OPEN ACCESS Freely available online

Journal of Theoretical and Computational Science

Mini Review

Anti-Money Laundering (AML) Detection Platform Leveraging Federated Learning with NVIDIA Federated Learning Application Runtime Environment (FLARE)

Venkatesh Upadrista[1*], Nitin Bhargava[2], Ram Gopal[3]

[1]School of Computing, Glasgow Caledonian University, Glasgow Scotland, United Kingdom; [2] Independent Researcher, Chief Operating Officer, Arab Bank for Investment & Foreign Trade (Al Masraf), Abu Dhabi, U.A.E; [3]Independent Researcher, Former Chief Executive Officer, Barclays Bank (India), Mumbai, India

ABSTRACT

Money laundering remains a significant challenge to the global financial system, employing complex and evolving methods that outpace current regulations. Despite the implementation of Anti-Money Laundering (AML) compliance measures such as Know Your Customer (KYC) and Customer Due Diligence (CDD), sophisticated laundering schemes continue to exploit gaps in existing systems. Traditional rule-based monitoring systems often result in high false positive rates, leading to inefficiencies and increased operational costs. While machine learning has been employed to enhance anomaly detection, issues such as imbalanced datasets, frequent false alarms and limited adaptability to new money laundering tactics still persist. To effectively combat money laundering, there is a need for more advanced, collaborative solutions that can adapt to emerging threats.

This research introduces an advanced AML detection platform utilizing the NVIDIA Federated Learning Application Runtime Environment (NVIDIA FLARE), integrating Graph Neural Networks (GNN) and eXtreme Gradient Boosting (XGBoost) machine learning models. The platform is designed to foster collaboration among financial institutions by encouraging them to share insights on AML threats and experiences through a common model without sharing raw data. This approach allows banks to collectively improve a shared model, enabling it to learn from the money laundering incidents faced by one institution so that similar incidents can be prevented at other banks, thereby mitigating overall risk and reducing financial losses. Federated learning enables the creation of such a platform without centralizing data, significantly enhancing detection capabilities through greater data diversity, reduced biases and increased adaptability to emerging money laundering patterns.

The developed platform achieved a detection accuracy of 96.22%, demonstrating the effectiveness of federated learning in enhancing AML detection while ensuring compliance with data privacy regulations. By promoting inter-bank learning rather than isolated operations, the platform incorporates the collective knowledge of money laundering attacks into the shared model, preventing similar impacts across institutions. This collaborative effort allows banks to reduce the overall impact of money laundering incidents and enhance the resilience of the financial system.

Keywords: Anti-Money Laundering; Machine learning; Federated Learning; Privacy-preserving; Anomaly detection; Transaction monitoring; Know Your Customer; Data privacy

Abbreviations: AE: Autoencoders; AI: Artificial Intelligence; AML: Anti-Money Laundering; ANN: Artificial Neural Networks; CDD: Customer Due Diligence; CNN: Convolutional Neural Networks; CUDA: Compute Unified Device Architecture; cuDNN: CUDA Deep Neural Network; DNN: Deep Neural Networks; EC2: Elastic Compute Cloud; FedAvg: Federated Averaging; FLARE: Federated Learning Application Runtime Environment; GAGNN: Group-Aware Deep Graph Learning; GDPR: General Data Protection Regulation; GiB: Gibibyte; GNN: Graph

Correspondence to: Venkatesh Upadrista, School of Computing, Glasgow Caledonian University, Glasgow Scotland, United Kingdom, E-mail: vupadr200@caledonian.ac.uk

Received: 15-Oct-2024, Manuscript No. JTCO-24-34618; **Editor assigned:** 18-Oct-2024, PreQC No. JTCO-24-34618 (PQ); **Reviewed:** 01-Nov-2024, QC No. JTCO-24-34618; **Revised:** 08-Nov-2024, Manuscript No. JTCO-24-34618 (R); **Published:** 15-Nov-2024, DOI: 10.35248/2376-130X.24.10.229

Citation: Upadrista V, Bhargava N, Gopal R (2024). Anti-Money Laundering (AML) Detection Platform Leveraging Federated Learning with NVIDIA Federated Learning Application Runtime Environment (FLARE). J Theor Comput Sci. 10:229

Neural Networks; GPUs: Graphics Processing Units; gp2: General Purpose Solid State Drives; HITS: Hyperlink Induced Topic Search; KYC: Know Your Customer; LIME: Local Interpretable Model-agnostic Explanations; ML: Machine Learning; NLG: Natural Language Generation; POC: Proof-of-Concept; PDPs: Partial Dependence Plots; SARs: Suspicious Activity Reports; SHAP: SHapley Additive exPlanations; SNA: Social Network Analysis; SVM: Support Vector Machines; VAE: Variational Auto encoders; WGAN: Wasserstein Generative Adversarial Networks; XAI: Explainable AI; XGBoost: eXtreme Gradient Boosting

INTRODUCTION

Money laundering poses a substantial threat to the global economy, with losses estimated between $800 billion and $2 trillion annually, accounting for 2%-5% of global Gross Domestic Product (GDP). Banks have implemented stringent AML measures, including KYC, CDD, transaction monitoring and the filing of Suspicious Activity Reports (SARs). Despite these efforts, money laundering remains prevalent due to the complexity of laundering schemes, emerging technologies and expansive global financial networks. Banks have further increased their investments in IT infrastructure to support AML compliance. Global banks spent over $180.9 billion in 2021 on AML-related IT, including solutions for real-time transaction monitoring, identity verification and regulatory reporting. Despite these investments, financial institutions continue to face substantial penalties for failing to comply with AML regulations. In 2023, AML-related penalties increased by 57% compared to 2022, reaching a total of $6.6 billion.

