Vol. 11, No. 1, April 2026, Page. 176-188
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Federated XGBoost for Ethereum AML: Performance Efficiency Trade off in a Multi Exchange Setting

Nazwa Auliarahman[*1], Nur Widiyasono[2] , Aradea[3]
[1,2,3]Informatics study program, Faculty of Engineering, Siliwangi University, Indonesia
Email: [1]227006007@student.unsil.ac.id, [2]nur.widiyasono@unsil.ac.id, [3]aradea@unsil.ac.id

Abstrak

Penelitian ini mengevaluasi efektivitas federated learning dalam mendeteksi alamat ilegal pada blockchain Ethereum untuk Anti Money Laundering (AML). Studi ini membandingkan XGBoost centralized dan federated dalam simulasi multi exchange, data dibagi secara horizontal menjadi 3 subset yang merepresentasikan entitas bursa independen. Untuk mengisolasi pengaruh pelatihan terdistribusi, pembagian data dilakukan secara homogen (IID), sehingga analisis difokuskan pada perbedaan mekanisme pembelajaran tanpa dipengaruhi heterogenitas data. Hasil menunjukkan bahwa model federated mencapai performa yang kompetitif dengan PR AUC 0,9962 dan akurasi 97,11%, dibandingkan model terpusat dengan PR AUC 0,9975 dan akurasi 97,75%. Namun, performa tersebut disertai peningkatan durasi pelatihan 4,01 detik dibandingkan 3,39 detik, yang disebabkan oleh beban komunikasi selama proses pembaruan model. Temuan ini menegaskan adanya trade off antara kinerja dan efisiensi dalam penerapan federated learning. Meskipun mampu mendekati performa pembelajaran centralized, pendekatan ini menimbulkan biaya operasional tambahan. Studi ini merupakan evaluasi dasar dalam kondisi terkontrol dan belum merepresentasikan skenario dunia nyata, sehingga diperlukan penelitian lanjutan pada data non IID dan skala sistem yang lebih besar.

Kata kunci: Anti Money Laundering, Blockchain, Centralized, Federated

Abstract

This study evaluates the effectiveness of federated learning in detecting illegal transactions on the Ethereum blockchain for Anti Money Laundering (AML) purposes. The study compares centralized and federated XGBoost models in a multi-exchange simulation, where the data is horizontally partitioned into three subsets representing independent exchange entities. To isolate the influence of distributed training, data partitioning was performed homogeneously (IID), so that the analysis focused on differences in learning mechanisms without being influenced by data heterogeneity. The results show that the federated model achieves competitive performance with a PR AUC of 0.9962 and an accuracy of 97.11%, compared to the centralized model with a PR AUC of 0.9975 and an accuracy of 97.75%. However, this performance comes with an increase in training duration 4.01 seconds compared to 3.39 seconds, caused by communication overhead during the model update process. These findings confirm the existence of a trade off between performance and efficiency in the application of federated learning. Although it can approach the performance of centralized learning, this approach incurs additional operational costs. This study constitutes a baseline evaluation under controlled conditions and does not yet represent real-world scenarios therefore, further research is needed on non IID data and larger scale systems.

Keywords: *Anti Money Laundering, Blockchain, Centralized, Federated*

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

The quick development of blockchain technology, especially Ethereum, has created new difficulties for financial surveillance, such as the identification of illegal activity like money laundering. Blockchain transactions are visible yet pseudonymous, in contrast to traditional financial systems, making it challenging to link transactions to actual persons [1]. As a result, one of the most important tasks in anti money laundering (AML) activities is to identify hazardous addresses based on transaction behavior.

A number of studies have investigated [2] and analyzed transaction patterns and behavioral characteristics using machine learning algorithms to identify illegal activity on blockchain networks [2]. Because models such as XGBoost are capable of identifying non-linear relationships in tabular data, these models have demonstrated excellent effectiveness in identifying suspicious accounts. However, most current methods rely on centralized datasets, which consolidate all transaction data into a single

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

system for training [2]. In real-world scenarios, this assumption is often impractical, particularly in multi-exchange systems where transaction data is scattered across multiple platforms and cannot be freely exchanged due to organizational, legal, or privacy constraints.

How to cooperatively create efficient detection models without necessitating data centralization is a major difficulty brought about by the dispersed nature of data across several entities [3]. Federated learning, which allows several entities to train a common model without exchanging raw data [4], Instead of sharing data, each participant performs local model training and sends model updates to a central server, where iterative aggregation is performed to construct a global model [5]. This process allows collaborative learning while preserving data locality and privacy, making it particularly suitable for multi-entity systems such as financial institutions and cryptocurrency exchanges.

Despite its potential, the application of federated learning in blockchain-based AML detection remains limited and insufficiently validated. Existing studies on illicit activity detection in Ethereum predominantly rely on centralized learning approaches [2], which assume that all transaction data can be aggregated into a single training environment. However, this assumption is often unrealistic in real-world multi-exchange systems, where data is inherently distributed and subject to privacy, legal, and organizational constraints. As a result, the applicability of centralized models in operational AML settings remains questionable.

Research on federated learning has demonstrated its capability to enable collaborative model training without data sharing [6]; [7]. Nevertheless, these studies are largely conducted on traditional financial datasets or generic tabular data, and do not account for the structural and behavioral characteristics unique to blockchain transactions. Consequently, their findings cannot be directly generalized to Ethereum-based AML detection tasks.

Prior research has primarily focused on predictive performance, while largely neglecting the operational implications of federated learning, such as communication overhead and training efficiency. This creates a critical gap, as federated learning systems inherently introduce additional system-level costs that may affect their practical feasibility. To date, there is a lack of systematic empirical studies that jointly evaluate predictive performance and operational efficiency in federated learning for blockchain AML, particularly under a multi-exchange setting.

