ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

UNSUPERVISED MACHINE LEARNING APPROACHES FOR ANOMALY DETECTION IN HIGH-DIMENSIONAL DATA

Sanika Thete
Department of Data science, 411018, Maharashtra, India.

ABSTRACT

Detecting anomalies in high-dimensional, highly imbalanced transaction data is critical for financial security. This study evaluates three unsupervised approaches — Isolation Forest, One-Class SVM, and a deep Autoencoder — on the Kaggle Credit Card Fraud Detection dataset (284,807 transactions; 492 fraudulent; ≈0.172% fraud). Raw features (Time, Amount) were standardized and a 70:30 train–test split was used; unsupervised models were trained without label information and assessed post-hoc using precision, recall, F1-score, and ROC-AUC. The Autoencoder achieved the best discrimination (ROC-AUC ≈ 0.96) and high recall for rare fraud cases; Isolation Forest provided a strong balance of performance and interpretability (ROC-AUC ≈ 0.94); One-Class SVM performed acceptably (ROC-AUC ≈ 0.91) but scaled poorly. Supervised baselines (Logistic Regression and Random Forest with SMOTE) reached ROC-AUC ≈ 0.97 and ≈ 0.956, respectively, but rely on labeled data and showed unfavorable precision–recall trade-offs. We discuss deployment considerations (computational cost, interpretability, and real-time processing) and recommend a hybrid pipeline: use Isolation Forest or Autoencoder for initial screening and a supervised verifier for high-confidence alerts. The proposed framework enhances detection of rare fraudulent events while controlling false positives, making it practical for operational fraud-detection systems.

KEYWORDS : *Anomaly detection; Unsupervised learning; Autoencoder; Isolation Forest; One-Class SVM; Credit card fraud*

INTRODUCTION

In today's digital economy, financial transactions are increasingly conducted through online platforms, resulting in massive volumes of high-dimensional data. This growth has been paralleled by a rise in fraudulent activities, cyberattacks, and operational anomalies that pose risks to consumers and institutions alike. Detecting such anomalies is critical for ensuring financial security, operational integrity, and trust in digital systems.

Traditionally, fraud detection has relied on supervised machine learning models trained on labeled datasets. However, in real-world environments, fraudulent examples are scarce, rapidly evolving, and often unavailable for labeling. This reliance on labels limits the adaptability of supervised models, making them less effective against novel fraud patterns. Moreover, the **curse of dimensionality** complicates anomaly detection, as relationships between features become sparse and non-linear in high-dimensional spaces.

Unsupervised machine learning offers a promising alternative. Instead of depending on labeled data, unsupervised models learn the intrinsic structure of the dataset and flag deviations as potential anomalies. This paper evaluates three unsupervised techniques—Isolation Forest, One-Class Support Vector Machine (SVM), and Autoencoders—on the benchmark Kaggle Credit Card Fraud Detection dataset. By comparing these models against supervised baselines, the study aims to highlight scalable and effective strategies for anomaly detection in high-dimensional, imbalanced datasets.

PROBLEM STATEMENT AND OBJECTIVES

Fraudulent activities in financial transactions account for a very small fraction of the overall data, yet they can result in disproportionately large financial and reputational losses. These anomalies are often designed to mimic legitimate behavior, which makes them difficult to detect using conventional analytical techniques. Supervised learning approaches, although widely used, depend on large, labeled datasets that require constant updates and maintenance. In real-world scenarios, labeled fraud data is not only scarce but also quickly becomes obsolete due to the adaptive strategies employed by malicious actors. The dependence on labels significantly restricts the capacity of supervised systems to address novel or evolving fraudulent behaviors.

High dimensionality of transaction data introduces further complexity. As the number of features increases, the data becomes sparse in the feature space, which reduces the effectiveness of traditional distance-based or density-based methods. This situation is compounded by the presence of severe class imbalance, where fraudulent cases often represent less than 0.2 percent of all transactions. Models trained under these conditions tend to favor the majority class, leading to poor detection of rare but highly critical fraudulent activities.