As financial crime evolves, there has been a growing reliance on Artificial Intelligence (AI) and Machine Learning (ML) to detect suspicious activities and improve compliance efficiency.

Leveraging AI and ML can help reduce the significant expenses associated with AML compliance. AI-based systems can automate transaction monitoring, drastically reduce false positives and improve customer risk profiling. Models such as Autoencoder Neural Networks (ANN) can detect anomalies in transactions by learning normal behavior patterns, flagging deviations, as potential money laundering attempts. Additionally, the use of advanced ML models like Graph Neural Networks (GNN) and eXtreme Gradient Boost (XGBoost) helps refine the classification of suspicious transactions, reducing manual reviews and improving decision-making efficiency. GNNs are designed to analyse relationships within data by modelling entities and their connections as graphs, which aids in identifying complex patterns. XGBoost enhances predictive accuracy by building an ensemble of decision trees that iteratively improve upon previous errors. AI-driven dynamic risk models allow banks to focus on high-risk transactions, reducing unnecessary investigations of low-risk activities. Natural Language Generation (NLG) models can automate the generation of SARs, streamlining regulatory reporting and saving significant operational time and cost. Furthermore, AI-powered audit trails provide enhanced transparency, reducing the manual workload associated with compliance audits. The integration of AI and ML into AML systems not only improves accuracy but also lowers the costs of compliance, making it a critical part of modernizing AML efforts. However, current AI-based AML solutions are limited in effectiveness because they are often trained on data from individual banks or entities, preventing them from learning the diverse money laundering patterns present across different institutions. This siloed approach weakens the strength of AI for AML detection, as it restricts knowledge sharing and prevents collaborative advancements against evolving money laundering techniques. Federated learning can overcome this challenge by enabling multiple banks to train a shared global model collaboratively, without sharing sensitive data. By forming a consortium, banks can tackle money laundering on a broader scale, enhancing the collective response to this global issue. Federated learning was proposed by Google researchers in 2016, as a innovative solution for addressing the issues of communication costs, data privacy and legalization [1-4]. Federated learning is a distributed ML approach where models are trained on local edge devices and sent to the central server for consolidation with other similar local models without sharing their local datasets, thus, ensuring privacy of data during the training process.

Federated learning presents a collaborative solution for AML by enabling multiple banks to jointly train AML models without sharing sensitive data. This privacy-preserving approach enhances model performance by leveraging insights from diverse datasets while ensuring compliance with data privacy regulations. Each bank trains a local model using its own AML data and only model updates-rather than raw data-are shared with a central aggregator, which combines these updates into a global model that captures broader patterns across all participating banks to detect AML. By allowing banks to learn from each other's data, federated learning significantly improves the detection of cross-border and sophisticated money laundering schemes that might go unnoticed by individual institutions. This approach addresses regulatory concerns, as it preserves data privacy by ensuring that raw data never leaves the institution.

In this paper, we have developed an NVIDIA FLARE based AML detection federated learning platform that allows multiple banks to join a consortium and collaboratively train a ML model using their own data to detect money laundering activities. This platform enables participating banks to improve the model while preserving data privacy, resulting in a more effective and adaptable AML detection system. The prototype developed on this platform achieved an accuracy of 96.22%, demonstrating the effectiveness of the collaborative approach in enhancing AML detection capabilities. By using this platform, banks can benefit from shared insights while maintaining the privacy of their own data, ensuring that if one bank encounters an AML threat or is impacted by an AML case, other banks can proactively avoid similar situations, thereby reducing the overall economic impact and costs associated with money laundering.

Using ML for AML Detection–A Literature Review

We conducted a literature review across various academic sources to investigate the use of AI and ML for AML. The review highlights several models for AML detection, including Explainable AI (XAI)-based methods, deep learning and sentiment analysis approaches. The highest accuracy achieved was 96% (AUC) by Group-Aware Deep Graph Learning (GAGNN) and 93% by the Variational Auto encoders (VAE) and Wasserstein Generative

Adversarial Networks (WGAN) models, with the DNN model reaching a recall of 0.94. These results indicate the need for further improvement in prediction accuracy. Additionally, current models face limitations such as data imbalance, lack of interpretability and limited adaptability to emerging patterns. Therefore, new models must be developed to enhance the effectiveness and reliability of AML detection systems.

MATERIALS AND METHODS

As part of their research, Li et al., proposed an Explainable AI (XAI)-based AML model to address the limitations of traditional deep learning techniques, which are often considered black-box models [5]. They utilized several algorithms including Decision Tree, Random Forest, XGBoost, and Deep Neural Networks (DNN). The DNN model achieved the highest recall of 0.94 for detecting money laundering activities. To enhance interpretability, the authors employed XAI tools like SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME) and Partial Dependence Plots (PDPs) to explain the model's predictions and identify key features, with "Payment Format" being the most significant.

Hamid proposed using Support Vector Machines (SVM) and Social Network Analysis (SNA) to enhance AML detection [6]. The SVM algorithm was applied to classify transactional patterns based on behavioural norms, creating a probabilistic adaptive matrix. Meanwhile, SNA was used to analyse the relationships within money laundering networks, enhancing prediction capabilities. The experimental results demonstrated an effective detection system, with accuracy figures related to identifying suspicious transactions or relationships in money laundering not explicitly stated. However, the proposed approach improved system efficiency and the detection of laundering activities.