To address these limitations, this study conducts a controlled empirical evaluation of federated XGBoost in a simulated multi-exchange environment. By employing a homogeneous (IID) data partitioning scheme, this study isolates the impact of distributed training from data heterogeneity, enabling a baseline understanding of federated learning behavior.

This study contributes by (1) providing a direct and controlled comparison between centralized and federated XGBoost for Ethereum AML detection, and (2) systematically analyzing the trade off between predictive performance and operational efficiency, including training duration and communication overhead. Rather than proposing a new model, this work establishes a foundational benchmark for evaluating the feasibility of federated boosting in blockchain-based AML systems, particularly in multi-entity environments.

2. RESEARCH METHOD

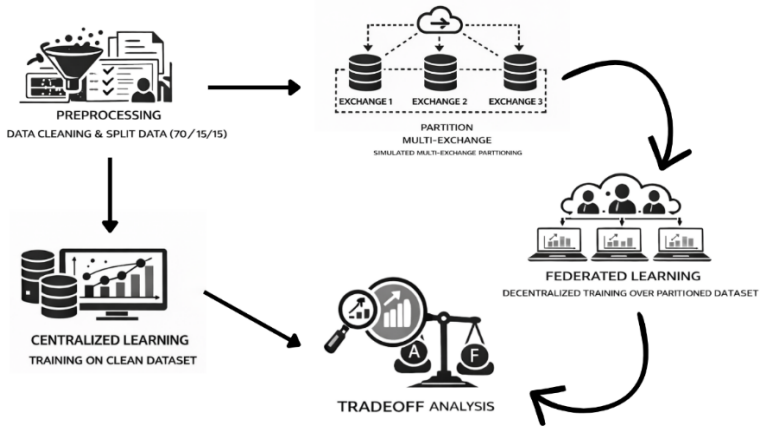


Figure 1. Research Method

Research methodology used in this study is illustrated in Figure 1, which includes preprocessing, multi exchange simulation partitioning, model training using centralized and federated approaches, and trade off analysis. This methodology is designed to ensure a fair comparison between the two learning paradigms within a consistent experimental framework. The following section describes the dataset and the preprocessing steps in detail.

2.1 Preprocessing

The dataset used in this study is derived from the publicly available Ethereum illicit account dataset introduced by Pahuja et al. [2], which has been widely used as a benchmark in blockchain AML research. The dataset consists of approximately 4.153 instances with 50 features, including transaction-based, temporal, and token interaction attributes. The features capture various behavioral patterns such as transaction frequency, value distribution, and ERC20 token interactions, which are relevant for illicit activity detection.

Before training the model, data preprocessing was performed to guarantee data quality and consistency. This procedure involved verifying data types, eliminating unrelated identifiers, addressing missing values, and assessing for possible data leakage. These procedures adhere to typical machine learning protocols to guarantee that the model identifies significant patterns without prejudice caused by data inconsistencies.

The dataset was split into training, validation, and testing sets utilizing a stratified division with a ratio of 70:15:15 to maintain class distribution throughout all subsets. The training set was utilized for model training, the validation set for selecting models and adjusting thresholds, and the test set for the ultimate assessment. The test set was completely separated from the training process to guarantee an impartial evaluation of model performance.

2.2 Multi Exchange Simulation Partition

To emulate a distributed data setting, the training dataset was divided horizontally into several subsets [6], with each subset representing a distinct exchange. In this environment, every exchange contains a segment of the data featuring the same attribute space but varying samples.

The data is partitioned under a homogeneous (IID) assumption, where each subset maintains similar statistical properties [4]. While this assumption does not fully reflect real-world federated learning scenarios, where data is typically non-IID, it is intentionally adopted to isolate the effect of distributed training from data heterogeneity. This controlled setting enables a clearer comparison between centralized and federated learning paradigms, serving as a baseline for future studies under more realistic conditions.

2.3 Centralized XGBoost

The centralized baseline model in this study was implemented using XGBoost, as illustrated in Figure 2. In this setup, all transaction data from multiple exchanges are aggregated into a single dataset, enabling the model to learn from the complete data distribution.

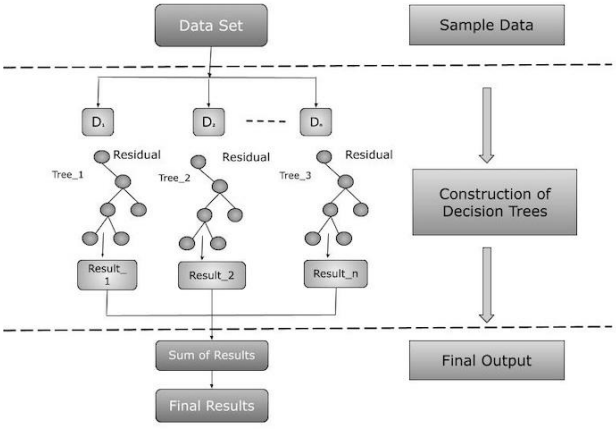


Figure 2. Centralized learning architecture

XGBoost utilizes a gradient boosting paradigm, building the model as a collection of decision trees that are trained one after the other. In every iteration, a new tree is created to represent the residual errors from the prior trees [7]. As illustrated in Figure 2, the training procedure starts with an initial forecast, subsequently involving repetitive tree development, where each tree aims to address the outstanding errors (residuals) from previous iterations.

Every decision tree generates a preliminary output, and the ultimate prediction is derived by combining (summing) the outputs from all trees. This incremental learning approach enables the model to gradually enhance its predictions and identify intricate [8], non linear relationships in the data.

