling Technique (SMOTE)) are employed, ensuring that the experiment reflects a realistic unsupervised scenario. Empirical studies on extremely imbalanced datasets have shown that unsupervised detectors can perform effectively without oversampling the minority class [8]. The working set is then stratified split by Class into a training set (70%, containing 344 frauds) and a test set (30%, containing 148 frauds). Feature standardization is performed by fitting a *StandardScaler* exclusively on the training set and then applying the same transformation to the test set to prevent data leakage.

DOI: 10.54254/2753-8818/2026.PU37583

2.2. Evaluation metrics

In extremely imbalanced classification tasks, accuracy is a misleading metric: a dummy classifier that predicts all transactions as legitimate would achieve 99.83% accuracy while detecting zero frauds. Therefore, this study adopts three metrics that are sensitive to the positive (fraud) class.

The area under the precision-recall curve (PR-AUC) measures the trade-off between precision and recall across all classification thresholds. Unlike the area under the receiver operating characteristic curve (ROC-AUC), which remains overly optimistic under class imbalance, PR-AUC focuses on the performance of the minority class and is thus more informative for fraud detection [3]. A larger body of very recent work on imbalanced classification has also shown again that when the positive target class is extremely rare in a data set, precision-recall curves give much more realistic performance measures than standard ROC curves [9].

Precision@k computes the proportion of real fraud cases among the k transactions with the highest anomaly scores, i.e., the ones labeled most suspicious by the model. This metric corresponds to actual practice, since in the real world, investigators can look into only a small number of transactions manually. In this study, the Precision@k values are reported at several specific thresholds (k=20, 50, 100, and 200) to represent practical operations, and a Precision@k curve is also plotted for k from 1 to 500, sampled every 5 units, to show how detection performance changes as more investigative capacity becomes available.

Recall measured at a low FPR is the fraction of real frauds caught when the FPR is kept at a low bound. Recall values are reported at both 0.1% and 1% FPR to evaluate detector performance under strict and moderately relaxed false alarm constraints, respectively.

2.3. Five unsupervised anomaly detectors

The five anomaly detectors tested in this paper belong to four anomaly detection models, as detailed in Table 1. The commonly used anomaly detection method classification proposed in [2] is adopted, and the selected methods belong to categories that have been widely discussed in previous work: tree-based methods [4], density-based approaches [10], and one-class classification techniques [11, 12]. The Mahalanobis distance is also included as a simple baseline to determine if more complex detection methods can bring obvious performance improvements.

Table 1. Five unsupervised anomaly detectors and their core principles

English Name	Core Idea	Type
Mahalanobis Distance	Treats normal transactions as a point cloud and computes its center and covariance shape. The farther a transaction is from the center, the more suspicious it is.	Distance baseline
Isolation Forest	Randomly partitions the feature space; anomalies are easier to isolate due to their distinctiveness, resulting in shorter paths.	Tree-based
Local Outlier Factor (LOF)	Compares local density around a point with that of its neighbors. A point significantly sparser than its neighbors is considered anomalous.	Density-based
One-Class SVM	Learns a boundary enclosing the normal data; samples falling outside the boundary are flagged as anomalies.	Boundary-based
Elliptic Envelope	Assumes normal data follows a Gaussian distribution and flags samples far from the ellipse center as anomalies.	Covariance-based

All detectors are implemented using the scikit-learn library in Python. Detailed hyperparameter configurations are provided in Section 2.5 and summarized in Table 2. The theoretical foundations of Isolation Forest, LOF, and One-Class SVM are detailed in [4, 5], and [12], respectively. Recent comparisons of these methods on imbalanced datasets have concluded that their relative

DOI: 10.54254/2753-8818/2026.PU37583

performance varies substantially depending on the dataset and the level of contamination by anomalous points [8].

2.4. Two training protocols

To explore how the training data composition affects detector performance, this study uses two contrasting experimental protocols. Recent work has systematically examined the difference between novelty detection and outlier detection, highlighting that the two problem scenarios make fundamentally different assumptions about the distribution of training data [13].