As part of their study, Raj et al., used AI to detect money laundering by employing ML, ANN and Convolutional Neural Networks (CNN) [7]. ML was used to identify suspicious transaction patterns, ANN was utilized to find complex relationships between variables and CNN helped analyse visual inputs such as document images. The study demonstrated the effectiveness of AI in improving AML processes, particularly in KYC, risk management and transaction monitoring. However, specific accuracy numbers for detection or prediction were not mentioned.

Wang et al., used a decision tree algorithm to evaluate money laundering risks based on customer profiles from a Chinese commercial bank [8]. The decision tree was built using data from 28 customers with four attributes: industry, location, business size and bank products. The algorithm generated rules to classify money laundering risk into low, middle or high categories. The results showed that the decision tree effectively classified 21 training samples and correctly predicted the risk for 6 out of 7 test samples, achieving an accuracy of approximately 85.7%.

As part of their study, Li et al., proposed an XAI-based AI AML model using deep learning to detect abnormal transactions [5]. They implemented decision tree, random forest, XGBoost, and deep neural network (DNN) models, with the DNN achieving the highest recall of 0.94. To improve model interpretability, they employed explainable AI tools such as SHAP, LIME, and PDPs. The feature "Payment Format" was found to be the most significant for identifying potential money laundering activities, with ACH and cash transactions contributing significantly to flagged activities.

As part of their study, Kute et al., reviewed deep learning and explainable AI (XAI) techniques applied to AML [9]. They found that deep learning techniques like CNN and Auto Encoder were preferred for detecting suspicious money laundering activities. Graph deep learning combined with natural language processing was identified as a promising approach for AML. The key challenges included data imbalance, lack of interpretability and limited access to recent transaction data. They concluded that the use of XAI could improve transparency and aid in adoption by financial institutions.

Thi et al., developed an AML system based on sentiment analysis using data from social media and online news sources [10]. They employed Google Cloud's AutoML for training the model, with data collected from Kaggle and other public sources. The trained model analysed the sentiment of news articles and tweets about companies, achieving an accuracy of 85%. The output allowed financial institutions to evaluate the reputation of new clients, thereby improving decision-making related to KYC processes.

As part of their study, El-Kilany et al., proposed an AML framework using a modified version of the Hyperlink Induced Topic Search (HITS) algorithm to detect suspicious customers [11]. The framework consisted of three stages: data modelling, detection and verification. A weighted HITS algorithm was used to calculate authority scores for each customer, ranking their suspiciousness. One-class SVM was then applied for verification. The experimental results showed a precision of 85.7% for the top 10 suspicious customers in a dataset of 1K vertices and 100K edges.

As part of their study, Oztas et al., used various ML techniques to enhance transaction monitoring controls for detecting money laundering [12]. Approaches included using supervised and unsupervised models such as Bayesian algorithms, SVM, Neural Networks, Isolation Forest and Random Forest. The Isolation Forest method was noted for its efficiency, achieving better accuracy compared to other approaches like one-class SVM and was preferred due to its shorter computation time and lower memory usage. The authors highlighted a reduction in false positives to 6.2%, with a true positive rate of 65.5%.

Cheng et al., proposed an AML detection system using Group-Aware Deep Graph Learning (GAGNN) [13]. The model used a community-centric encoder to analyse both nodes and group behaviour within financial transactions, identifying patterns indicative of organized money laundering. Experimental results, conducted on a dataset from Union Pay, showed that GAGNN outperformed existing graph learning benchmarks in detecting suspicious transactions, achieving an AUC of 96% and a recall rate of 86% for the top-ranked suspicious cases.

As part of their study, Chen et al., proposed using VAE and WGAN for improving AML processes [14]. The models were used to address the challenge of detecting fraud in highly imbalanced datasets by generating synthetic fraud transactions using WGAN, which were then mixed with the original dataset to train VAE and Autoencoders (AE). The experimental results showed a false positive rate as low as 7% for the multi-loss AE model, while achieving 93% accuracy and 100% recall of fraud transactions.

Federated learning platform architecture to detect anti-money laundering

Traditional ML algorithms often require centralizing large amounts of data for training, which poses significant privacy,

compliance and security challenges, especially in financial sectors. When banks share sensitive customer transaction data, they risk exposing it to potential data breaches and fail to comply with regulations such as General Data Protection Regulation (GDPR). Moreover, models trained on a single bank's data may have limited accuracy and generalizability due to the lack of diversity and volume, resulting in more false positives and missed detections. False positives refer to instances where a test or model incorrectly identifies something as positive when it is actually negative. In other words, it is an error where the system incorrectly flags an action, condition or instance as significant when it is not.

We propose a platform-based architecture for AML detection where multiple banks can form a consortium to collaboratively train a ML model. Each bank can join the platform, pull the model from the central server, train it using their own dataset and then put the updated model back on the platform. The central authority aggregates these updated models from all participating banks to create a more refined global model. This approach not only enhances the prediction accuracy of the ML models to detect AML but also ensures data privacy. Moreover, banks can join or leave the platform at any time, allowing for flexible participation based on their needs. By using this platform, the impact of AML threats faced by one bank can be reduced for others, ensuring better overall risk management, making it an essential tool for reducing the overall risks and costs associated with money laundering globally.

Platform architecture

The banking sector is increasingly adopting advanced AI/ML technologies to enhance security and prevent financial crimes such as money laundering. Federated learning, in particular, presents a significant advancement in this area, enabling multiple banks to collaborate on building superior models for detecting fraudulent transactions without sharing sensitive customer data. This approach not only enhances privacy but also overcomes challenges associated with centralized data storage, regulatory compliance and cross-institution collaboration.