The model is trained using predefined hyperparameters on the combined dataset, without any limitations concerning data division or communication. The XGBoost model is configured with the following hyperparameters number of trees 742 best iteration after early stopping, maximum depth 4, learning rate 0.1, subsample 0.8, and colsample_bytree 0.8. These parameters are selected based on preliminary experiments to balance predictive performance and computational efficiency.

This centralized XGBoost configuration exemplifies a perfect learning environment, since it can directly utilize all accessible data and fully take advantage of global transaction trends. Consequently, it acts as a reference point and upper limit for comparing with the federated learning method.

2.4 Federated XGBoost

The federated learning approach in this study is implemented using a sequential boosting scheme, as illustrated in Figure 3. The architecture consists of a central server that maintains the global XGBoost model and multiple participating exchanges that collaboratively update the model without sharing raw data.

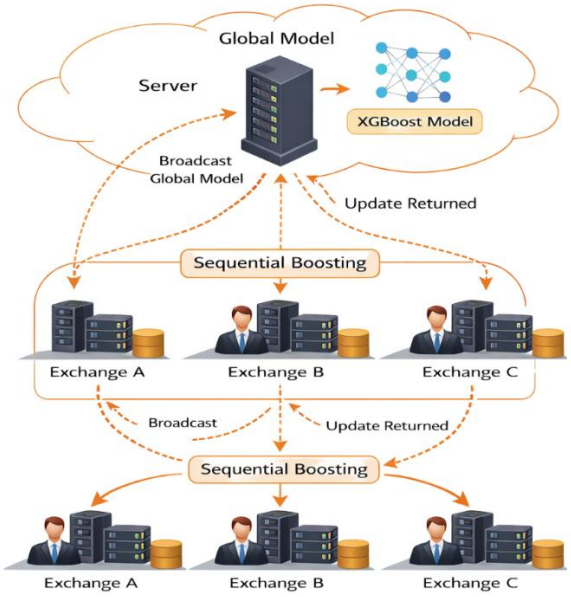


Figure 3. Federated learning architecture

The federated learning approach in this research adheres to the typical framework of distributed model training, where a central server sets up a global model and shares it with several participating exchanges. Every exchange conducts local training with its specific dataset, guaranteeing that sensitive transaction information stays decentralized, which is a standard practice in [9] federated learning frameworks.

This repetitive process persists for a fixed number of federated rounds. In this study, the federated training process is conducted over 10 communication rounds, involving 3 simulated exchanges. In each round, the model is sequentially updated across participating clients before being returned to the central server. This process is repeated until the maximum number of rounds is reached or convergence is observed.

Unlike standard federated learning approaches such as FedAvg, which rely on parallel aggregation of local updates, this study adopts a sequential boosting strategy designed for the experimental context. This approach is chosen to align with the iterative nature of gradient boosting,

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

where model updates are applied progressively. Additionally, the sequential setup allows controlled observation of how model knowledge is incrementally refined across distributed data sources, making it suitable for the experimental objectives of this study.

During each training iteration, the global model is progressively updated across exchanges. The model is initially trained on a single exchange (e.g., Exchange A), after which the revised model is sequentially transferred to other exchanges (Exchange B and Exchange C), permitting each participant to incrementally improve the model using its local data.

Upon finishing the sequential updates on all exchanges, the final model is sent back to the server and regarded as the revised global model. The server subsequently spreads this model to all exchanges to start the next training round. This repetitive process persists for a set number of federated rounds.

This sequential boosting approach allows the model to gradually incorporate insights from various distributed data sources while preserving data locality. This method illustrates a controlled experimental framework instead of a typical federated learning protocol, aiming to enable the examination of model performance in distributed training environments. The training process is observed with validation data, and the best round is chosen based on evaluation metrics like PR-AUC. Threshold tuning is then carried out using the validation set, and the ultimate evaluation is performed on the test set to evaluate the generalization capabilities of the global model.

2.5 Evaluation Metrics and Trade off Analysis

The evaluation of model performance was conducted using various metrics such as Accuracy, Precision, Recall, F1-score, and PR AUC. Of these metrics, PR AUC was selected as the main metric because [10] it effectively reflects the trade off between precision and recall, especially in classification tasks with possibly imbalanced class distributions.

Besides predictive performance, this research also assesses the operational aspects of the learning process. The analysis focuses on convergence behavior across federated rounds, training duration, and communication overhead to provide a more comprehensive evaluation of the learning method. Training duration is measured in seconds, representing the total end-to-end time required for model training [11]. Communication overhead is quantified as the total amount of data exchanged between clients and the central server during the federated training process, measured in megabytes (MB) per communication round.

This evaluation framework aims to assess classification performance while also facilitating comparative analysis between centralized and federated learning. Specifically, the integration of performance metrics and operational measures directly aids in evaluating the trade off between model utility and operational overhead, which is crucial to the research goals.

3. LITERATURE STUDY

Detecting illicit activities in blockchain systems poses distinctive challenges because of the pseudonymous characteristics of transactions and the decentralized ownership of information among various entities. To address these challenges, the study of detecting illicit activities has developed through multiple supportive avenues. The initial set of studies concentrates on analytics at the address level within Ethereum. Demonstrated that a single real-world entity can manage several addresses, suggesting that on chain suspicious activity is frequently scattered among accounts that appear to be independent. This discovery is significant as it indicates that illegal activities on Ethereum cannot solely be comprehended through individual addresses, but should be analyzed within a larger transactional framework [3]. Demonstrated that behavioral traits engineered from Ethereum transaction histories can enable precise illicit-account classification via XGBoost, while further validated the efficacy of boosting-based models on Ethereum tabular data. Nonetheless, these studies were conducted based on a centralized learning premise, in which all data is combined into one analytical setting prior to training [2].

