Weighted Ensemble Outlier Detection for Anti-Money Laundering

Anagha Srikrishna[1], Akshaya Srikrishna[1], Abhi Lavanya[1], Krishnendu M R[1], and S. Chitrakala[1]

Dept. of Computer Science and Engineering,
College of Engineering Guindy, Anna University, Chennai, India
2022103066@student.annauniv.edu, 2022103065@student.annauniv.edu,
2022103068@student.annauniv.edu, 2022103081@student.annauniv.edu,
chitras@annauniv.edu

Abstract. Money laundering poses a major challenge to financial integrity as illicit transactions often evade traditional Anti-Money Laundering (AML) systems. Rule-based methods typically generate excessive false positives, while static machine learning models struggle to capture complex and evolving patterns in transaction networks. This work proposes a Performance-Weighted Ensemble for Anti-Money Laundering Detection (PWED-AML), a framework that integrates four complementary algorithms: statistical (Z-Score with Random Forest), density-based (Local Outlier Factor), distance-based (Isolation Forest), and graph-based (Graph Neural Networks). Its novelty lies in uniting interpretable outlier-detection approaches with Graph Neural Networks (GNNs), enabling the system to capture both global anomalies and relational patterns in transaction networks. Each algorithm's contribution is weighted according to sensitivity, specificity, and ROC-AUC, producing an ensemble that balances diverse detection capabilities. Experiments on the IBM AML dataset demonstrate that individual models vary in effectiveness, with Graph Neural Networks achieving the strongest performance (92% sensitivity, 81% specificity). By leveraging weighted contributions, PWED-AML produces risk scores that reduce false positives while preserving detection accuracy. The proposed PWED-AML framework provides financial institutions with a scalable and robust approach to next-generation AML monitoring.

Keywords: anti-money laundering · outlier detection · graph neural networks · anomaly detection · financial fraud detection · machine learning · risk assessment

1 Introduction

Money laundering is a persistent and complex threat to the global financial system, enabling the concealment of illicit funds and undermining the stability of economic and regulatory structures. Estimates from the United Nations Office on Drugs and Crime suggest that between $800 billion and $2 trillion is

laundered annually, representing 2–5% of global GDP [9]. The laundering process—placement, layering, and integration—is deliberately designed to obscure the origin of funds, making detection increasingly difficult in large-scale, high-velocity digital transaction environments.

Conventional Anti-Money Laundering (AML) systems rely heavily on static rules and threshold-based monitoring. While these systems provide a baseline level of protection, they face significant limitations: high false positive rates, difficulty adapting to new laundering strategies, and an inability to capture subtle or relational patterns across complex transaction networks. These shortcomings impose substantial operational burdens on financial institutions and highlight the need for adaptive, data-driven methods that can identify diverse manifestations of laundering activity.

In this work, we propose Performance-Weighted Ensemble for AML Detection (PWED-AML) for outlier detection that integrates complementary analytical perspectives. The framework combines: Z-Score with Random Forest for identifying global statistical deviations, Local Outlier Factor (LOF) for density-based irregularities, Isolation Forest for isolating sparse anomalies, and Graph Neural Networks (GNNs) for capturing relational behaviors across accounts and transaction flows. Each model's contribution is weighted according to sensitivity, specificity, and ROC-AUC performance, ensuring that stronger detectors have proportionally higher influence in the ensemble.

Experimental evaluation on the IBM AML dataset reveals significant performance variations among individual algorithms, with Graph Neural Networks achieving the strongest metrics (90.3% sensitivity, 84.7% specificity). The ensemble generates risk scores for transaction classification while maintaining low false positive rates. This performance-weighted approach represents a significant advancement in AML technology, providing financial institutions with an adaptive, scalable solution for next-generation transaction monitoring systems capable of evolving with emerging laundering strategies.

2 Related Work

Research on outlier detection for Anti-Money Laundering (AML) spans multiple methodological paradigms, including graph analysis, ensemble learning, deep unsupervised models, and domain-specific data-mining approaches. Transactional data remains the primary input for these studies, but methods differ in the anomaly types they target—ranging from isolated value outliers and local-density anomalies to relational and subgraph-level irregularities.