Federated learning allows banks to jointly develop a global model to detect AML transaction by training individual models locally and aggregating only the learned parameters. This ensures that data stays within each bank, thereby maintaining the confidentiality of customer information and significantly reducing the risk of data breaches. Some of the benefits of federated learning in the AML context:

Enhanced data privacy and compliance: Financial institutions are subject to strict data privacy and protection regulations, which often limit the extent of data sharing across entities. Federated learning addresses these challenges by ensuring that customer data remains on-premises. Only the model updates (parameters) are shared, not the data itself. This ensures compliance with privacy laws such as GDPR and enables collaboration without the need for direct data transfer between institutions.

Broader and more robust data representation: Anti-money laundering models require a diverse and comprehensive dataset to accurately detect suspicious transactions and adapt to evolving money laundering techniques. By using federated learning, banks can train a global model using data from multiple institutions without sharing raw data. This collaborative approach allows each bank to benefit from the experiences and data of others, effectively eradicating the impact of money laundering tactics encountered by one bank as other banks learn from them. Consequently, this leads to a more comprehensive model that captures diverse transaction patterns and trends observed across institutions, resulting in better detection rates and fewer false positives.

Collaborative learning without risk: Money laundering schemes are becoming more sophisticated, often involving activities that span across multiple banks. Federated learning allows financial institutions to collaborate on AML initiatives by sharing insights without revealing sensitive information. This means that banks can collaboratively train a model to detect complex, multi-bank money laundering schemes, providing a more holistic approach to crime detection and prevention.

Reduction in data breaches and centralized vulnerabilities: Centralized data storage often becomes a target for cyber-attacks and large-scale breaches can have catastrophic consequences for banks and their customers. Federated learning mitigates this risk by avoiding centralized data collection altogether. Since each bank retains its own data, the attack surface is minimized and the potential impact of a breach is significantly reduced.

Real-time detection and adaptation: Federated learning can also be used to create models that are capable of near real-time updates. As different institutions train their models on the latest local data and share model parameters, the global model is updated frequently, allowing for the rapid identification of emerging money laundering techniques. This adaptive approach is particularly important given the fast-evolving nature of financial crimes.

Efficient use of resources: By training local models and aggregating updates, federated learning reduces the need to transfer large amounts of data between different locations. This leads to lower network costs and faster model training, as each institution only needs to share model parameters rather than the complete dataset. This also reduces the infrastructure burden, as banks do not need to manage the security and storage complexities associated with large-scale centralized datasets.

Preserving customer trust: With increased customer awareness regarding data privacy, federated learning provides a transparent approach that reassures customers about how their data is handled. Knowing that sensitive financial information remains within the bank and is not shared externally can help banks maintain customer trust while still benefiting from advanced collaborative security measures.

A scalable federated learning platform architecture for AML consists of three layers, as shown in Figure 1.

Banks landing station

This layer consists of data provided by various financial institutions. The raw transactional data of bank customers, including suspicious activity reports, transaction histories and customer profiles, is gathered by the banks.

Local training layer

This layer handles the training of individual models using data collected by each bank. A base ML model, known as the "initial model," is stored on a central server. Participating banks (e.g., Bank A, as shown in figure 1) download the initial model and train it locally using their transaction data. Once the training is complete, only the updated model parameters are sent back to the central server.

The advantage of this approach is that each bank's data is never shared directly, ensuring customer privacy and regulatory compliance. By aggregating learning's from multiple institutions, federated learning creates a more comprehensive and effective model for detecting money laundering, without compromising data privacy.

Central server (aggregator)

A central server coordinates the AML detection learning process by aggregating the model parameters (not raw data) sent by the banks. The server computes a global model and distributes updated parameters back to each participating bank. The centralized server aggregates the models trained locally at different banks. It coordinates the training across participating institutions by receiving model parameters from individual banks and consolidating them into a global model. The central server may be deployed on the cloud or in a secure local data center to facilitate the sharing of the updated model.

The key benefit for is that only model parameters-not customer data-are exchanged between nodes, maintaining data privacy while enhancing the robustness of the global model to detect AML. This ensures that financial institutions can collaboratively improve AML measures while adhering to strict privacy regulations.

Machine learning models to detect AML

Anti-money laundering is becoming an increasingly challenging problem due to the complexity and sophistication of modern financial crimes. Banks and financial institutions often operate in isolation, which limits the effectiveness of traditional machine learning models for detecting suspicious activities. However, if banks collaborate and learn from each other, they can significantly reduce the impact of money laundering activities. Federated learning offers a new way forward by enabling these institutions to collaboratively train a model without sharing sensitive customer data. This approach not only protects data privacy but also enhances the model's performance by aggregating insights from different banks, making it more difficult for money launderers to exploit system weaknesses.

The following ML models have proven to be highly effective when integrated with federated learning for AML detection.

Graph Neural Networks (GNN): Due to the interconnected nature of financial transactions, GNNs are especially effective for AML [15]. Money laundering schemes often involve intricate webs of transactions that are spread across multiple accounts and entities. GNNs are particularly powerful in this setting because they can model and learn from relationships between various entities, such as accounts, transactions and customer profiles.

When trained in a federated learning setting, GNNs benefit from the expanded graph data of multiple financial institutions, allowing them to identify complex money laundering patterns that might not be evident within a single institution's data. Studies indicate that GNN-based models can increase the detection accuracy by 15%-20% compared to conventional methods for financial crime detection [16]. The combination of GNNs and federated learning strengthens the model's capacity to detect hidden relationships, making it a superior choice for combating sophisticated AML schemes.