A second stream of literature moves beyond tabular address features toward graph-based and subgraph-based analysis. Model Ethereum account behavior using interaction graphs to capture complex relations among addresses [2], while applied graph embeddings to detect laundering-related accounts on Ethereum [12]. Further argued that money laundering on blockchain is often expressed as suspicious transaction subgraphs, rather than as isolated account behavior [13]. These studies deepen the

understanding of laundering as a relational and structural phenomenon. Nevertheless, their main contribution lies in richer graph representation and detection structure, not in addressing the practical problem of cross entity data distribution [14]. In other words, they improve what to detect, but still largely assume that relevant data can be analyzed in a centralized setting.

A third area of research explores federated learning for detecting financial crimes or AML. Demonstrated that federated learning enables joint AML analytics while avoiding the need for institutions to share raw data [5]. This is a crucial measure as it tackles the institutional limitation that sensitive financial information is frequently dispersed and cannot be easily centralized. Nonetheless, these studies primarily rely on conventional financial datasets or simulated banking contexts, rather than blockchain transaction information. Consequently, their results fail to accurately reflect the structural and behavioral aspects of Ethereum transactions, including token interactions, smart contract activities, and address level pseudonymity. Consequently, while federated learning has appeared promising in AML overall, its practical effectiveness for identifying illicit addresses on Ethereum has not been adequately confirmed.

Federated gradient boosting and privacy conscious boosting frameworks, such as SecureBoost [15], its expansion SecureBoost+ [9], and modifications of FedXGBoost [6], are the focus of a fourth line of inquiry. These studies demonstrate that gradient boosting models may be tailored for federated contexts, allowing cooperative model training without sharing raw data and, in certain cases, improving communication effectiveness or maintaining predictive performance.

The main emphasis of these studies is methodological, highlighting elements like privacy protection, the structure of vertical or horizontal federated learning, and enhancing communication efficiency. They are not uniquely designed for AML detection in blockchain environments, nor do they directly assess model effectiveness on Ethereum illicit-address datasets or across multi exchange situations. Consequently, although this body of work establishes a solid technical basis for federated boosting, it does not specifically tackle the unique challenges of blockchain-related AML detection, creating an opportunity that this study seeks to investigate

Based on these four strands, the current literature reveals a clear gap. Existing Ethereum AML studies are strong in behavioral and structural detection, but are mostly centralized. Existing federated AML studies are strong in collaborative learning design, but are mostly grounded in non blockchain financial data. Meanwhile, federated boosting research is strong in algorithmic adaptation, but is not yet specialized for Ethereum illicit address detection. Consequently, the intersection of these three dimensions Ethereum AML detection, XGBoost based classification, and federated learning under a multi exchange setting remains underexplored.

This study is positioned precisely at that intersection. Rather than proposing a new privacy preserving protocol or a new graph architecture, it contributes an empirical evaluation of whether Federated XGBoost can maintain competitive model utility relative to Centralized XGBoost for illicit-address detection on Ethereum, while also accounting for operational overhead such as training runtime and communication cost. In this sense, the novelty of the study lies not in inventing a new detection model, but in providing a controlled baseline for understanding the feasibility and limitations of federated boosting for blockchain AML analytics in a simulated multi-exchange environment. Because the experiment is conducted under an IID partitioning scheme and without additional privacy-preserving cryptographic mechanisms, the results should be interpreted as a baseline empirical study, not as definitive evidence of real-world deployment readiness.

4. RESULT AND DISCUSSION

Based on the identified research gap, this study evaluates the effectiveness of federated XGBoost for illicit address detection on Ethereum under a simulated multi exchange setting. The evaluation focuses not only on predictive performance compared to centralized learning, but also on the operational trade off associated with the federated training process. The results of this evaluation are presented and discussed in the following section.

4.1 Performance Comparison between Centralized and Federated Learning

The comparative performance between centralized and federated learning is illustrated in Figure 4. Overall, the results show that the federated XGBoost model is able to maintain a competitive level of predictive performance across all evaluation metrics when compared to the centralized baseline.

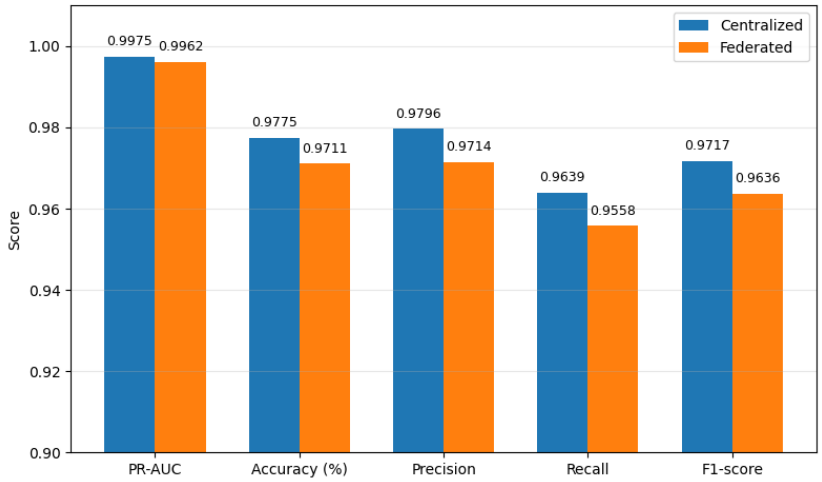


Figure 4. Utility Comparison on Centralized and Federated baselines

The federated model's PR-AUC score of 0.9962 is just somewhat less than that of the centralized model 0.9975. This suggests that even in the absence of direct access to centralized data, the federated method maintains a significant capacity to differentiate between illegitimate and legitimate addresses. In a similar vein, the federated model's accuracy is 97.11% as opposed to 97.75% in the centralized configuration, showing just a little decline in overall classification performance.