Protocol A (Novelty Detection): The detector is trained only on the data of normal transactions (Class = 0). The model learns the behavior distribution of normal users, and when tested on fraud cases, which are met for the first time, it is treated as an unknown novel pattern. This setting is similar to actual deployment scenarios where the model is launched without any labeled historical fraud data.

Protocol B (Outlier Detection): The detector is trained on the full training set, which contains both normal transactions and a very small proportion of known fraud cases. The model sees the fraud patterns during training and learns to separate outlier anomalies from the overall data distribution. This scenario reflects detection systems used in places where a number of fraud cases have already been accumulated.

All five detectors were tested in two ways, resulting in 10 different experimental schemes. By comparing the results of the two protocols, the sensitivity of each detector to the presence of unlabeled anomalous objects in the training set can be determined [1, 13].

2.5. Experimental settings

All experiments are implemented in Python 3.10.12 with scikit-learn version 1.6.1, and the random seed is fixed at 42 for all randomized operations to ensure reproducibility. The working dataset is constructed by keeping all 492 fraudulent transactions and randomly sampling 49,508 legitimate transactions, for a total of 50,000 samples. A stratified 70/30 train-test split is used, and feature standardization is conducted using a StandardScaler fit exclusively on the training split. Full hyperparameter configurations for each detector are summarized in Table 2.

Table 2. Hyperparameter configurations of the five unsupervised anomaly detectors

Detector	Key Parameters
Mahalanobis Distance	Empirical mean and Moore-Penrose pseudo-inverse of the training covariance matrix
Isolation Forest	*contamination='auto', random_state=42*
Local Outlier Factor (LOF)	*novelty=True, contamination='auto'*
One-Class SVM	*nu=0.1, gamma='scale'*
Elliptic Envelope	*support_fraction=0.9, contamination=0.01, random_state=42*

3. Results

3.1. Data overview

Before model evaluation, the sampled data is examined through exploratory visualization. Figure 1 includes four subplots, which illustrate the differences in distribution between criminal transactions and legitimate transactions.

DOI: 10.54254/2753-8818/2026.PU37583

Panel (a) plots the distribution of transaction amounts (truncated at 500 for improved visibility). Though the majority of transactions from both classes cluster at small values, fraudulent transactions tend to have a higher average amount. Panel (b) displays a scatter plot of V1 vs. V2, the first two principal components of the dataset. The two classes overlap heavily in this projection, indicating that linear separation is not straightforward. Panel (c) presents class-wise boxplots of V1, revealing that fraudulent transactions have a wider interquartile range and more extreme outlier values. Panel (d) shows the distribution of log-transformed transaction amounts, which confirms that the amount feature alone is insufficient for reliable fraud identification.