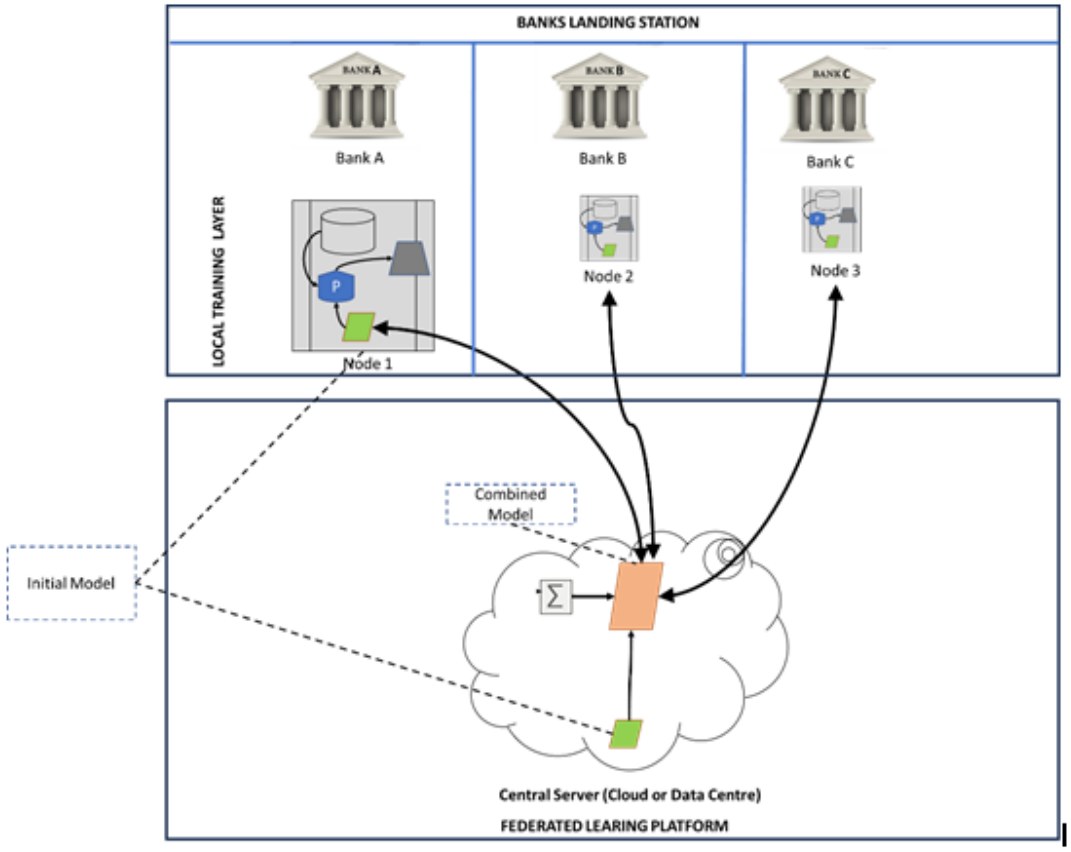

Figure 1: Federated learning data flow.

XGBoost for classification

For classifying transactions as either suspicious or non-suspicious, XGBoost is an ideal candidate due to its high accuracy and scalability. XGBoost is particularly effective in environments with noisy and imbalanced data, which is common in financial datasets. The gradient boosting approach used by XGBoost allows the model to learn from misclassified instances, further improving its performance.

When used in federated learning, XGBoost benefits significantly from the diverse datasets of multiple banks. It has been observed that federated learning with XGBoost can lead to prediction accuracies exceeding 90% for transaction classification [17,18]. By leveraging federated learning, the model can generalize better across different environments, effectively capturing variations in transaction patterns and ensuring high performance for AML.

The combination of federated learning with GNN and XGBoost provides a cutting-edge approach to AML, significantly enhancing both prediction accuracy to detect AML and privacy. Federated learning enables collaboration across institutions without compromising sensitive data, making these models a robust defense against evolving money laundering techniques. The prediction accuracy of these models, potentially can exceed 90%, highlights the promise of federated learning as an effective and privacy-preserving solution to financial crime detection.

Prototype

To evaluate the effectiveness of using federated learning for AML detection, we developed a federated learning platform using NVIDIA FLAIR and conducted multiple experiments to compare the performance of a traditional single-node setup with that of a federated learning setup. We utilized two key ML algorithms: GNN for capturing the complex interconnections between entities, and XGBoost for accurate transaction classification. The single-node setup simulated a traditional ML environment using data from one institution, while the federated learning setup involved collaboration between multiple financial institutions. The results indicated a significant improvement in model performance when utilizing federated learning, especially in accuracy and generalizability.

The datasets used in our experiments consisted of transaction histories, customer profiles, and suspicious activity reports (SARs) that were downloaded from Kaggle. Datasets were split into two parts representing two banks-Bank A and Bank B. This allowed the models to be trained with a more comprehensive set of financial behaviors.

We utilized NVIDIA FLARE as the federated learning framework, allowing multiple institutions to train a shared global model without centralizing sensitive data. Using FLARE, we integrated secure model aggregation and differential privacy techniques into the federated logistic regression model, thereby enhancing data privacy and regulatory compliance. FLARE's Proof-of-Concept (POC) mode was used to simulate real-world deployment on a local host with two simulated nodes representing two financial institutions, enabling us to conduct realistic experiments with ease. Furthermore, the FLARE simulator provided a quick response and debugging environment to refine our federated learning workflow.

The application was deployed on Amazon Web Services (AWS) with the following configurations, including NVIDIA GPU support:

- EC2 Compute Optimized (P4d.24xlarge) instances with NVIDIA A100 Tensor Core GPUs for backend and federated learning model training, providing the high computational power required for large-scale ML workloads.
- Memory Optimized (R5.large) instances with 2 vCPUs and 16 GiB memory for Elasticsearch, optimized for handling large datasets efficiently.
- General Purpose (T3.medium) instances with 2 vCPUs and 4 GiB memory for the front-end interface.