Other measures, such as Precision (0.9714 vs. 0.9796), Recall (0.9558 vs. 0.9639), and F1-score (0.9636 vs. 0.9717), also show a same pattern. Even though the centralized model continuously performs better than the federated model, the performance difference is still rather tiny across all measures, suggesting that, in the current experimental scenario, the information loss resulting from data decentralization is negligible.

This very tiny performance difference indicates that the federated training method may successfully integrate information from several data partitions under homogenous data distribution (IID). Despite the lack of raw data exchange, the global model is able to capture comparable decision limits because the aggregation of local model updates effectively approximates the behavior of centralized learning.

These findings are consistent with prior studies that demonstrate the effectiveness of XGBoost-based models for illicit address detection on Ethereum [2], where high classification performance is achieved using centralized datasets. However, unlike these studies, which assume full data availability, this work shows that comparable performance can still be achieved under a distributed training setting without direct data sharing.

These findings suggest that, within the context of a simulated multi-exchange environment, federated learning can serve as a viable alternative to centralized approaches when data sharing is restricted due to privacy or regulatory constraints. The ability to achieve near-centralized performance while preserving data locality highlights its potential applicability in real-world AML systems. However, this implication should be interpreted with caution, as the current evaluation is conducted under an IID assumption.

It is crucial to stress, nonetheless, that the experimental environment has a significant impact on this outcome. An idealized situation where local data distributions are constant between exchanges is represented by the usage of IID data partitioning. Data is usually heterogeneous (non-IID) in real-world multi exchange situations, which can pose serious problems such delayed convergence, model bias, and performance deterioration. Therefore, rather than being a direct depiction of deployment settings in the actual world, the study's results should be viewed as a best-case baseline for federated learning performance.

4.2 Convergence Behavior in Federated Training

The convergence behavior of the federated model can be observed from the performance trends across training rounds, as illustrated in Figure 5 and Figure 6. Figure 5 shows the evolution of validation PR AUC, while Figure 6 presents the corresponding validation Accuracy across federated rounds.

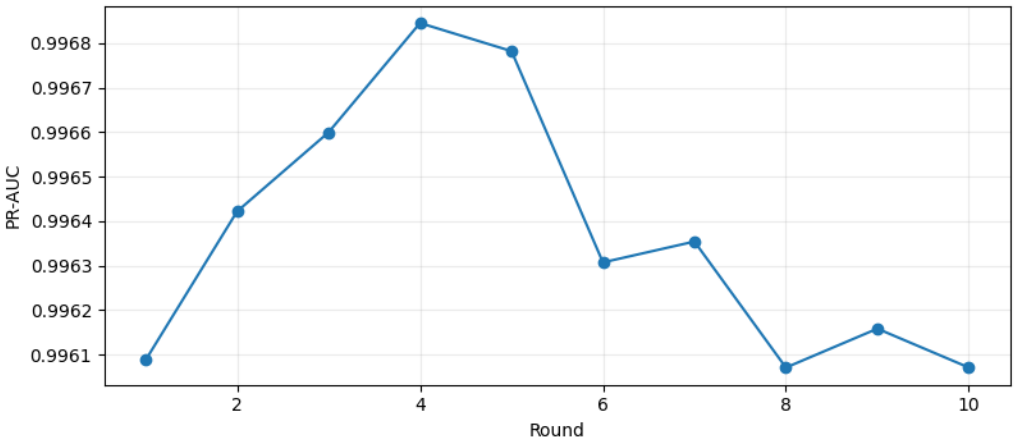


Figure 5. Convergence Development in PR-AUC Validation

From Figure 5, the PR AUC exhibits a rapid increase during the early rounds, reaching its peak around round 4. This indicates that the model is able to quickly capture the underlying patterns of illicit activity from distributed data. After this point, the PR AUC begins to fluctuate slightly and gradually stabilizes, suggesting that the model has approached its optimal representation and further updates provide only marginal improvements.

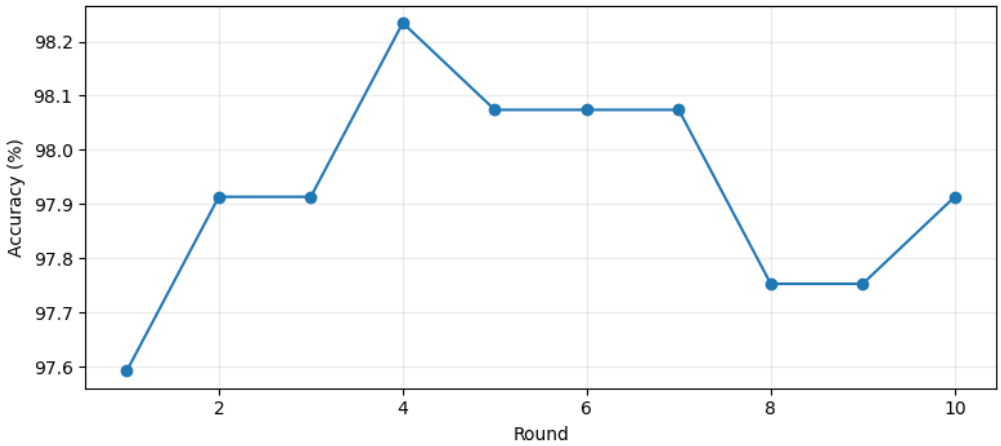


Figure 6. Convergence Development in Accuracy Validation

A comparable convergence trend is seen in Figure 6 for validation Accuracy. The precision markedly enhances during the early rounds and peaks around round 4, then displays a fairly steady pattern with slight variations in the later rounds. The minor drop and fluctuation following the peak show that extra federated updates do not reliably enhance performance, a typical feature of convergence in iterative learning systems.