Graph- and network-based analyses have been employed to expose structural anomalies in transaction flows. For example, network-analysis methods applied to cross-country wire transfers highlight abrupt changes in node roles as potential indicators of illicit behavior [10]. Systematic reviews of graph anomaly detection further synthesize deep-learning approaches for nodes, edges, subgraphs, and whole-graph anomalies, underscoring both opportunities and challenges in applying these methods to transactional networks [4].

Deep and hybrid unsupervised models have also been investigated for detecting complex transaction patterns. Unsupervised architectures—such as hybrid convolutional–recurrent models—have shown promise in cross-border transaction anomaly detection, outperforming simpler baselines in certain contexts [13]. In parallel, graph neural networks (GNNs) have been studied for their ability to capture relational laundering behaviors that tabular models often miss [4, 10].

Classical data-mining and machine-learning techniques continue to play an important role in AML detection. Studies comparing decision trees, neural networks, and clustering methods highlight the value of feature engineering and scalable pattern discovery in large transaction datasets [7]. Supervised and semi-supervised adaptations of clustering and classification (e.g., supervised K-means variants) have also been evaluated for transaction-level fraud detection, particularly when labeled data are available [6]. Domain-specific approaches extend these ideas; for instance, multi-granularity representation and information-set concepts have been used to better characterize data distributions and identify abnormal points [14, 11].

Ensemble and hybrid approaches are widely recognized for improving robustness in anomaly detection. Several studies demonstrate that combining heterogeneous detectors can mitigate individual weaknesses and increase coverage, although deployment remains challenged by computational cost and integration complexity [5, 4]. In blockchain-specific contexts, comparative studies find that tree-based boosters often achieve strong performance, though outcomes depend heavily on feature design and dataset characteristics [2]. However, most existing ensemble methods rely on static weighting schemes that fail to account for varying detector performance across data types or temporal contexts.

To make trade-offs explicit: statistical profiling is low-cost and captures global deviations but struggles with subtle or relational anomalies; density- and distance-based detectors uncover local and sparse outliers but are sensitive to feature scaling and neighborhood selection; and GNNs excel at network-level patterns but incur greater computational and labeling costs. Ensembles broaden anomaly coverage, but static weighting often limits adaptability in real-world AML monitoring.

Beyond financial fraud, anomaly-detection techniques have been applied in domains such as government auditing and cybersecurity, showing cross-domain utility of density, time-series, and graph-based methods [11, 8]. Other applications emphasize forecasting and risk assessment (e.g., market trend prediction and e-commerce risk), demonstrating how model selection and preprocessing materially affect performance [3, 1].

Despite progress, persistent gaps remain. Graph-based and deep-learning methods provide structural insight but are resource-intensive and often depend on labeled data or negative-sampling strategies. Ensemble methods improve coverage but introduce deployment overhead. Domain-specific approaches can lack generality, while explainability remains a pressing concern—particularly in regulated AML environments where auditability and interpretability are essential.

Moreover, existing ensembles lack adaptive mechanisms to dynamically adjust detector contributions in response to real-time performance metrics.

Collectively, this motivates a unified framework that leverages complementary detectors while addressing operational constraints. The framework we propose integrates statistical profiling (Z-Score with Random Forest), density-based (LOF), distance-based (Isolation Forest), and graph-based (GNN) perspectives. Each detector adds distinct value: statistical profiling highlights extreme deviations, density- and distance-based methods uncover local and global anomalies, and graph models capture relational and structural irregularities.

Unlike conventional ensembles that assign fixed or arbitrary weights, our approach employs a performance-weighted scheme in which detector contributions are determined by sensitivity, specificity, and ROC-AUC from individual model evaluation. This ensures that strong performers—such as the GNN in our experiments—receive proportionally greater influence, while complementary methods provide broader anomaly coverage. The resulting framework enhances robustness across diverse transaction patterns and improves practical suitability for AML deployment in large-scale financial environments.