Storage was provided by General Purpose Solid State Drives (gp2) with sizes adjusted to meet data requirements. NVIDIA's CUDA and cuDNN libraries were employed for GPU acceleration, significantly speeding up the training of ML models within the federated learning framework. The entire NVIDIA FLARE environment was containerized using Docker, ensuring seamless integration with the other system components in AWS. The use of NVIDIA A100 GPUs combined with CUDA allowed us to achieve fast, efficient training performance, making the federated learning approach both scalable and effective.

Model accuracy evaluation process

To evaluate the model's accuracy in detecting money laundering patterns, we used a federated learning setup simulating two banks using FedAvg algorithm. In federated learning, the model is trained across multiple clients without sharing their actual data, which is crucial for preserving privacy. Below are the steps for each phase:

Model parameters broadcasting: The central server initially distributed the latest version of the model, along with training instructions; to both banks using NVIDIA FLAIR's broadcasting capabilities. This phase involved sharing model parameters (such as weights and biases) so that each bank had a synchronized starting point for local training while ensuring secure transmission.

Client training: Each bank refined the model locally using NVIDIA FLAIR's federated training framework. We utilized GNN for capturing the relationships between entities in the transactions and XGBoost for accurately classifying transactions based on these relationships. By using a combination of real and synthetic datasets, the banks ensured a comprehensive and diverse training set while maintaining data privacy. During this phase, each bank retained its raw data and only updated model parameters were shared with the central server. This local training phase ensured that sensitive information remained securely within each bank's premises.

Aggregation phase: Upon completion of local training, NVIDIA FLAIR facilitated the aggregation of model updates from both banks at the central server. In this phase, model parameters were combined without transferring any raw data, ensuring that confidentiality was preserved.

Model update: The central server employed the Federated Averaging (FedAvg) algorithm, integrated with NVIDIA FLAIR, to merge the updates received from both banks. FedAvg computed a weighted average of the model updates from all clients to produce an improved global model that leveraged the insights from each institution's data, maintaining privacy.

In our experiments, we used NVIDIA FLAIR's POC mode to simulate real-world deployment with two nodes, each representing a financial institution. This approach enabled us to conduct realistic experiments efficiently and allowed for debugging and

refinement of the federated learning workflow using FLAIR simulator.

The datasets originally included transaction records from 50 accounts at Bank A, which were expanded to over 150 synthetic datasets using data augmentation techniques. Similarly, Bank B's dataset of 50 accounts was expanded to an additional 100 synthetic datasets. These datasets contained both continuous features (such as transaction amounts) and categorical features (such as transaction types, customer demographics and flagged suspicious behaviors). For evaluation purposes, 67% of the data was used for training, while 33% was held back for testing, which allowed for robust model training across various scenarios, enhancing its capability to detect complex money laundering activities.

We conducted multiple experiments by incrementally increasing the number of communication rounds, which, in federated learning, involves the central server broadcasting the model, clients training it locally and clients sending back updates for aggregation using NVIDIA FLAIR. Increasing the number of communication rounds progressively minimized the model's loss function and enhanced prediction accuracy. This led to a steady improvement in accuracy, eventually stabilizing at a high level, effectively improving the model's ability to detect anti-money laundering activities.

RESULTS

Evaluation results: Model accuracy evaluation using a non-federated learning setup

To compare the federated learning setup with a traditional single-node setup, we simulated a centralized learning environment using the same FedAvg algorithm. While FedAvg is primarily designed for distributed learning, we applied it in a non-federated setup to explore:

- Whether the federated approach yields higher accuracy compared to a non-federated approach.
- The basic functionality of FedAvg without the complexities of distributed data sources.
- How accuracy changes as the number of simulated data partitions (nodes) increases.

We conducted three experiments within this setup to validate the accuracy of detecting anti-money laundering differences between the single-node and federated learning approaches.

Experiment A: Single node setup using data from Bank A

In this experiment, NVIDIA FLAIR was used to train a model for AML detection using only data from Bank A. We utilized two models in a complementary fashion: GNN were first used to analyze the complex interrelationships between different entities such as accounts and customers within the dataset, while XGBoost was subsequently used to classify individual transactions based on the insights derived from GNN. This approach combined the strengths of both models-leveraging GNN's capability to capture entity relationships and XGBoost's accuracy in classification.

The training was conducted over 14 epochs, with each epoch representing a complete pass through the entire local dataset, as illustrated in Figure 2. Initially, the model's accuracy was 71.11%, but after several communication rounds, the accuracy improved to 89.9%. This improvement demonstrates the model's steady learning progress; however, it also highlights the limitations of relying on data from a single institution, as it lacks the diversity needed to effectively generalize the anti-money laundering detection model across different environments. Using NVIDIA's GPU-accelerated infrastructure, with CUDA and cuDNN, expedited the training process, but the constrained dataset presented challenges in maximizing the model's detection capabilities.

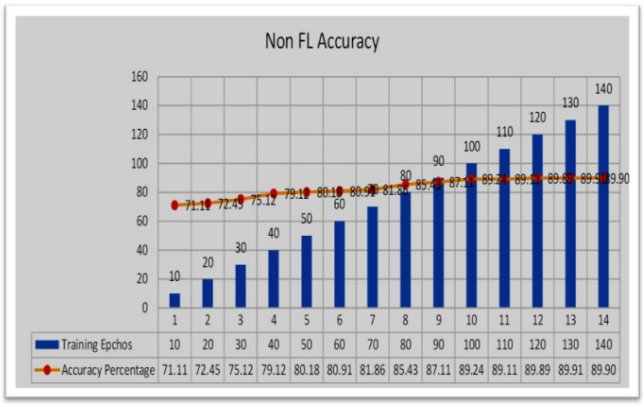

Figure 2: Accuracy results in single node setup using data from Bank A.