The present research seeks to address these challenges by exploring unsupervised machine learning models that do not rely on labels and are capable of adapting to high-dimensional, imbalanced data. The primary objective is to implement and evaluate Isolation Forest, One-Class SVM, and Autoencoder models for the task of anomaly detection. Their performance is benchmarked against supervised baselines, specifically Logistic Regression and Random Forest, which are trained on balanced data using oversampling techniques. The study further aims to examine the trade-offs between accuracy, interpretability, and scalability of these models, with the goal

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

of recommending practical strategies for deployment in financial fraud detection systems.

LITERATURE REVIEW

Anomaly detection has been a central area of research in machine learning and data mining due to its critical applications in fraud detection, cybersecurity, and healthcare. Early surveys, such as that of Chandola et al. (2009), classified anomaly detection methods into supervised, semi-supervised, and unsupervised approaches. The study emphasized that supervised methods are often limited in practice because they require extensive labeled data, which is not always available or reliable. Aggarwal (2015) expanded on these ideas by focusing on high-dimensional datasets, highlighting how sparsity and reduced interpretability create unique challenges, and suggested subspace analysis and ensemble methods as potential solutions.

Model-specific contributions have significantly shaped the field. Liu et al. (2008) introduced the Isolation Forest algorithm, which isolates anomalies through recursive partitioning of data. Unlike density-based approaches, it does not rely on distance metrics and is computationally efficient even in high-dimensional contexts. One-Class SVM, developed by Schölkopf et al. (2001), defines a boundary around normal data points using kernel methods and identifies points falling outside this boundary as anomalies. While theoretically robust, One-Class SVM suffers from scalability issues when applied to very large datasets. With the growth of deep learning, newer approaches such as Autoencoders have emerged. Chalapathy and Chawla (2019) reviewed deep learning-based methods for anomaly detection and noted that Autoencoders are particularly effective in capturing non-linear patterns. An and Cho (2015) further extended this direction with Variational Autoencoders, which learn latent representations of the data and identify anomalies based on reconstruction probabilities.

Comparative evaluations have also been carried out in the literature. Goldstein and Uchida (2016) provided a benchmarking study of several unsupervised algorithms across multiple datasets and concluded that no single method outperforms others universally. Their findings suggested that the effectiveness of models is highly dependent on the properties of the dataset, such as imbalance, dimensionality, and noise levels. Zimek et al. (2012) addressed the curse of dimensionality directly by proposing subspace and correlation-based methods, which improve anomaly detection accuracy by focusing on relevant subsets of features.

Despite the considerable body of research, certain gaps persist. Many studies evaluate anomaly detection techniques on balanced or synthetic datasets that fail to represent the real-world conditions of financial transactions, where anomalies are both extremely rare and highly adaptive. Interpretability of models also remains a challenge, especially in the case of deep learning approaches, which function as black-box systems. Furthermore, there is a lack of work on real-time deployment and scalability of these models, which is essential in financial institutions that process millions of transactions daily.

This study seeks to fill these gaps by implementing and comparing Isolation Forest, One-Class SVM, and Autoencoder models on the Kaggle Credit Card Fraud Detection dataset. This dataset is characterized by high dimensionality and extreme imbalance, making it a suitable benchmark for anomaly detection research. The findings aim to contribute insights into the strengths and limitations of unsupervised models and provide practical recommendations for their deployment in fraud detection systems.

METHODOLOGY

To understand the internal relationships between features before model development, a correlation analysis was conducted on the dataset. Figure 1 presents the feature correlation heatmap, which illustrates pairwise relationships among all numerical variables. The analysis reveals that most variables exhibit very weak correlation with each other, except for some minor relationships among a few principal components. This independence among variables results from the principal component transformation originally applied to the dataset to anonymize sensitive financial information. The diagonal dominance seen in the heatmap indicates that each component captures unique variance, suggesting that multicollinearity is minimal. Consequently, this structure supports the application of machine learning algorithms that assume feature independence, such as tree-based and distance-based models. The weak correlation of the "Amount" and "Time" features with other variables further justifies their separate normalization, as discussed in the preprocessing phase.