3 Proposed Methodology

The proposed PWED-AML follows a systematic pipeline consisting of preprocessing, anomaly detector integration, performance-based weighting, and ensemble risk scoring. Each step is carefully designed to handle the unique challenges of Anti-Money Laundering (AML) detection, such as high class imbalance, heterogeneity of transaction data, and the subtle nature of laundering patterns. The overall framework is illustrated in Figure 1.

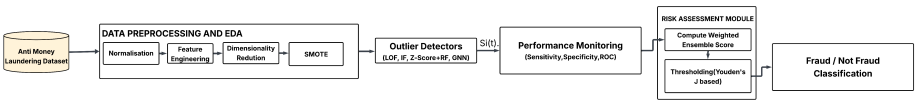


Fig. 1: Overall architecture of the proposed PWED-AML framework

3.1 Data Preprocessing

Prior to modeling, preprocessing was performed to ensure data quality and improve downstream detection:

- **Missing values** were imputed using feature-wise medians.
- **Categorical variables** (e.g., account type, transaction channel) were one-hot encoded.

- **Normalization:** All continuous features were scaled to zero mean and unit variance:
$$x' = \frac{x - \mu}{\sigma} \tag{1}$$
where μ and σ are the feature mean and standard deviation.
- **Feature engineering:** Domain-specific features such as transaction velocity, aggregated incoming/outgoing values, and frequency ratios were derived.
- **Dimensionality reduction:** Principal Component Analysis (PCA) was applied to reduce redundant correlations and retain components explaining 95% of variance.

Since money laundering transactions are rare compared to legitimate ones, the dataset exhibited severe class imbalance. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic minority samples and reducing model bias toward the majority class.

3.2 Anomaly Detectors

To capture different manifestations of suspicious behavior, four complementary detectors were employed. Each detector focuses on a unique perspective of the data, ensuring robustness against diverse laundering strategies.

1. **Z-Score with Random Forest:** Transactions with high Z-score deviations in monetary values are flagged as anomalies and then classified using a Random Forest, which leverages contextual patterns such as account activity and feature interactions.
2. **Local Outlier Factor (LOF):** LOF is a density-based method that detects points in sparse regions compared to their neighbors, making it effective for exposing small collusive laundering networks that appear normal globally but abnormal locally.
3. **Isolation Forest (IF):** IF partitions the data recursively, with anomalies requiring fewer splits for isolation. Its efficiency and scalability make it well-suited for detecting rare behaviors across millions of IBM transactions.
4. **Graph Neural Networks:** Graph Neural Networks (GNNs) model accounts and transactions as a graph, aggregating information from neighboring transactions to capture structural laundering patterns, such as circular transfers and account clustering, that vector-based methods miss. This allows the network to focus on suspicious transaction flows between accounts.

3.3 Ensemble Strategy and Weighting

The weight w_i for each detector i is computed as a performance-based combination of sensitivity, specificity, and AUC-ROC:

$$w_i = \frac{\alpha \cdot \text{Sensitivity}_i + \beta \cdot \text{Specificity}_i + \gamma \cdot \text{AUC}_i}{\sum_{j=1}^{k} (\alpha \cdot \text{Sensitivity}_j + \beta \cdot \text{Specificity}_j + \gamma \cdot \text{AUC}_j)} \tag{2}$$

where α, β, γ are hyperparameters controlling the relative importance of each metric. Based on validation set optimization, the selected values were $\alpha = 0.5$, $\beta = 0.2$, and $\gamma = 0.3$, giving higher priority to sensitivity while maintaining specificity and overall detector reliability.

3.4 Threshold Selection using Youden's J

To transform the continuous ensemble risk scores into binary suspicious/legitimate labels, we employ *Youden's J statistic* [12], a robust criterion widely used in diagnostic testing. The statistic is defined as:

$$J = \text{Sensitivity} + \text{Specificity} - 1 \tag{3}$$

Maximizing J identifies the threshold θ^* that jointly optimizes sensitivity (true positive rate) and specificity (true negative rate):

$$\theta^* = \arg\max_{\theta} \Big(\text{Sensitivity}(\theta) + \text{Specificity}(\theta) - 1\Big) \tag{4}$$