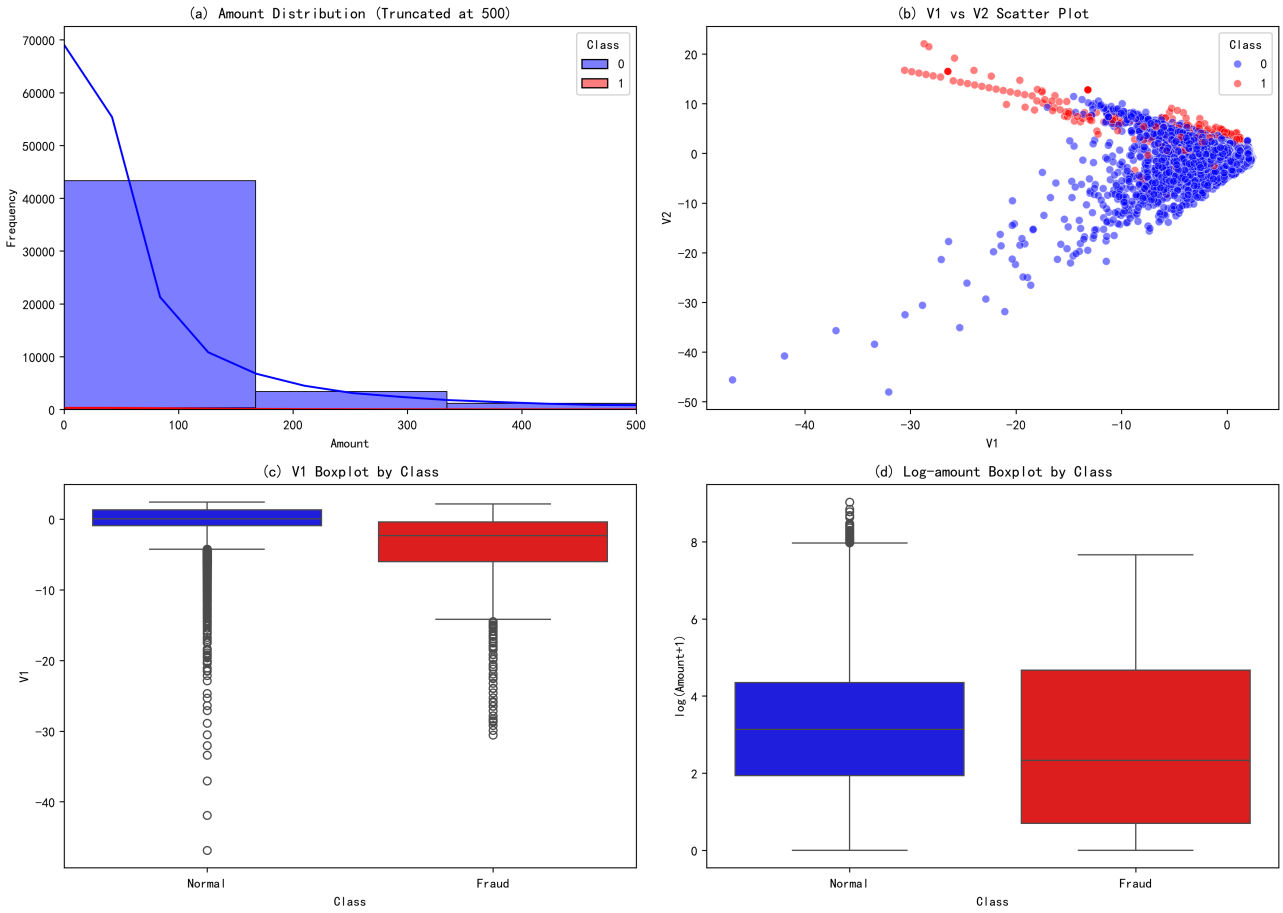

Figure 1. Exploratory analysis of the sampled dataset: (a) amount distribution by class, (b) V1-V2 scatter plot, (c) V1 boxplot by class, (d) log-amount boxplot by class

These observations show that an effective fraud detection method for this dataset is a multivariate anomaly detector, not a simple univariate threshold rule.

3.2. Main results

Table 3 reports the performance of all five detectors under both training protocols across multiple evaluation metrics. Several patterns emerge from the results.

Table 3. Performance of five detectors under two training protocols

Detector	Protocol	PR-AUC	ROC-AUC	Prec@20	Prec@50	Prec@100	Prec@200	Recall@0.1%FPR	Recall@1%FPR
Mahalanobis	Normal-Only	0.7855	0.9575	1.00	0.92	0.88	0.615	0.0946	0.8041

DOI: 10.54254/2753-8818/2026.PU37583

Table 3. (continued)

Mahalanobis	All-Data	0.2153	0.9358	0.25	0.38	0.33	0.255	0.0203	0.2568
Isolation Forest	Normal-Only	0.3953	0.9485	0.60	0.64	0.53	0.395	0.0541	0.4392
Isolation Forest	All-Data	0.4432	0.9424	0.85	0.70	0.52	0.400	0.0743	0.4595
LOF	Normal-Only	0.7324	0.9635	0.95	0.94	0.84	0.585	0.0878	0.7297
LOF	All-Data	0.0130	0.5042	0.05	0.04	0.03	0.025	0.0068	0.0270
One-Class SVM	Normal-Only	0.5166	0.9469	0.65	0.64	0.65	0.495	0.0676	0.5878
One-Class SVM	All-Data	0.2749	0.9423	0.05	0.24	0.32	0.375	0.0000	0.3649
Elliptic Envelope	Normal-Only	0.3734	0.9526	0.55	0.52	0.48	0.380	0.0338	0.4662
Elliptic Envelope	All-Data	0.3798	0.9533	0.55	0.52	0.50	0.380	0.0338	0.4662