Experiment B: Single node setup using data from Bank B

In this experiment we trained the model with data from Bank B over 14 epochs, as shown in Figure 3. The accuracy of the model improved from 72.11% to 91.75% over the training process, following a similar trend observed with Bank A. This demonstrates that isolated training benefits the model, but there are clear constraints due to the limited data diversity available from a single institution. Despite leveraging NVIDIA's GPU-accelerated infrastructure, the lack of varied data restricted the model's ability to generalize effectively, underlining the need for more diverse datasets to improve the robustness of AML detection.

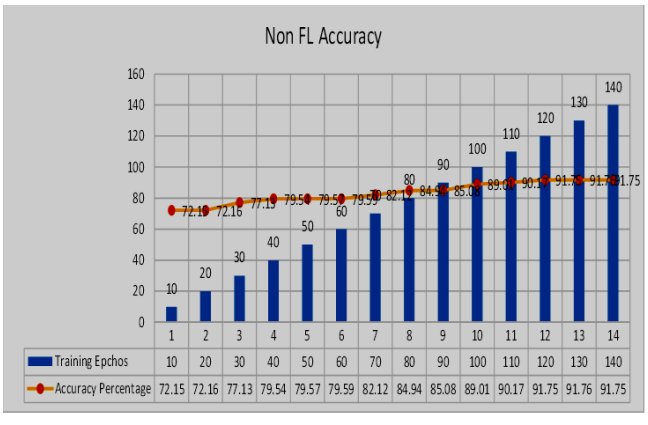

Figure 3: Accuracy results in single node setup using data from Bank B.

Experiment C: Combined data from Bank A and Bank B in a Federated learning setup

In this experiment, we trained the models using data from both Bank A and Bank B to collaboratively develop an anti-money laundering detection model. NVIDIA FLAIR facilitated this

collaboration by allowing multiple institutions to train a shared global model without sharing raw data, thereby maintaining data privacy and enhancing security. Rather than transmitting sensitive transaction data between banks, only the model parameters were exchanged, ensuring confidentiality. This distributed learning framework employed the FedAvg algorithm available in NVIDIA FLAIR to aggregate the model updates, ensuring that customer data remained securely within each bank while improving the model's performance through shared learning.

The use of NVIDIA A100 GPUs, along with CUDA and cuDNN for GPU acceleration, significantly improved training efficiency. As shown in Figure 4, the federated learning approach with NVIDIA FLARE led to substantial improvements in model accuracy compared to the single-node setup, where each bank trained the model independently. With the federated learning setup powered by NVIDIA technology, the initial accuracy was 78.2%, and after 14 communication rounds, it increased to 96.22%. This demonstrates the effectiveness of NVIDIA FLARE in leveraging diverse datasets from multiple institutions, achieving enhanced detection performance and accuracy while preserving data privacy (Table 1).

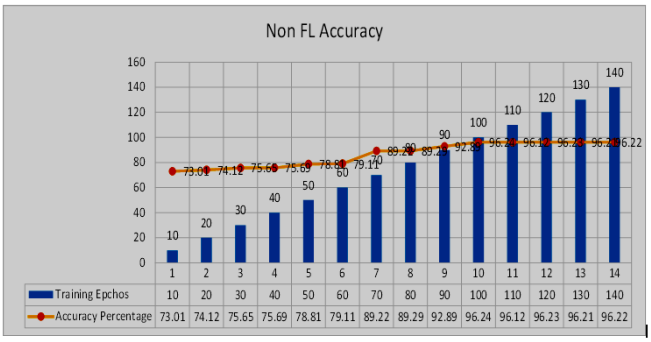

Figure 4: Accuracy results in single node setup combining data from Banks A and B.

Table 1: Provides a summary of the results from the experiments.

Experiment	Data source	Initial accuracy	Final accuracy
Experiment A	Bank A (Single Node)	71.11%	89.90%
Experiment B	Bank B (Single Node)	72.11%	91.75%
Experiment C (Federated Learning)	Banks A and B (Combined)	78.20%	96.22%

These results highlight that the federated learning setup outperformed single-node configurations by leveraging diverse data sources while ensuring privacy. The accuracy improvements in federated learning underscore the potential of NVIDIA FLARE to deliver more accurate and generalizable AML detection across multiple financial institutions.

DISCUSSION

This research presents the development of an advanced AML detection platform using NVIDIA FLARE for federated learning, incorporating GNN and XGBoost. The platform's primary objective is to facilitate collaboration among banks through a consortium, allowing institutions to share insights about AML threats without exposing sensitive data or compromising privacy. By enabling the sharing of knowledge without direct data exchange, the platform aims to prevent similar incidents across participating banks. Through this collaboration, the banking sector can develop a consolidated AML detection model that integrates the experiences of all institutions, thereby eradicating recurring money laundering patterns.

Federated learning with NVIDIA FLARE enhances detection capabilities by enabling financial institutions to collaboratively train a global model without centralizing data. Each bank trains the model locally, maintaining privacy while improving the system's ability to detect complex money laundering patterns. By using this platform, AML threats experienced by one bank can be proactively mitigated by others, making it an essential tool for reducing risks and costs related to money laundering.