3.5 Proposed Algorithmic Workflow

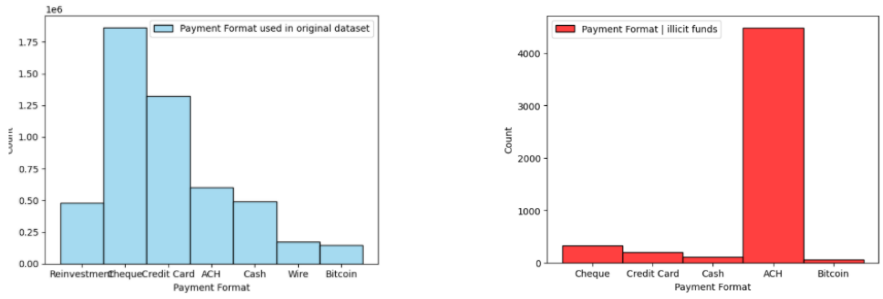
The proposed PWED-AML methodology for fraud detection is outlined in Algorithm 1, which integrates multiple detectors through performance-weighted ensembling.

Algorithm 1: Performance-Weighted Ensemble for AML Detection

Input: IBM AML dataset D
Output: Fraud/Not Fraud Classification
Split D into training, validation, and test sets;
Preprocess data (handle missing values, encode features, normalize, feature engineer, apply PCA);
Apply SMOTE **only to the training set** to balance classes;
foreach *detector* $i \in$ {*Z-Score+RF, LOF, IF, GNN*} **do**
    Train detector i on training set;
    Compute anomaly scores $S_i(t)$ for all transactions t;
    Evaluate Sens_i, Spec_i, AUC_i on validation set;
Compute weights w_i using Eq. 2;
foreach *transaction* $t \in D$ **do**
    Compute ensemble risk score: $R(t) = \sum_i w_i \cdot S_i(t)$;
    if $R(t) > \theta$ *(chosen using Youden's J)* **then**
        Flag t as suspicious;

4 Experimental Results and Analysis

This study employs the IBM Anti-Money Laundering (AML) dataset containing **5,078,345 transactions** across multiple accounts, banks, and currencies. Each record is labeled as legitimate or laundering, with **5,073,159 (99.9%)** legitimate and only **5,177 (0.1%)** illicit, demonstrating the extreme class imbalance typical in AML tasks. This imbalance is further reflected in the distribution of **payment formats**, where **ACH transactions—though relatively infrequent overall—dominate the illicit subset** (Figure 2), highlighting both systemic skew and channel-specific vulnerability.



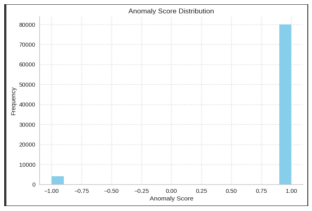
(a) All Transactions (b) Illicit Transactions Only

Fig. 2: Distribution of payment formats in the IBM AML dataset, with ACH transactions disproportionately represented in illicit cases

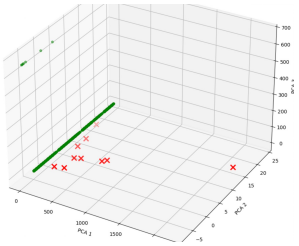
Table 1 summarizes the key features of the dataset.

Table 1: IBM AML Dataset Features

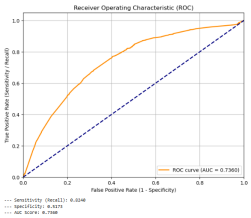
No.	Feature	Description
1	Timestamp	Transaction timestamp
2	From Bank	Originating bank identifier
3	From Account	Originating account
4	To Bank	Receiving bank identifier
5	To Account	Receiving account
6	Amount Received	Amount credited
7	Receiving Currency	Currency of credit
8	Amount Paid	Amount debited
9	Payment Currency	Currency of debit
10	Payment Format	Mode of transfer
11	Is Laundering	Label: 1 = laundering, 0 = legitimate