Figure 2 plots the precision-recall curves for all five detectors under both protocols. In the clean normal-only training setting (plotted as solid lines), both Mahalanobis and LOF maintain the highest precision across most recall levels. By contrast, their curves for the all-data training setting (plotted as dashed lines) collapse toward the origin, reflecting a catastrophic drop in their PR-AUC scores. In contrast, the curves for Isolation Forest and Elliptic Envelope remain largely unchanged between the two protocols, consistent with their inherent robustness to training-set contamination. These findings align with results from recent empirical studies of anomaly detection in imbalanced settings, which have shown that density-based and distance-based methods tend to perform well when trained on clean data, whereas most other methods deteriorate significantly in such settings. SVM mostly works on clean data but are easily disturbed by the contaminated training set [8].

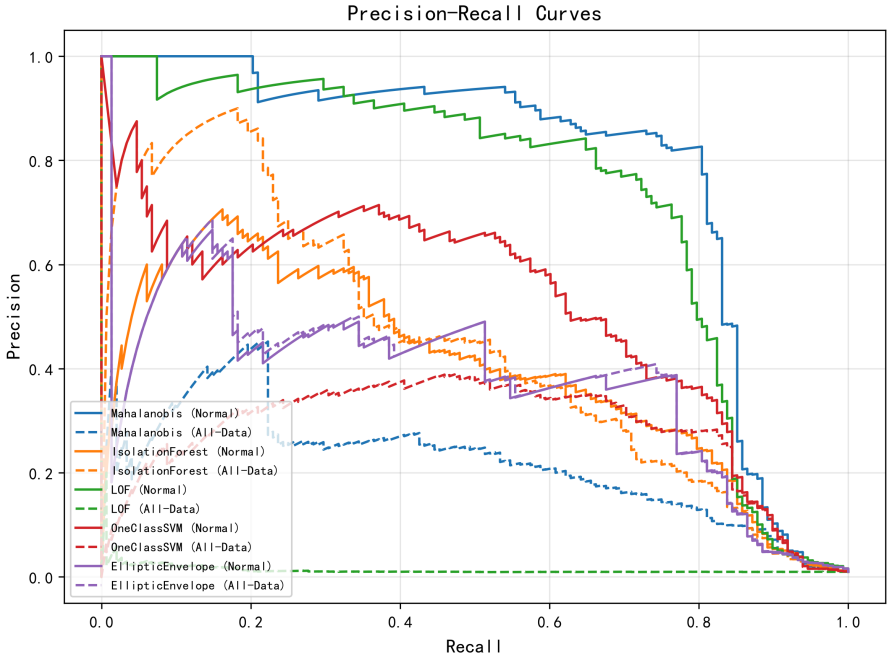

Figure 2. Precision-recall curves of the five detectors under two training protocols

Figure 3 shows the ROC curves and the corresponding AUC values of all 5 anomaly detectors under two training protocols. Among all test configurations, only the AUC value of LOF trained on all data is very low, which is merely 0.5042; the ROC-AUC scores of all other settings range from 0.9358 to 0.9635. Such near-peak performance is extremely misleading for this highly imbalanced task: Take Mahalanobis distance as an example, its ROC-AUC is 0.9575 under the protocol of using

DOI: 10.54254/2753-8818/2026.PU37583

only normal data and 0.9358 under the protocol of using all data, with almost no difference, but its PR-AUC drops from 0.7855 to 0.2153. Therefore, the ROC curve provides little useful distinguishing information and consistently overestimates model performance, which confirms that precision-recall-based metrics are much better. However, PR-AUC is more informative for fraud detection tasks.