GNNs and XGBoost models were used with NVIDIA FLARE. GNN were particularly effective in capturing the interconnected nature of financial transactions, identifying intricate patterns indicative of organized money laundering. XGBoost was utilized for classifying transactions as suspicious or non-suspicious, benefiting from its high accuracy and efficiency in handling noisy and imbalanced datasets. The federated learning approach enabled both GNN and XGBoost models to leverage diverse datasets from multiple banks, significantly improving the detection of sophisticated money laundering schemes.

The results indicated a significant improvement in model performance, especially in terms of accuracy, precision and recall. With the support of NVIDIA's GPU-accelerated technology, the federated learning approach achieved an overall accuracy of 96.22%, demonstrating its effectiveness compared to traditional, centralized training methods. By training on diverse datasets from multiple institutions, the system can detect complex cross-border money laundering schemes that might otherwise go unnoticed by individual banks. The collaboration enabled by federated learning resulted in a more comprehensive and robust model, reducing false positives and increasing adaptability to emerging laundering tactics.

The findings of this study demonstrate that federated learning with NVIDIA FLARE can significantly enhanced the performance of GNN and XGBoost while ensuring data privacy. This privacy-preserving approach not only improved the effectiveness of the AML model but also helped address regulatory requirements related to data sharing across institutions. NVIDIA's GPU-accelerated FLARE framework proved crucial in efficiently training the model, optimizing the time and resources required and enabling a scalable solution for combating money laundering at a global level.

CONCLUSION

In this paper, we have embraced the principle of shared learning, enabling banks to benefit from each other's experiences with anti-money laundering threats. By fostering collaboration, our platform ensures that knowledge about money laundering patterns is shared across institutions, ultimately preventing the recurrence of similar incidents and improving the overall resilience of the financial sector. The use of NVIDIA FLARE for federated learning, combined with GNN and XGBoost offers significant advantages for improving AML detection. By enabling multiple financial institutions to collaborate while maintaining strict data privacy, federated learning allows for the creation of

a more accurate and generalizable global model. This approach ensures that sensitive customer information remains protected, while the shared model benefits from the diversity of data from various banks, making it more effective in detecting complex money laundering patterns.

In conclusion, this study underscores the transformative potential of collaborative learning and the development of a shared model for detecting anti-money laundering patterns, leveraging NVIDIA's federated learning platform alongside GNN and XGBoost. By offering strong performance metrics, privacy-preserving features and scalability, the proposed system provides a robust and forward-looking solution that can be widely adopted in the financial sector to improve compliance and reduce the risks associated with money laundering. The integration of NVIDIA's GPU acceleration further enhances model training efficiency, making this approach a highly viable solution for addressing evolving AML challenges.

AUTHOR'S CONTRIBUTIONS

All authors read and approved the final manuscript. The corresponding author was responsible for the study's conception and design. Venkatesh Upadrista authored the first draft of the manuscript. All other authors have reviewed, read and approved the final version of the manuscript.

CONFLICT OF INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

1. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: system design. arXiv preprint. 2019.
2. McMahan HB, Yu FX, Richtarik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. NIPS. 2016:5-10.
3. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. InArtificial intelligence and statistics 2017:1273-1282.
4. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST). 2019;10(2):1-9.
5. Li PY, Chang TT, Kuo YC, Lin CY, Chang HY. Unveiling the black box: An XAI-based Anti-Money Laundering Model (AML). ICCE 2024:293-294. .
6. Hamid OH. Breaking through opacity: A context-aware data-driven conceptual design for a predictive anti money laundering system. IEEE-GCC Conference 2017:1-9.
7. Raj M, Khan H, Kathuria S, Chanti Y, Sahu M. The Use of Artificial Intelligence (AI) in Anti-Money Laundering (AML). ICSADL. 2024:272-277.
8. Wang SN, Yang JG. A money laundering risk evaluation method based on decision tree. ICMLC 2007;01:283-286.
9. Kute DV, Pradhan B, Shukla N, Alamri A. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. IEEE Access. 2021;9:82300-82317.
10. Thi MH, Withana C, Quynh NT, Vinh NT. A novel solution for anti-money laundering system. CITISIA. 2020:1-6.
11. El-Kilany A, Ayoub AM, El Kadi HM. Detecting suspicious customers in money laundering activities using weighted HITS algorithm. AIRC. 2024:112-117.
12. Oztas B, Cetinkaya D, Adedoyin F, Budka M. Enhancing transaction monitoring controls to detect money laundering using machine learning. ICEBE. 2022:26-28.
13. Cheng D, Ye Y, Xiang S, Ma Z, Zhang Y, Jiang C. Anti-money laundering by group-aware deep graph learning. IEEE TKDE. 2023;35(12):12444-12457.
14. Chen Z, Soliman WM, Nazir A, Shorfuzzaman M. Variational autoencoders and Wasserstein generative adversarial networks for improving the anti-money laundering process. IEEE Access. 2021;9:83762-83785.
15. Singha SP, Hossain MM, Rahman MA, Sharmin N. Investigation of graph-based clustering approaches along with graph neural networks for modeling armed conflict in Bangladesh. Int. J. Data Sci. Anal. 2024:1-7.
16. Benslimane S, Azé J, Bringay S, Servajean M, Mollevi C. A text and GNN based controversy detection method on social media. WWW. 2023;26(2):799-825.
17. Santoro D, Ciano T, Ferrara M. A comparison between machine and deep learning models on high stationarity data. Sci Rep. 2024;14(1):19409.
18. Sun B, Huang Y, Liu G, Wang W. Prediction of compressive strength of concrete under various curing conditions: a comparison of machine learning models and empirical mathematical models. Innov. 2024;9(7):262.