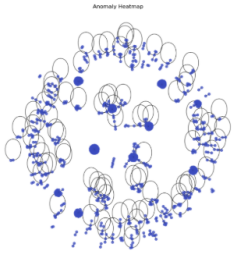
(a) LOF Anomaly Scores



(b) Isolation Forest



(c) Z-Score + RF ROC Curve



(d) GNN Anomalies

Fig. 3: Visualization of anomalies captured by each detection method

Model performance was evaluated using Sensitivity, Specificity, and AUC, standard metrics in fraud and anomaly detection tasks where class imbalance and false negatives are critical. Sensitivity (Eq. 5) measures the ability to correctly identify illicit transactions, while Specificity (Eq. 6) quantifies the ability to reject legitimate ones. AUC (Eq. 7) summarizes the overall discriminative capability of the model. The anomaly patterns captured by each detector are visualized in Figure 3.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR \tag{7}$$

As shown in Table 2, GNN achieved the highest sensitivity (92%) and strong specificity (81%), effectively capturing network-level laundering patterns. Z-Score + RF offered a balanced performance, combining statistical interpretability with machine learning refinement. Isolation Forest achieved high sensitivity but lower specificity, while LOF was more effective in detecting local neighborhood anomalies.

Table 2: Comprehensive Algorithm Performance Metrics

Algorithm	**Sensitivity**	**Specificity**	**AUC**
LOF	0.62	0.70	0.66
Isolation Forest	0.84	0.57	0.70
GNN	0.92	0.81	0.87
Z-Score + RF	0.84	0.68	0.76

Weights for each detector were computed using the performance-based formula (Eq. 2) with $\alpha = 0.5$, $\beta = 0.2$, and $\gamma = 0.3$. Several weight combinations were explored during validation, and one representative setting based on the metrics in Table 2 is shown below:

$$w_{\text{LOF}} = 0.5(0.62) + 0.2(0.70) + 0.3(0.66) = 0.648 \tag{8}$$

$$w_{\text{IF}} = 0.5(0.84) + 0.2(0.57) + 0.3(0.70) = 0.744 \tag{9}$$

$$w_{\text{Z+RF}} = 0.5(0.84) + 0.2(0.68) + 0.3(0.76) = 0.784 \tag{10}$$

$$w_{\text{GNN}} = 0.5(0.92) + 0.2(0.81) + 0.3(0.87) = 0.883 \tag{11}$$

After normalization:

$$w_{\text{LOF}} \approx 0.21, \quad w_{\text{IF}} \approx 0.24, \quad w_{\text{Z+RF}} \approx 0.26, \quad w_{\text{GNN}} \approx 0.29$$

Consider a transaction with scores

$$s_{\text{LOF}} = 0.3, \quad s_{\text{IF}} = 0.8, \quad s_{\text{Z+RF}} = 0.5, \quad s_{\text{GNN}} = 0.9.$$

The ensemble score is then

$$R(t) = 0.21(0.3) + 0.24(0.8) + 0.26(0.5) + 0.29(0.9) \approx 0.66.$$

This score is compared against the threshold chosen via Youden's J statistic (Eq. 3). In this case, $\theta^* = 0.65$ was selected on the validation set. Since $R(t) = 0.66 > \theta^*$, the transaction is flagged as suspicious.

This calculation illustrates the weighting and thresholding procedure; the same methodology was applied systematically across the full dataset to obtain the final classification results.

4.1 Conclusion and Future Work

This work introduces PWED-AML, a performance-guided ensemble for Anti-Money Laundering detection that combines individual detector outputs using validation-driven weights. Its novelty lies in uniting interpretable outlier-detection approaches with Graph Neural Networks (GNNs), enabling the system to capture both global anomalies and relational patterns in transaction networks. Experimental results showed that while GNNs achieved the strongest individual

performance, the ensemble delivered more balanced and robust detection outcomes. Nonetheless, limitations remain, including challenges in scaling to very large graphs and a dependence on validation AUC, which may not fully capture operational trade-offs in real-world deployments.

Future research will focus on improving the scalability of GNN models through sampling and distributed training, developing real-time detection pipelines for high-throughput environments, and exploring sequential transaction modeling using transformer-based architectures. Privacy-preserving approaches such as federated learning and more advanced ensemble strategies are also promising directions to enhance both effectiveness and compliance in AML applications.