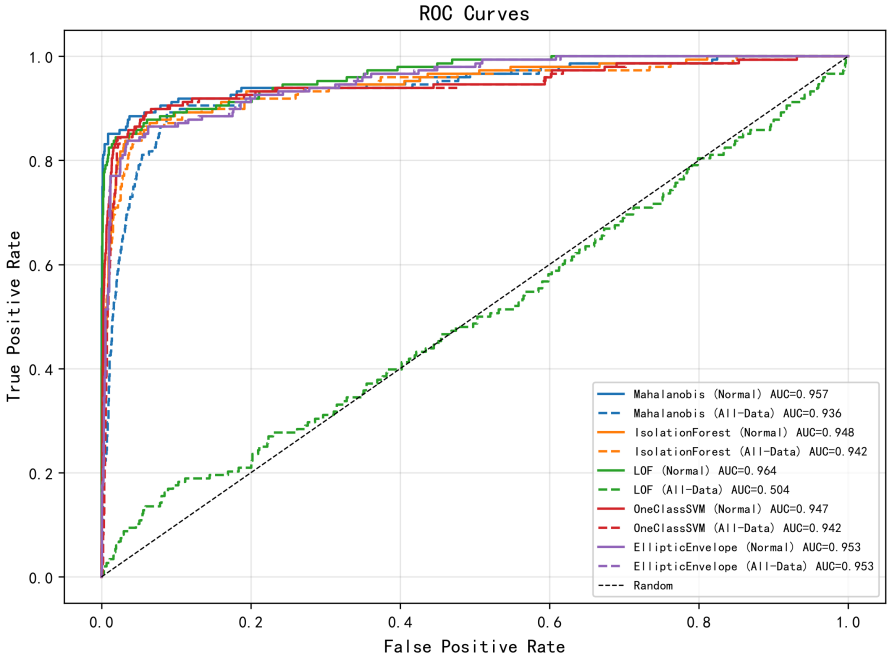

Figure 3. Receiver operating characteristic (ROC) curves of the five detectors under two training protocols

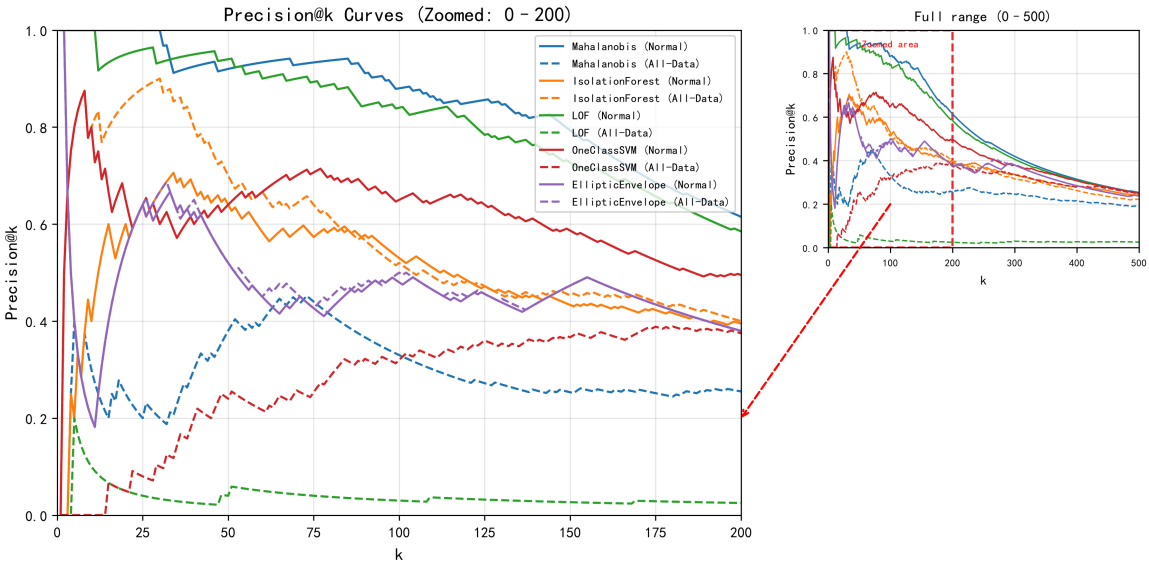

Figure 4. Precision@k curves for all detectors under two training protocols

For this reason, all subsequent analysis centers on PR-AUC and Precision@k. Figure 4 plots the Precision@k curves for all evaluated detectors. The main panel zooms in on the 0–200 range to