These findings validate that the sequential federated boosting approach successfully enhances the global model through iterative improvements during exchanges. The model can maintain consistent performance over a restricted number of federated rounds, showcasing effective learning regardless of the data being distributed.

From an operational viewpoint, this convergence behavior holds significant importance. As the improvement in performance diminishes after a specific number of rounds, one can restrict the number of communication rounds to decrease training duration and communication expenses without

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

considerably compromising model performance. This result reinforces the feasibility of federated learning in multi-exchange situations where communication costs are a significant limitation.

It is crucial to highlight that the identified convergence takes place within an IID data distribution framework. In practical situations, where data from exchanges is usually non-IID, the convergence patterns could vary greatly. Differences in local data distributions may result in slower convergence, increased instability, and possible decline in the performance of the global model. Consequently, the findings of this study must be understood as a baseline situation, and additional research is necessary to assess convergence under more realistic non-IID circumstances.

4.3 Trade off between Model Utility and Operational Overhead

Beyond predictive performance, this study evaluates the operational implications of federated learning, particularly the trade off between model utility and training time, as illustrated in Figure 7.

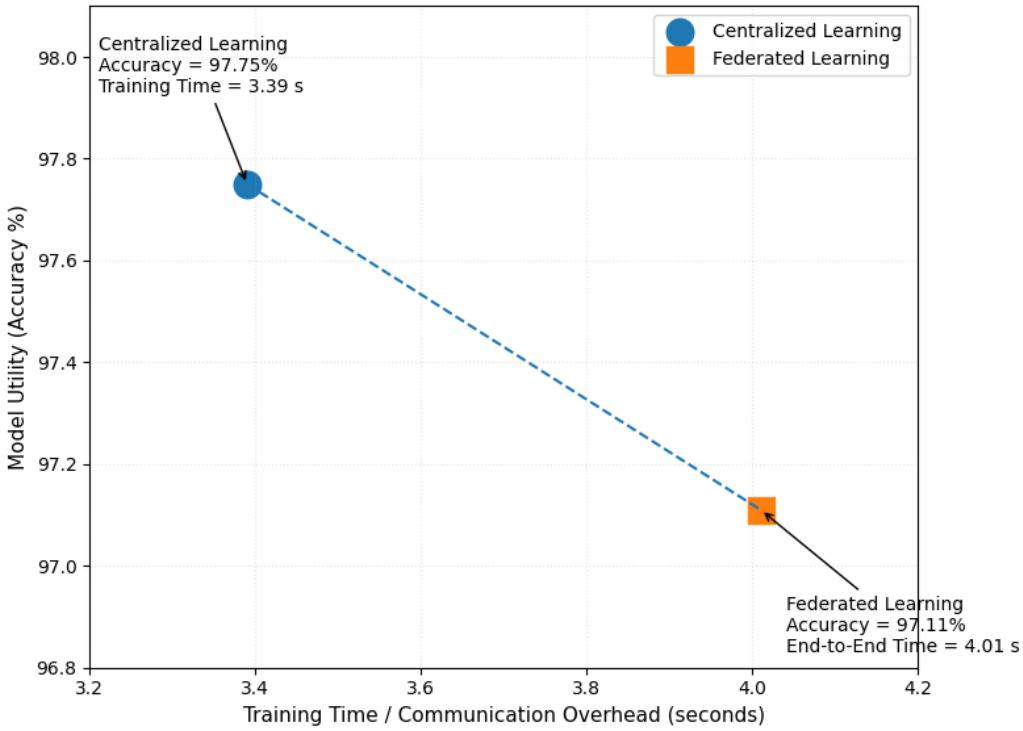


Figure 7. Trade off between model utility and training time against centralized and federated baselines

The results in Figure 7 show that centralized learning achieves slightly better accuracy 97.75% with a shorter training duration of 3.39 seconds, whereas the federated method achieves similar accuracy 97.11% but requires a longer end to end training time of 4.01 seconds. This suggests that federated learning imposes an additional computational burden despite achieving comparable predictive performance.

The increased training duration in the federated framework is primarily due to communication overhead, not local computation. Although each exchange performs local training efficiently, the repeated need to share model updates between clients and the aggregation server over multiple rounds significantly impacts the total runtime. This behavior is clearly evident in the changes shown in Figure 7, where the federated model lies further along the training time axis despite having only slightly lower accuracy.

This observation highlights a core trade off in federated learning systems: achieving decentralized training and data privacy results in greater communication demands and longer training durations. Although the utility loss is small, this additional overhead must be considered when implementing federated learning in real world settings.

In contrast to existing federated learning studies in the financial domain [6]; [7], which primarily emphasize predictive performance, this study explicitly highlights the operational implications of federated learning, particularly communication overhead and training duration. These system-level costs

are often underreported in prior work, yet they play a critical role in determining the feasibility of real-world deployment.

In this study, communication costs ranged from 1 to 5 MB per round, which remains feasible in a simulated environment. In large-scale implementations involving additional data exchange, increased model complexity, or limited network bandwidth, this additional burden could become a major obstacle hindering the system's scalability and responsiveness.

Consequently, these results highlight that the success of federated learning should not be evaluated solely based on prediction accuracy. Instead, system level factors such as communication efficiency, network conditions, and training cycle optimization are crucial in assessing the feasibility of federated learning for real-world AML applications.

4.4 Feature Importance Analysis

The feature importance analysis of the federated XGBoost model is presented in Figure 8, which highlights the top 20 features ranked by their contribution gain to the classification process. The results show that features related to token interactions, particularly ERC20-based activities, dominate the model's decision-making.

