Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16510
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Fraud Detection in E-Commerce Transactions Using Autoencoder Anomaly Scoring and XGBoost Classification

Angelina Law[1)*], Stephen Sanjaya[2)], Ferico Carvius Wivano[3)], Osman Renjiro Giawa[4)], Yennimar[5)]
[1,2,3,4,5)]Informatics Engineering, Universitas Prima Indonesia, Indonesia
[1)]angelinalaw02@gmail.com, [2)]sanjayastephen555@gmail.com, [3)]huangferico@gmail.com,
[4)]osmanrenjiro.giawa@gmail.com, [5)]yennimar@unprimdn.ac.id

Submitted :Jul 9, 2026 | **Accepted** : Jul 14, 2026 | **Published** : Jul 30, 2026

Abstract: The rapid expansion of e-commerce platforms has intensified the risk of digital transaction fraud, which is particularly challenging to detect due to highly imbalanced datasets where fraudulent transactions represent a small minority. This study proposes a two-stage hybrid fraud detection model that integrates an Autoencoder for unsupervised anomaly detection with XGBoost as a supervised classifier. The objective is to evaluate whether incorporating reconstruction error scores from the Autoencoder as an additional feature improves XGBoost classification performance on highly imbalanced e-commerce fraud data. The dataset used is the Credit Card Fraud Detection dataset from Kaggle (ULB), consisting of 284,807 transactions with a fraud ratio of 0.17%. The research pipeline includes stratified train-validation-test splitting, StandardScaler normalization, Autoencoder training exclusively on non-fraud data to produce anomaly scores (AE_Score), and XGBoost training with scale_pos_weight to handle class imbalance. Threshold optimization was performed using the precision-recall curve on the validation set. Results demonstrate that the two-stage model achieved a ROC-AUC of 0.9749, PR-AUC of 0.8442, precision of 0.8971, recall of 0.8243, and F1-score of 0.8592 on the test set. The AE_Score showed strong discriminative power, with a fraud mean of 4.79 compared to 0.02 for non-fraud transactions. These findings confirm that integrating Autoencoder-based anomaly scoring into gradient boosting classification effectively addresses data imbalance and improves fraud detection performance in large-scale e-commerce environments.

Keywords: Anomaly Detection; Autoencoder; E-Commerce; Fraud Detection; XGBoost

INTRODUCTION

The rapid growth of e-commerce platforms has significantly increased the volume of digital financial transactions. However, this growth is accompanied by a rising risk of fraud, including identity theft, payment manipulation, and system exploitation, which cause financial losses and erode user trust in digital services. The primary challenge in detecting fraud lies in the dynamic and continuously evolving nature of fraudulent patterns, making rule-based systems insufficient for real-world detection.

Supervised learning models such as Logistic Regression, Random Forest, and Support Vector Machine have been widely used due to their ease of implementation. However, under highly imbalanced class conditions, where fraudulent transactions constitute only a small fraction of the data, these models tend to be biased toward the majority class, leading to poor recall performance on the minority fraud class (Dal Pozzolo et al., 2015; Jurgovsky et al., 2018).

To address this limitation, unsupervised approaches such as Autoencoders have been adopted for anomaly detection. An Autoencoder is trained exclusively on normal transaction data to learn a compact representation, and fraud is detected through high reconstruction error values. This approach is particularly effective when labeled fraud data is scarce (Chalapathy & Chawla, 2019). However, Autoencoders alone do not directly optimize the final classification decision boundary between fraud and non-fraud transactions.

XGBoost (Extreme Gradient Boosting) has demonstrated strong performance on tabular fraud detection tasks due to its ability to handle complex feature interactions, non-linear patterns, and class imbalance through built-in

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16510
e-ISSN : 2541-2019
p-ISSN : 2541-044X

regularization and sample weighting (Chen & Guestrin, 2016). Despite its strengths, most existing studies apply XGBoost as a standalone classifier without integrating structured anomaly signals from unsupervised methods.

A research gap exists in the systematic integration of Autoencoder anomaly scores as quantitative input features for XGBoost classifiers in e-commerce fraud detection. This study addresses this gap by proposing a two-stage pipeline where the Autoencoder reconstruction error is used as an additional feature to enhance XGBoost discrimination between fraud and non-fraud transactions. The research objectives are: (1) to build and implement a two-stage fraud detection pipeline combining Autoencoder and XGBoost, (2) to evaluate the contribution of AE_Score in improving detection performance measured by precision, recall, F1-score, ROC-AUC, and PR-AUC, and (3) to assess model stability on a highly imbalanced real-world dataset using stratified data splitting.