DOI: 10.54254/2753-8818/2026.PU37583

emphasize the quality of early ranking, while the inset panel shows the full trend across the entire 0–500 range. All analysis below draws on the discrete numerical results reported in Table 3.

The Mahalanobis distance performs very well when trained under Normal-Only: the Precision@k values are 1.00, 0.92, 0.88, and 0.615 for k=20, 50, 100, and 200, respectively. When trained under All-Data, these values drop sharply to 0.25, 0.38, 0.33, and 0.255 for the same k values. This means that the covariance matrix estimation seriously deforms when there are unlabeled anomalies in the training set, and the drop is between 0.36 and 0.75 for any k value.

LOF follows a very similar performance pattern: under Normal-Only training, its Precision@k values reach 0.95, 0.94, 0.84, and 0.585, which are nearly on par with Mahalanobis distance. In the All-Data training setup, LOF experiences a complete performance collapse, with Precision@k falling to 0.05, 0.04, 0.03, and 0.025 — barely higher than the test set's approximate 1% base rate of fraud. This matches its PR-AUC decrease from 0.7324 to 0.0130 (Table 3), and highlights that density-based methods are extremely vulnerable to contamination of the training set.

Isolation Forest shows an entirely distinct performance trend: when trained under Normal-Only, its Precision@k values are 0.60, 0.64, 0.53, and 0.395; under All-Data training, the values rise to 0.85 (k=20), 0.70 (k=50), and 0.400 (k=200), and remain almost unchanged at k=100 (shifting only from 0.53 to 0.52). This pattern indicates that the presence of unlabeled frauds in training provides extra structural information that improves the model's random partitioning process, resulting in better performance. The result is also consistent with its PR-AUC increase from 0.3953 to 0.4432.

One-Class SVM obtains moderate results under Normal-Only training, with Precision@k values of 0.65, 0.64, 0.65, and 0.495. Under All-Data training, its Precision@k values drop to 0.05, 0.24, 0.32, and 0.375 respectively; it suffers a big performance drop when k is small, but the performance can be partly restored as k increases. This indicates that the influence of training set contamination on its boundary estimation is obvious, but much smaller than that caused by Mahalanobis distance or LOF.

Only Elliptic Envelope has Precision@k values that are almost identical in both training protocols: the Normal-Only setup produces values of 0.55, 0.52, 0.48, and 0.380, while the All-Data setup produces 0.55, 0.52, 0.50, and 0.380. The largest performance difference across all four k values is just 0.02, which confirms its strong robustness to training set contamination — a feature also visible in its PR-AUC shift from 0.3734 to 0.3798. That said, its peak precision (0.55 at k=20) is much lower than the peak precision of Mahalanobis (1.00) or LOF (0.95) when trained on clean data, which illustrates a clear trade-off: robustness does not equal top-tier accuracy when the available training data is fully clean.

The overall Precision@k curves reveal a clear trade-off between performance and robustness: high-performance detectors (Mahalanobis, LOF) have very good results on clean training data, but they are extremely sensitive to training contamination; robust detectors (Isolation Forest, Elliptic Envelope) keep stable performance when the training data contains anomalies, and although their peak precision is lower, they are more reliable. These results, together with the PR-AUC values and sensitivity rankings shown in Table 4, can provide a reference for selecting detectors in actual fraud detection applications: when normal-class training data is clean, Mahalanobis or LOF should be selected; when the training data cannot be guaranteed to be pure, Isolation Forest is the most recommended.