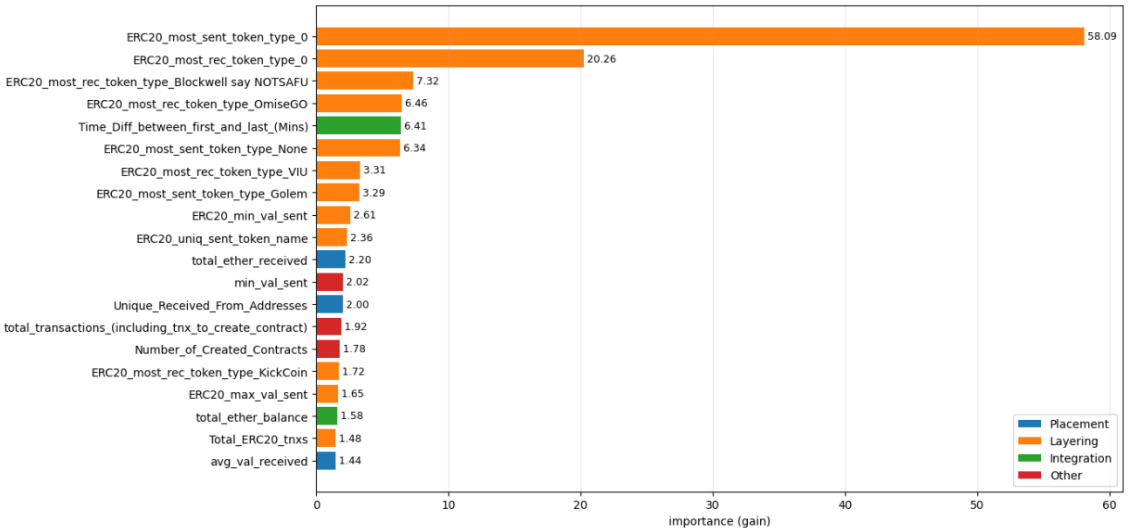


Figure 8. Feature Importance Analysis

The most impactful feature is ERC20_most_sent_token_type_0, with ERC20_most_rec_token_type_0 close behind, as both exhibit considerably greater significance than other features. This suggests that token transfer patterns, instead of mere Ether (ETH) transactions, offer the most powerful signals for recognizing illicit addresses. Extra token-related attributes, like particular token categories (e.g., Blacklist, OmiseGO, along with various ERC20 types), further emphasize the significance of interaction patterns at the token level.

The prevalence of ERC20-related functions indicates that illegal activities within the Ethereum ecosystem tend to include intricate token transactions rather than standalone ETH transfers. Such actions align with established money laundering tactics in blockchain networks, where assets are frequently channeled through various tokens and transactions to conceal their source. In this setting, token interaction characteristics can be seen as indicators of layering behavior, wherein attackers try to obscure transaction trails via organized and multi-step processes.

Conversely, attributes pertaining to fundamental transaction metrics, like total Ether acquired, lowest transaction amount, or overall balance, show comparatively lesser significance. This suggests that basic monetary aggregates are not as effective in identifying illegal behavior as interaction-based characteristics that reflect structural complexity.

From an AML viewpoint, the allocation of feature significance can likewise be linked to various phases of unlawful financial conduct. Token interaction attributes are mainly associated with the layering stage, where transaction concealment takes place, whereas elements like time discrepancies and balance-related characteristics might reflect signals concerning placement or integration. This

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

underscores the model's capacity to subtly discern patterns throughout various phases of illegal transaction flows.

It is crucial to understand that feature importance indicates statistical significance in the model rather than causal connections. The recognized characteristics do not necessarily suggest that particular tokens or transaction types are intrinsically illegal instead, they signify that these trends are the most telling for categorization within the specified dataset and experimental environment. Consequently, the interpretation ought to be considered as an analytical understanding of model performance instead of conclusive proof of unlawful processes.

4.5 Discussion and Research Positioning

The results of this research offer significant insights into the practical applicability of using federated learning for AML detection on blockchain. Although previous studies have demonstrated strong predictive performance using centralized machine learning models for illicit address detection on Ethereum [2], these approaches rely on full data aggregation and do not account for distributed data constraints in real-world multi-exchange environments.

This research directly tackles this issue by offering a controlled empirical evaluation of centralized versus federated XGBoost within a simulated multi-exchange scenario. The findings indicate that federated learning can closely match centralized performance in IID conditions, suggesting that the decline due to data decentralization may be minimal when distributional variations are managed. This discovery is important because it indicates that the main restriction of federated learning in this scenario is not so much the model's abilities, but instead the limitations at the system level.

In comparison to prior studies, this finding highlights a key distinction. Centralized approaches in [2] achieve high performance under the assumption of complete data availability, whereas this study demonstrates that similar performance can still be achieved in a distributed setting without raw data sharing. However, unlike many federated learning studies in the financial domain [6];[7].,which primarily emphasize predictive accuracy, this work explicitly reveals the operational costs associated with federated training, particularly communication overhead and increased training duration. This suggests that while federated learning maintains predictive capability, its practical feasibility is strongly influenced by system-level constraints that are often overlooked in previous research.

Nevertheless, the analysis indicates that the expenses associated with decentralization are mainly operational. The noted rise in training duration, caused by communication overhead, emphasizes a core trade off between model effectiveness and system efficiency. This emphasizes that assessing federated learning only through predictive performance is inadequate, especially in practical multi-entity settings where communication and scalability are vital considerations. Crucially, this research does not intend to suggest a completely optimized federated learning framework, but instead to create a foundational comprehension of federated boosting dynamics in blockchain AML assignments. The application of IID data partitioning and the lack of privacy-preserving measures create a deliberately managed experimental environment, enabling the investigation to focus on the impact of the learning paradigm alone.