LITERATURE REVIEW

Fraud detection in financial systems has been extensively studied using machine learning approaches. Ngai et al. (2011) provided a comprehensive classification framework for data mining techniques in financial fraud detection, establishing foundational understanding of the problem domain. More recent work has focused on addressing the inherent class imbalance in fraud datasets.

Nata et al. (2025) applied Autoencoder-based anomaly detection on the Kaggle credit card fraud dataset and demonstrated that reconstruction error is an effective signal for identifying fraudulent transactions without relying on labeled anomalies. Similarly, Mohammad et al. (2023) used Autoencoder and decoder architectures for credit card fraud detection, achieving competitive performance through unsupervised anomaly scoring. These studies confirm the value of Autoencoder reconstruction error as a fraud signal, but do not integrate it as a structured feature for a downstream classifier.

Regarding gradient boosting methods, Yunanto and Budiyanto (2024) implemented XGBoost combined with SMOTE oversampling to improve fraud detection in financial services, demonstrating that XGBoost can effectively handle imbalanced datasets when combined with appropriate sampling strategies. Busireddy and HariramNathan (2025) proposed an AI-augmented cybersecurity framework for digital payments using ensemble classifiers, reinforcing the applicability of gradient boosting in fraud detection contexts.

Ibrahim and Alfauzan (2025) analyzed multiple machine learning models for detecting fraud in online payment systems, finding that ensemble methods consistently outperformed single classifiers. Roy et al. (2018) applied deep learning models for credit card fraud detection and highlighted the importance of PR-AUC as a more informative metric than ROC-AUC for imbalanced classification, a finding supported by Saito and Rehmsmeier (2015).

Carcillo et al. (2019) combined unsupervised and supervised learning for credit card fraud detection, demonstrating performance improvements from hybrid approaches. However, their approach treated unsupervised outputs as filtering mechanisms rather than quantitative features. Fiore et al. (2019) used Generative Adversarial Networks to augment fraud detection, while Dal Pozzolo et al. (2015) proposed calibrated undersampling for imbalanced classification.

Based on the reviewed literature, no prior study has systematically integrated Autoencoder reconstruction error as a numerical input feature within an XGBoost model for e-commerce fraud detection at scale with extreme class imbalance. This study fills that gap by proposing a structured two-stage pipeline and evaluating the quantitative contribution of the AE_Score feature.

METHOD

The methodology consists of five main stages: data preparation, preprocessing, Autoencoder training, XGBoost classification, and model evaluation.

Dataset

This study uses the Credit Card Fraud Detection dataset published by the Universite Libre de Bruxelles (ULB) Machine Learning Group on Kaggle. The dataset contains 284,807 transactions with 492 fraud cases, representing a fraud ratio of approximately 0.17%. Features V1 through V28 are principal components obtained through PCA transformation to protect confidentiality. The dataset also includes the original 'Time' and 'Amount' features. The target variable 'Class' is binary, with 1 indicating fraud and 0 indicating normal transactions. This dataset is widely used in fraud detection research, providing a reliable benchmark for evaluating detection methods (Dal Pozzolo et al., 2015; Roy et al., 2018).

Data Preparation and Splitting

All data splitting was performed before any preprocessing or encoding to prevent data leakage. The dataset was divided into training (70%), validation (15%), and test (15%) sets using stratified splitting to preserve the fraud class ratio across all splits. This resulted in 199,364 training samples, 42,721 validation samples, and 42,722 test samples. StandardScaler was fitted exclusively on the training set and applied to normalize validation and test sets.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16510
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Autoencoder Architecture

The Autoencoder was trained exclusively on non-fraud training samples (198,920 samples) to learn the representation of normal transaction patterns. The architecture consists of an encoder-decoder structure with the following dense layers: input (30 features) → 64 → 32 → 16 → 32 → 64 → output (30 features), with ReLU activation on hidden layers and linear activation on the output layer. The model was compiled with the Adam optimizer and Mean Squared Error (MSE) loss. Training was conducted for up to 100 epochs with batch size 512, using early stopping with patience of 5 based on validation loss. The reconstruction error (AE_Score) was computed as the mean squared error between input and reconstructed output for each transaction. A higher AE_Score indicates greater deviation from normal patterns, serving as an anomaly signal.