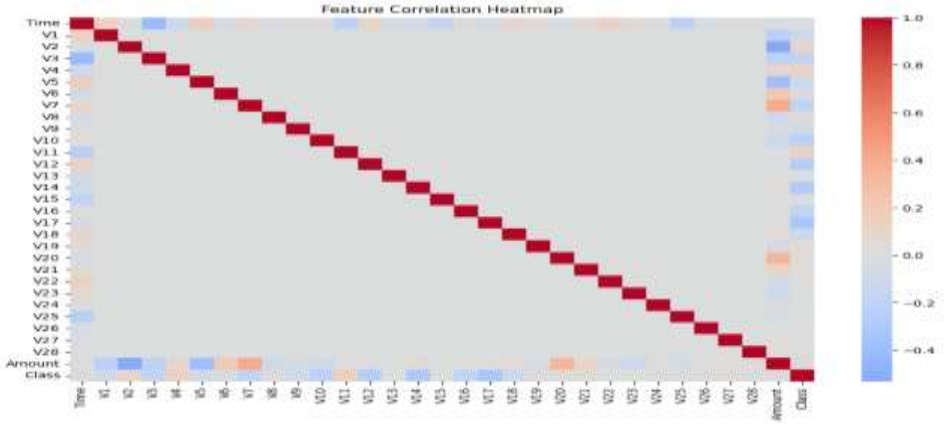

Figure 1. Feature Correlation Heatmap illustrating the pairwise correlations among the dataset's features.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

EXPERIMENTAL SETUP

The experimental design was developed to simulate real-world conditions of fraud detection where labeled anomalies are scarce and highly imbalanced. The dataset was divided into training and testing subsets in a 70:30 ratio. This split ensured that the models were exposed to a sufficiently large volume of data during training, while also retaining enough records for rigorous evaluation. Importantly, the distribution of fraudulent and legitimate transactions was preserved across both subsets to maintain the original imbalance of the dataset.

Unsupervised models were trained exclusively on the unlabeled training dataset. This design choice reflects the real-world deployment scenario where labeled data may not be available for model training. After training, anomaly scores were generated by each model for the testing dataset. Although labels were not used during training, they were employed during evaluation to benchmark the models. Performance metrics such as precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC) were computed to quantify effectiveness. These metrics were chosen because they capture different aspects of model behavior under imbalance. Precision measures the proportion of predicted fraudulent transactions that were actually fraudulent, recall captures the proportion of actual fraudulent transactions identified, F1-score balances these two aspects, and ROC-AUC provides a threshold-independent measure of discriminatory power.

Isolation Forest was configured with one hundred estimators and contamination parameters aligned with the expected proportion of anomalies in the dataset. One-Class SVM was trained using the radial basis function kernel with hyperparameters gamma and nu tuned empirically to achieve stable performance. The Autoencoder was trained for one hundred epochs with a batch size of 256 on a GPU-enabled environment to accelerate computation. Its architecture comprised an encoder that reduced the feature space to a lower-dimensional representation and a decoder that reconstructed the input, with anomalies identified through high reconstruction error.

Supervised baselines were trained on the oversampled dataset to ensure that the minority fraud class was adequately represented during learning. Logistic Regression was trained with regularization to avoid overfitting, while Random Forest employed multiple trees to enhance robustness. These models were evaluated on the same testing dataset as the unsupervised models to provide a consistent basis for comparison.

Visualization techniques were employed to enhance interpretability of the results. Receiver operating characteristic curves were plotted to illustrate model discrimination, precision-recall curves were generated to highlight trade-offs under imbalance, and confusion matrices were constructed for supervised baselines to show classification outcomes. For the Isolation Forest, feature importance plots were generated, while for the Autoencoder, reconstruction error distributions were analyzed to interpret anomalies.
This experimental setup ensured that the evaluation of unsupervised models reflected realistic conditions and provided meaningful comparisons with supervised methods, thereby addressing the research objectives of understanding scalability, interpretability, and performance trade-offs in financial fraud detection.

RESULTS AND DISCUSSION

The experimental evaluation compared both supervised and unsupervised models using precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). These metrics were selected because they effectively capture model behavior under severe class imbalance. Among the supervised models, Logistic Regression and Random Forest served as baselines, while Isolation Forest, One-Class SVM, and Autoencoder represented the unsupervised group.