References

1. Chen, X., Wang, B.: Risk assessment of processing cross border ecommerce using improved radial basis function kernel-support vector machine in data mining. In: 2024 Second International Conference on Data Science and Information System (ICDSIS). pp. 1–4 (2024). https://doi.org/10.1109/ICDSIS61070.2024.10594263
2. El-Attar, N.E., Salama, M.H., Abdelfattah, M., Taha, S.: A comparative analysis for anomaly detection in blockchain networks using machine learning techniques. In: 2024 34th International Conference on Computer Theory and Applications (ICCTA). pp. 171–176 (2024). https://doi.org/10.1109/ICCTA64612.2024.10974876
3. Hosseini, S.A., Niccolai, A., Lorenzo, M., Casamatta, F., Grimaccia, F.: Advanced pattern detection and trend forecasting in european carbon markets using machine learning algorithms. In: 2024 IEEE International Conference on Artificial Intelligence Green Energy (ICAIGE). pp. 1–6 (2024). https://doi.org/10.1109/ICAIGE62696.2024.10776670
4. Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q.Z., Xiong, H., Akoglu, L.: A comprehensive survey on graph anomaly detection with deep learning. IEEE Transactions on Knowledge and Data Engineering **35**(12), 12012–12038 (2023). https://doi.org/10.1109/TKDE.2021.3118815
5. Raj, M., Khan, H., Kathuria, S., Chanti, Y., Sahu, M.: The use of artificial intelligence in anti-money laundering (aml). In: 2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL). pp. 272–277 (2024). https://doi.org/10.1109/ICSADL61749.2024.00050
6. Rakhra, M., Puri, R., Sarkar, T., Maurya, S., Chakrabarty, P., Jairath, K.: Deploying k-mean cluster utilizing ml approach for detecting fraud of financial landscape. In: 2025 3rd International Conference on Disruptive Technologies (ICDT). pp. 900–906 (2025). https://doi.org/10.1109/ICDT63985.2025.10986519
7. Rathore, R., Sharma, Y., Ambika, P., Upadhyay, A.K., Mahajan, S., Kumar, R.: Data mining techniques in financial fraud detection. In: 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N). pp. 1300–1305 (2024). https://doi.org/10.1109/ICAC2N63387.2024.10895091
8. Snehal, M.S.C., Nagoor, V., Rohit, S., Raghunandan, S., Thangavel, S.K., Srinivasan, K., Kapoor, P.: Towards explainability using ml and deep learning models for malware threat detection. In: 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT). pp. 1–6 (2024). https://doi.org/10.1109/AIIoT58432.2024.10574689

9. United Nations Office on Drugs and Crime: Money-laundering and globalization (2025), https://www.unodc.org/unodc/en/money-laundering/overview.html, accessed: 2025-08-29
10. Vilella, S., Capozzi Lupi, A.T.E., Ruffo, G., Fornasiero, M., Moncalvo, D., Ricci, V., Ronchiadin, S.: Exploiting graph metrics to detect anomalies in cross-country money transfer temporal networks. In: Companion Proceedings of the ACM Web Conference 2023. p. 1245–1248. WWW '23 Companion, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3543873.3587602, https://doi.org/10.1145/3543873.3587602
11. Wu, Z., Mao, Y., Fang, C.: Data-driven analytics in government auditing: A first look at real-world cases. In: 2023 10th International Conference on Behavioural and Social Computing (BESC). pp. 1–6 (2023). https://doi.org/10.1109/BESC59560.2023.10386553
12. Youden, W.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
13. Yu, Q., Xu, Z., Ke, Z.: Deep learning for cross-border transaction anomaly detection in anti-money laundering systems. In: 2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). pp. 244–248 (2024). https://doi.org/10.1109/MLBDBI63974.2024.10823769
14. Zhang, P., He, Z., Wang, D., Jiang, T., Li, B., Liu, J., Huang, W., Li, T.: Odmgis: An outlier detection method based on multigranularity information sets. IEEE Transactions on Fuzzy Systems **33**(7), 2050–2061 (2025). https://doi.org/10.1109/TFUZZ.2025.3550749