Consequently, the findings ought to be seen as an ideal baseline scenario, instead of a direct depiction of actual deployment circumstances. In reality, non-IID data distributions, larger systems, and tighter privacy requirements are likely to pose extra challenges that could greatly impact model performance and system efficiency.

In this context, the primary contribution of this study lies in bridging the gap between algorithmic federated learning research and domain-specific AML applications on blockchain. By jointly evaluating predictive performance and operational overhead, this work provides a foundational reference for future research aiming to deploy federated learning in realistic multi-exchange environments.

Therefore, this study extends previous work by not only validating the applicability of federated learning in blockchain-based AML, but also by explicitly quantifying the performance–efficiency trade-off in a multi-exchange setting. By jointly evaluating predictive performance and operational overhead, this work provides a more comprehensive perspective compared to prior studies that predominantly focus on predictive metrics alone.

5. CONCLUSION

This study evaluates the effectiveness of federated learning for illicit address detection on the Ethereum blockchain through a comparative analysis of centralized and federated XGBoost in a simulated multi-exchange setting. The results show that federated learning achieves competitive predictive performance, with a PR-AUC of 0.9962 and an accuracy of 97.11%, compared to 0.9975 and 97.75% in the centralized setting.

Despite this near-equivalent performance, federated learning introduces additional operational costs, reflected in longer training time 4.01s vs. 3.39s, primarily due to communication overhead. This confirms the presence of a performance efficiency trade off, where decentralized training preserves predictive capability but incurs system-level costs.

The main contribution of this study lies in providing a controlled empirical evaluation of this trade off in a blockchain AML context, highlighting that the primary limitation of federated learning is not model capability, but operational efficiency.

However, these findings are derived from a controlled IID setting and may not fully represent real-world scenarios with heterogeneous data distributions. Future work should investigate non-IID conditions, larger-scale systems, and communication efficient or privacy preserving mechanisms to assess the practical deployment of federated learning in real-world AML environments.

REFERENCES

[1] The Financial Action Task Force (FATF), "Risk Based Approach Guidance for the Real Estate Sector," *Fatf*, no. July, 2022, [Online]. Available: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Guidance-rba-real-estate-sector.html

[2] L. Pahuja and A. Kamal, "EnLEFD-DM: Ensemble Learning based Ethereum Fraud Detection using CRISP-DM framework," *Expert Syst.*, vol. 40, no. 9, pp. 1–18, 2023, doi: 10.1111/exsy.13379.

[3] Z. Gu and O. Dib, "Enhancing fraud detection in the Ethereum blockchain using ensemble learning," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2716.

[4] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowl. Based. Syst.*, vol. 216, p. 106775, 2021, doi: 10.1016/j.knosys.2021.106775.

[5] B. Guembe, A. Azeta, V. Osamor, and R. Ekpo, "A Federated Machine Learning Approaches For Anti-Money Laundering Detection," *SSRN Electronic Journal*, vol. 1, pp. 1–13, 2022, doi: 10.2139/ssrn.4669561.

[6] C. Ma, X. Qiu, D. Beutel, and N. Lane, "Gradient-less Federated Gradient Boosting Tree with Learnable Learning Rates," pp. 56–63, 2023, doi: 10.1145/3578356.3592579.

[7] A. A. Ahmed and O. O. Alabi, "Secure and Scalable Blockchain-Based Federated Learning for Cryptocurrency Fraud Detection: A Systematic Review," *IEEE Access*, vol. 12, no. August, pp. 102219–102241, 2024, doi: 10.1109/ACCESS.2024.3429205.

[8] W. Lindskog and C. Prehofer, "A Federated Learning Benchmark on Tabular Data: Comparing Tree-Based Models and Neural Networks," *2023 8th International Conference on Fog and Mobile Edge Computing, FMEC 2023*, pp. 239–246, 2023, doi: 10.1109/FMEC59375.2023.10305887.

[9] T. Fan, W. Chen, G. Ma, Y. Kang, L. Fan, and Q. Yang, "SecureBoost+: Large Scale and High-Performance Vertical Federated Gradient Boosting Decision Tree," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14647 LNAI, pp. 237–249, 2024, doi: 10.1007/978-981-97-2259-4_18.

[10] C. C. Wang, P. H. Kuo, and G. Y. Chen, "Machine Learning Prediction of Turning Precision Using Optimized XGBoost Model," *Applied Sciences (Switzerland)*, vol. 12, no. 15, 2022, doi: 10.3390/app12157739.

[11] S. Garst, J. Dekker, and M. Reinders, "A comprehensive experimental comparison between federated and centralized learning," *Database*, vol. 2025, no. 00, pp. 1–10, 2025, doi: 10.1093/database/baaf016.

[12] J. Liu *et al.*, "Graph Embedding-Based Money Laundering Detection for Ethereum," *Electronics (Switzerland)*, vol. 12, no. 14, pp. 1–14, 2023, doi: 10.3390/electronics12143180.

[13] C. Y. Dan Lin, Jiajing Wu, Yunmei Yu, Qishuang Fu, Zibin Zheng, "DenseFlow: Spotting Cryptocurrency Money Laundering in Ethereum Transaction Graphs," *Proceedings of the ACM Web Conference 2024*, 2024, [Online]. Available: https://github.com/DenseFlow.

[14] K. Song *et al.*, *Identifying Money Laundering Subgraphs on the Blockchain*, vol. 1, no. 1. Association for Computing Machinery, 2024. doi: 10.1145/3677052.3698635.

[15] K. Cheng *et al.*, "SecureBoost: A Lossless Federated Learning Framework," *IEEE Intell. Syst.*, vol. 36, no. 6, pp. 87–98, 2021, doi: 10.1109/MIS.2021.3082561.