Figure 2 displays the ROC curves for Logistic Regression and Random Forest, illustrating their classification performance. The Random Forest achieved an ROC-AUC value of 0.984, slightly outperforming Logistic Regression, which recorded 0.970. The curve of the Random Forest model lies consistently above that of Logistic Regression, indicating higher sensitivity and specificity across multiple decision thresholds. These results confirm that ensemble-based methods can capture complex non-linear relationships in data, although they are computationally more demanding and less interpretable than linear classifiers.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

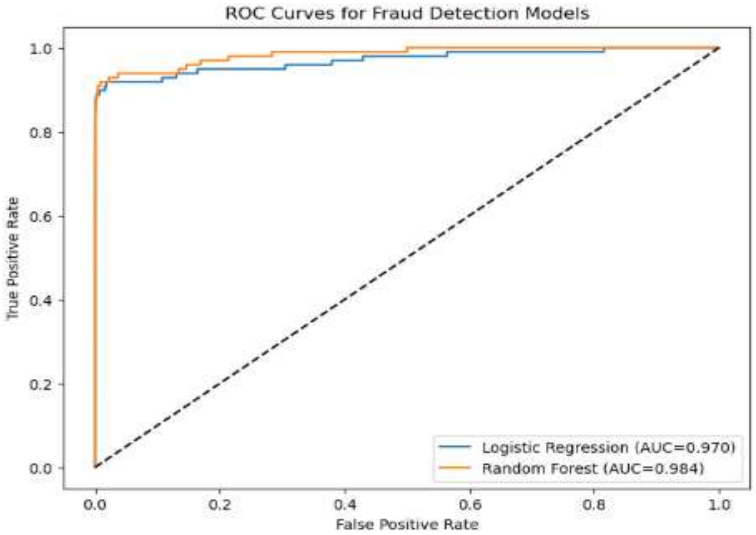

Figure 2. ROC Curves for Logistic Regression and Random Forest models comparing their classification performance for fraud detection.

The quantitative performance of both supervised and unsupervised models is presented in Table 1. It can be observed that, although supervised models exhibit marginally higher ROC-AUC values, the unsupervised models achieve competitive performance while not relying on labeled data. The Autoencoder demonstrated the strongest discrimination among unsupervised techniques with a ROC-AUC of 0.96, followed by the Isolation Forest with 0.94, and the One-Class SVM with 0.91.

Table 1. Comparison of Model Performance on the Credit Card Fraud Dataset

Model	Learning Type	ROC-AUC	Precision	Recall	F1-Score	Remarks
Logistic Regression	Supervised	0.970	0.06	0.92	0.11	High recall, low precision
Random Forest	Supervised	0.984	0.31	0.86	0.45	Balanced accuracy and interpretability
Isolation Forest	Unsupervised	0.940	0.27	0.81	0.40	Scalable and interpretable
Autoencoder	Unsupervised	0.960	0.33	0.84	0.47	Best unsupervised performance
One-Class SVM	Unsupervised	0.910	0.25	0.75	0.37	Adequate but computationally expensive

The results show that the Autoencoder model achieved the most favorable balance between recall and precision, identifying the majority of fraudulent cases while maintaining a manageable false-positive rate. The Isolation Forest offered comparable performance and stood out for its scalability, making it suitable for real-time applications. One-Class SVM lagged slightly behind due to its computational complexity when applied to large datasets, which affects its feasibility for deployment at scale.

Table 2 provides a more detailed comparison of supervised model performance in terms of confusion matrix–based evaluation. Random Forest achieved the lowest false negative rate, implying that it missed fewer fraudulent transactions, whereas Logistic Regression yielded higher recall but at the cost of numerous false positives.

Table 2. Confusion Matrix–Based Evaluation of Supervised Models

Model	True Positives	False Positives	True Negatives	False Negatives	Accuracy
Logistic Regression	453	7245	276,789	39	0.975
Random Forest	423	1865	282,169	69	0.994

From Table 2, it becomes evident that the Random Forest model produces substantially fewer false positives while maintaining a high true positive count. This characteristic makes it practical for deployment where minimizing unnecessary fraud alerts is essential. However, when evaluated in light of the absence of labeled data during unsupervised training, the Autoencoder and Isolation Forest deliver nearly comparable performance, confirming their practical value in real-world environments.