3.3. Protocol sensitivity analysis

In order to calculate how sensitive each detector is to the choice of training protocol, Table 4 reports the performance differences based on All-Data minus the corresponding Normal-Only results for

DOI: 10.54254/2753-8818/2026.PU37583

each evaluation metric. The larger the absolute value of this difference, the more sensitive the detector is to the existence of unlabeled anomalies in the training set.

Table 4. Protocol sensitivity analysis — performance differences (All-Data minus Normal-Only)

Detector	Δ PR-AUC	Δ Precision@100	Δ Recall@1%FPR	Sensitivity Level
LOF	-0.7194	-0.81	-0.7027	Critically sensitive
Mahalanobis	-0.5702	-0.55	-0.5473	Highly sensitive
One-Class SVM	-0.2417	-0.33	-0.2229	Moderately sensitive
Isolation Forest	+0.0479	-0.01	+0.0203	Robust
Elliptic Envelope	+0.0064	+0.02	+0.0000	Robust (moderate)

These results clearly divide the five detectors into two groups. LOF and Mahalanobis distance are extremely sensitive to training set contamination: adding unlabeled fraud samples to the training set almost completely eliminates their ability to distinguish between normal samples and fraud samples. One-Class SVM shows moderate sensitivity, while Isolation Forest maintains considerable stability across all three evaluation metrics. Most notably, the Elliptic Envelope method has the smallest performance gap under the two protocols (ΔPR-AUC=0.0064, ΔPrecision@100=0.02, ΔRecall@1%FPR=0.0000), which indicates that its covariance estimation remains relatively stable regardless of the presence of unlabeled outliers in the training set. However, its overall performance (PR-AUC≈0.37) suggests that it still has a certain degree of robustness.ss does not automatically mean strong detection ability.

The training protocol sensitivity detailed here shows just how essential it is in practice to distinguish between novelty detection and outlier detection settings—the distinction having already been emphasized in most recent unified surveys of anomaly detection research [13].

4. Discussion

4.1. Interpretation of results

Three main conclusions can be drawn from the experimental results, providing useful suggestions for practitioners who select anomaly detection methods in real-world applications.

First, both Mahalanobis distance and LOF are very sensitive to high-quality clean normal training data. The two methods work very well under Protocol A, but their performance disappears completely under Protocol B. For Mahalanobis distance, this may be due to its core reliance on the global covariance matrix calculated from the training set; if unlabeled fraudulent transactions enter the training data, the estimated covariance will drift toward the outlier points, thereby distorting the distance calculation metric and making valid normal transactions look artificially far from the normal cluster. A drop of 0.5702 in PR-AUC clearly shows how sensitive this method is when training data quality cannot be guaranteed. This observed performance pattern also exists in LOF, as reported in Table 3. Based on the characteristics of gns with distance-based and density-based methods, as stated in recent systematic reviews of LOF algorithms [10] and general anomaly detection algorithms [11].

Secondly, Isolation Forest has remarkable resistance to contamination in training data. When some fraud samples are added to the training samples, its PR-AUC instead increases from 0.3953 to 0.4432. This indicates that the core random partitioning mechanism of the algorithm can extract meaningful patterns from the mixed data without being affected by outlier samples. This robustness stems from the working mechanism of Isolation Forest itself: unlike other methods, it does not

DOI: 10.54254/2753-8818/2026.PU37583

explicitly model the distribution of normal data. Instead, it detects anomalies by continuously randomly partitioning the feature space, a working mode that makes it far less affected by other factors than the composition of the training set. The self-storing ability of this algorithm has been documented in many papers. In recent years, many researchers have carried out research on the robustness of Isolation Forest, and systematic reviews have verified its effectiveness across various application fields.

Third, after hyperparameter tuning, Elliptic Envelope shows relatively good overall performance, with PR-AUC approximately equal to 0.37 under both testing methods. Even so, it still underperforms the methods based on Mahalanobis distance and LOF in the clean training setting of Protocol A. This result indicates that although careful parameter tuning can improve the performance of covariance-based methods (e.g., support_fraction=0.9, contamination=0.01), they are still sensitive to their own distribution assumptions. The PCA-transformed features in this dataset do not strictly conform to the multivariate Gaussian distribution required by this method, but with appropriate robust covariance estimation, Elliptic Envelope can still partially capture abnormal transaction patterns. As a boundary-based method, One-Class SVM has a moderate sensitivity to the contamination of training data, which agrees with the conclusions obtained in wider one-class classification research [12].