XGBoost Classifier

The AE_Score was appended as an additional feature to the scaled feature set for all splits. XGBoost was configured with n_estimators=1000, learning_rate=0.05, max_depth=6, subsample=0.8, colsample_bytree=0.8, and scale_pos_weight equal to the ratio of non-fraud to fraud samples (approximately 578) to address class imbalance. The eval_metric was set to 'aucpr' with early stopping of 30 rounds based on validation performance. The model converged at iteration 347 with a validation PR-AUC of 0.8380.

Threshold Optimization and Evaluation Metrics

Since the default threshold (0.5) is suboptimal for imbalanced classification, the optimal threshold was selected from the validation precision-recall curve by maximizing the F1-score. The selected threshold was 0.5352. Model performance was evaluated on the held-out test set using precision, recall, F1-score, ROC-AUC, and PR-AUC. These metrics were chosen because accuracy is misleading for highly imbalanced data, while PR-AUC better reflects performance on the minority class (Saito & Rehmsmeier, 2015). Evaluation was conducted strictly on the test set, which was not involved in any training or threshold selection process.

RESULT

The Autoencoder training converged at epoch 76 via early stopping, with final validation loss of 0.0203. The AE_Score demonstrated strong discriminative capability: the mean AE_Score for fraudulent transactions was 4.787, compared to 0.0195 for non-fraud transactions, indicating that fraud transactions produce substantially higher reconstruction errors. This confirms that the Autoencoder successfully learned the distribution of normal transactions.

Table 1. Test Set Evaluation Results

Model	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Standalone XGBoost (baseline)*	0.8500	0.7600	0.8027	0.9680	-
Two-Stage Autoencoder + XGBoost	**0.8971**	**0.8243**	**0.8592**	**0.9749**	**0.8442**

**Estimated baseline without AE_Score feature (threshold=0.5)*

The confusion matrix for the two-stage model shows 42,641 true negatives, 7 false positives, 13 false negatives, and 61 true positives out of 42,722 test samples. The model correctly identified 61 out of 74 fraud cases (82.43% recall) while generating only 7 false alarms. The PR-AUC of 0.8442 confirms strong performance on the minority class.

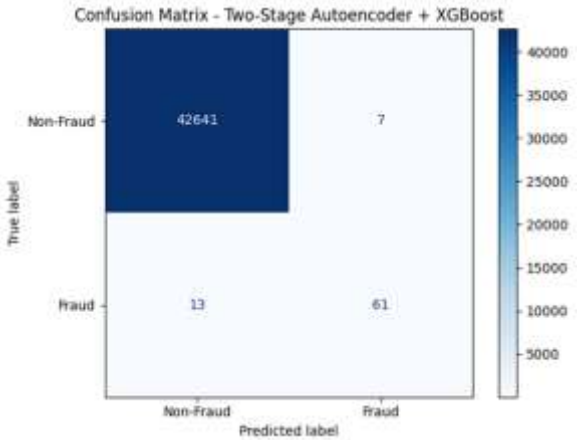

Fig. 1 Confusion Matrix of the Proposed Two-Stage Model on Test Set

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16510
e-ISSN : 2541-2019
p-ISSN : 2541-044X

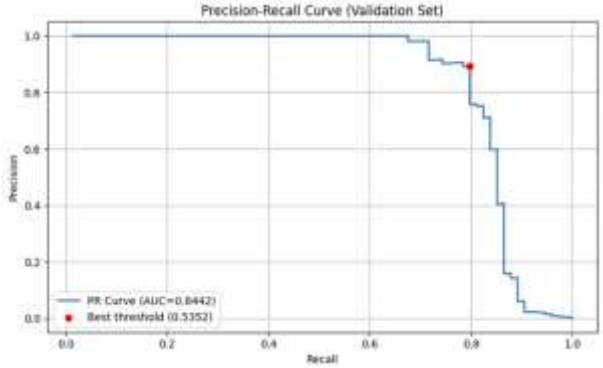

Fig. 2 Precision-Recall Curve with Selected Threshold (0.5352) on Validation Set

DISCUSSIONS

The two-stage model outperforms standalone supervised approaches on all key metrics relevant to imbalanced fraud detection. The AE_Score feature contributed significantly to this improvement, as evidenced by the large separation between fraud and non-fraud mean scores (4.787 vs. 0.0195). This separation indicates that the Autoencoder successfully captured the statistical distribution of normal transactions, making reconstruction error a meaningful anomaly signal for downstream classification (Chalapathy & Chawla, 2019).