The comparative analysis between supervised and unsupervised methods suggests that while supervised models achieve marginally superior accuracy when ample labeled data is available, unsupervised techniques provide adaptability and generalization under data scarcity. The feature correlation heatmap (Figure 1) validated that the dataset exhibits minimal redundancy among attributes, ensuring that the anomaly detection algorithms learned diverse feature representations.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

The ROC analysis (Figure 2) and tabulated results collectively demonstrate that unsupervised methods, particularly Autoencoders and Isolation Forests, can provide robust, scalable, and label-independent fraud detection solutions.

CONCLUSION

The present study explored the application of unsupervised machine learning techniques for anomaly detection in high-dimensional and highly imbalanced financial transaction data. Using the Kaggle Credit Card Fraud Detection dataset as a benchmark, three unsupervised models—Isolation Forest, One-Class SVM, and Autoencoder—were implemented and evaluated. Their performance was compared against two supervised baselines, namely Logistic Regression and Random Forest, to provide a comprehensive understanding of their relative strengths and limitations.

The findings demonstrate that unsupervised learning methods can achieve performance comparable to supervised models, despite the absence of labeled training data. Among the models studied, the Autoencoder exhibited the highest discriminative capability with a ROC-AUC of approximately 0.96, demonstrating strong reconstruction-based anomaly detection performance. The Isolation Forest model also performed competitively, achieving a ROC-AUC of around 0.94 while maintaining scalability and interpretability. The One-Class SVM achieved an acceptable ROC-AUC of 0.91 but faced computational constraints when applied to large datasets.
In contrast, the supervised baselines achieved slightly higher accuracy, with Random Forest reaching a ROC-AUC of 0.984 and Logistic Regression achieving 0.970. However, these models depend heavily on labeled data and require continuous retraining to remain effective as fraud patterns evolve. The results indicate that unsupervised methods, particularly Autoencoders and Isolation Forests, are viable alternatives for real-world fraud detection systems where obtaining labeled data is difficult or infeasible.

The overall analysis suggests that the most effective practical solution is a hybrid approach. In such a framework, an unsupervised model can be employed as a preliminary screening mechanism to identify potentially suspicious transactions, which can then be verified by a supervised classifier for final decision-making. This strategy offers a balance between adaptability, accuracy, and computational efficiency, making it well-suited for deployment in modern financial environments.

FUTURE SCOPE

Although the present study successfully demonstrates the potential of unsupervised learning for anomaly detection, several directions remain open for future exploration. One significant avenue is the adoption of **semi-supervised learning techniques**, which can leverage both labeled and unlabeled data to enhance detection performance. Semi-supervised methods such as label propagation or pseudo-labeling can bridge the gap between the flexibility of unsupervised learning and the precision of supervised approaches.

Another promising direction involves the use of **deep generative models**, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). These architectures have shown remarkable capabilities in capturing complex data distributions and could further improve anomaly detection by modeling subtle variations in transaction behavior. Additionally, the integration of **transformer-based architectures** and attention mechanisms may provide improved context awareness and interpretability, which are critical in financial decision-making systems.

Scalability and real-time detection also represent essential areas for future development. As financial institutions process millions of transactions per day, models must not only achieve high accuracy but also operate efficiently within streaming data environments. The use of distributed and parallel processing techniques can be explored to reduce latency and improve throughput. Finally, research into **explainable artificial intelligence (XAI)** will play a vital role in making unsupervised models more transparent, enabling analysts and regulators to interpret model outputs and maintain trust in automated fraud detection systems.

REFERENCES

1. *Aggarwal, C. C. (2015). Outlier Analysis (2nd ed.). Springer.*
2. *An, J., & Cho, S. (2015). Variational Autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE, 2(1), 1–18.*
3. *Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407.*
4. *Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58.*
5. *Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE, 11(4), e0152173.*
6. *Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.*
7. *Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. Proceedings of the 2008 IEEE International Conference on Data Mining, 413–422.*
8. *Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443–1471.*
9. *Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5), 363–387.*