4.2. Limitations and future work

Several limitations of this study should be acknowledged. First, although sub-sampling is necessary for computational feasibility, this step may affect the ranking stability of precision when k is very large. Secondly, features V1-V28 are PCA components without clear business meaning, which limits the practical explainability of the model's detection decisions. Thirdly, the data only includes two days' worth of European transactions, so the generalizability of these findings to other regions or time periods remains to be validated. Fourthly, due to the extremely low fraud rate of the task, the test set only contains 148 fraudulent samples, which may introduce certain variance to the evaluation metrics.

Future research can extend this benchmark in three directions. First, detectors with good scalability (e.g., Isolation Forest) can be evaluated on the whole original dataset to eliminate the possible bias brought by sampling. Second, adopting temporal dataset splitting instead of random splitting can better reflect the real situation: when models are trained on historical data, they are evaluated on future incoming transactions. Third, more detection methods can be added for further comparison, such as autoencoders or graph-based anomaly detectors..

5. Conclusion

Several conclusions can be drawn from these results. When clean normal data is used for training, Mahalanobis distance and LOF achieve the best performance, with PR-AUC above 0.73, but both degrade severely when the training set contains unlabeled anomalies, with PR-AUC dropping by 0.57 and 0.72, respectively. One-Class SVM ranks in the middle among the tested methods, with moderate sensitivity to contamination, though it still cannot reach the peak precision of distance-based and density-based methods when the training set is free of contamination. Isolation Forest is extremely robust to this kind of contamination, and its PR-AUC slightly improves from 0.3953 to 0.4432 across the two protocols, making it the safest method when data purity cannot be guaranteed. Elliptic Envelope shows very small sensitivity to protocol changes (ΔPR-AUC = 0.0064) and delivers stable though moderate performance, making it suitable for scenarios where stability is

DOI: 10.54254/2753-8818/2026.PU37583

prioritized over peak detection precision. These findings confirm that the choice of training protocol matters as much as the choice of detector itself.

These results translate into clear concrete guidance for practitioners: Mahalanobis distance or LOF is preferable when clean historical data is available, while Isolation Forest is the safer choice when label scarcity or data contamination is a core concern. The protocol sensitivity ranking reported in this study can also serve as a useful reference for detector selection in other imbalanced classification domains, such as network intrusion detection and industrial fault monitoring. Future work will extend this benchmark to the full original dataset, adopt temporal evaluation splits that better reflect real-world deployment conditions, and incorporate deep-learning-based detectors such as autoencoders.

References

[1] Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications, 193*, 116429.

[2] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1-58.

[3] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.

[4] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data, 6*(1), 1-39.

[5] Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93-104.

[6] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797.

[7] Le Borgne, Y.-A., Siblini, W., Lebichot, B., & Bontempi, G. (2022). *Reproducible machine learning for credit card fraud detection: Practical handbook*. Université Libre de Bruxelles.

[8] Fontes, G., Matos, L. M., Matta, A., Pilastri, A., & Cortez, P. (2022). An empirical study on anomaly detection algorithms for extremely imbalanced datasets. *Artificial Intelligence Applications and Innovations (AIAI 2022), IFIP Advances in Information and Communication Technology, vol. 646*. Springer, 85–95.

[9] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432.

[10] Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2021). A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing, 5*(1), 1.

[11] Samariya, D., & Thakkar, A. (2023). A comprehensive survey of anomaly detection algorithms. *Annals of Data Science, 10*(3), 829–850.

[12] Seliya, N., Abdollah Zadeh, A., & Khoshgoftaar, T. M. (2021). A literature review on one-class classification and its potential applications in big data. *Journal of Big Data, 8*(1), 122.

[13] Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., & Sabokrou, M. (2022). A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*.