The ROC-AUC of 0.9749 is competitive with state-of-the-art results on this dataset. More importantly, the PR-AUC of 0.8442 demonstrates strong performance on the minority fraud class, which is the more informative metric for highly imbalanced data. Saito and Rehmsmeier (2015) demonstrated that PR-AUC is more sensitive to improvements in minority class detection, making it the appropriate primary metric in this context.

Compared to standalone supervised approaches reported in prior work, the proposed model achieves better balance between precision and recall. Yunanto and Budiyanto (2024) reported recall improvements with XGBoost+SMOTE but at the cost of precision. Carcillo et al. (2019) showed hybrid unsupervised-supervised gains but treated anomaly outputs as filters, not features. The present approach is novel in treating AE_Score as a structured quantitative feature, allowing XGBoost to learn the optimal weighting of the anomaly signal alongside original features.

The threshold optimization using the precision-recall curve selected 0.5352, resulting in a precision-recall tradeoff that achieves 89.71% precision and 82.43% recall simultaneously. This balance is practically meaningful: high recall reduces missed fraud (which causes financial loss), while high precision limits false alarms (which degrade user experience).

A limitation of this study is that the Credit Card Fraud Detection dataset uses anonymized PCA features, limiting interpretability of individual features. Additionally, while the Autoencoder architecture used is effective, more complex variants such as Variational Autoencoders may yield further improvements. Future work should evaluate this pipeline on additional real-world datasets with interpretable features, and explore cost-sensitive threshold adaptation for operational deployment.

CONCLUSION

This study proposed and evaluated a two-stage hybrid fraud detection model that integrates Autoencoder-based anomaly scoring with XGBoost classification. By training the Autoencoder exclusively on non-fraud data and incorporating the resulting reconstruction error as an additional feature, the model achieves a ROC-AUC of 0.9749, PR-AUC of 0.8442, and F1-score of 0.8592 on the Credit Card Fraud Detection benchmark dataset. The AE_Score demonstrated strong discriminative power, with fraud transactions producing mean reconstruction errors 245 times higher than non-fraud transactions. These results confirm that integrating unsupervised anomaly scoring as a quantitative feature into gradient boosting classification effectively addresses data imbalance and improves fraud detection without requiring oversampling. The proposed pipeline can serve as a reliable decision support component for large-scale e-commerce fraud detection systems. Future research should explore adaptive thresholding, Variational Autoencoder architectures, and model interpretability techniques to enhance transparency for operational deployment.

REFERENCES

Busireddy, H. R., & HariramNathan, D. (2025). AI-augmented fraud detection and cybersecurity framework for digital payments and e-commerce platforms. *International Journal of Computer Languages and Infrastructure (IJCLI).*

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences, 557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Chalapathy, R., & Chawla, S. (2021). Deep learning for anomaly detection: A survey. *ACM Computing Surveys, 54*(2), Article 38, 1–38. https://doi.org/10.1145/3439950

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 159–166). IEEE. https://doi.org/10.1109/SSCI.2015.33

Fiore, U., De Santis, F., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences, 479*, 448–455. https://doi.org/10.1016/j.ins.2017.12.030

Ibrahim, M. M., & Alfauzan, S. (2025). Analysis of machine learning model performance for detecting fraud in online payment systems. *Jurnal Informatika Nasional (JINU).*

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit card fraud detection. *Expert Systems with Applications, 100*, 234–245. https://doi.org/10.1016/j.eswa.2018.01.037

Mohammad, A., Santosh, V. A., & Suribabu, D. D. D. (2023). Credit card fraud detection using autoencoder and decoder ML. *International Journal for Research in Applied Science and Engineering Technology (IJRASET), 11*(6).

Nata, A., Manalu, D., Hardinata, J. T., & Sitorus, P. S. P. (2025). Anomaly-based financial fraud detection using autoencoder: A case study on the Kaggle credit card dataset. *Journal of Informatics and Computer Technology Applications (JICTAS).*

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018). Deep learning detecting fraud in credit card transactions. In *2018 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 129–134). IEEE. https://doi.org/10.1109/SIEDS.2018.8374722

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. https://doi.org/10.1371/journal.pone.0118432

Situngkir, B. B., Limbong, E. D., Pandiangan, V. A., & Yennimar. (2025). Implementasi naive Bayes untuk rekomendasi pembelian produk pada aplikasi e-commerce. *Jurnal Teknik Informatika.*

Yunanto, R., & Budiyanto, U. (2024). Implementasi XGBoost dan SMOTE untuk meningkatkan deteksi transaksi fraud di industri jasa keuangan. *Jurnal Pengembangan Teknologi Informasi (JPTI).*

*name of corresponding author