


# Detection Of Bank Transaction Anomalies Using Gradient Boosted Federated Learning

Rohan C 

*Dept. of Computer Science and Engineering*  
*PES University*  
Bengaluru, India  
chandrashekar.rohans@gmail.com

Rithvik Grandhi 

*Dept. of Computer Science and Engineering*  
*PES University*  
Bengaluru, India  
grandhirithwik@gmail.com

Rahul Roshan Ganesh 

*Dept. of Computer Science and Engineering*  
*PES University*  
Bengaluru, India  
rahulroshanganesh2002@gmail.com

**Abstract**—In today’s digital era, financial security, especially online banking has become a prominent security challenge to tackle. The accurate detection and labelling of bank transactions as fraudulent or not is crucial for safeguarding the integrity of our systems as well as maintaining customer trust. In this study, we attempt to introduce a novel approach to this problem. Here, we utilise the decentralised nature of the Federated Learning (FL) framework to enhance data security and privacy in our analysis of sensitive and confidential banking transaction data. Unique to this approach is our use of Gradient Boosting (GB) algorithms. GB algorithms are known for their efficiency in handling diverse and heavily imbalanced datasets. This approach to banking transactions data helps us to significantly improve the detection of anomalies, reducing false positives and increasing accuracy. By utilising this novel approach, we not only address data privacy concerns as the data stays with each bank client but also uncover a significant advancement in the efficiency of fraud detection in banking. The results achieved through this study clearly show a marked enhancement in the identification of anomalies and the efficiency in doing so. Paving way for a transformative potential for this method in our modern online banking security infrastructure.

**Index Terms**—federated learning, gradient boost, xgboost, decentralisation, machine learning, data privacy, banking

## I. INTRODUCTION

The current landscape of modern banking and associated payment transactions have become more digital. This shift brings new chances but also new problems, especially in keeping money safe. With online transactions now common, it is really important to have strong security to stop fraud. Banks and other financial organisations are always looking for better technology to protect their work and keep the trust of the customers. Paying more attention to security is a key way to deal with the changing methods of financial fraud and to make sure the financial system stays honest and reliable.

Federated Learning (FL) is an innovative approach that traces its roots back to the 2010s [1]. The original idea behind FL was to tackle issues related to data privacy and the decentralisation of machine learning. As time passed, FL has undergone considerable development. It has become a key technology, allowing various participants to jointly develop a shared predictive model. The unique aspect of this approach is that it keeps the training data on the user’s device, ensuring data privacy. This advancement has been particularly beneficial in industries such as Internet of Things (IoT) and Banking, where data security and privacy are of paramount importance.

Gradient Boosting models are quite precise and are known for dealing with complex, non-linear data [2]. They’ve been really effective in many areas, especially in finance. When we use these models in a Federated Learning setting, they become even more advantageous. The strength of Gradient Boosting, along with Federated Learning’s decentralised and privacy-focused approach, makes a strong combination. This helps in spotting unusual patterns in banking transactions. The mix of these two technologies lets the model learn from varied data from different places, all while keeping data privacy and security intact.

## II. BANKING AND FINANCIAL DATA SECURITY

### A. Current Mechanisms for Data Privacy and Security

In the realm of banking and financial services, the safeguarding of data privacy and security during machine learning and training activities is a critical concern. At present, a range of strategies and systems are in place to achieve these goals. This encompasses methods like established data encryption, secure multi-party computation, and differential privacy techniques. Typically, banks rely on centralized data repositories and processing frameworks, where rigorous data security measures are applied. The primary aim of these

frameworks is to shield confidential financial information from unauthorized access and breaches. Additionally, these systems are meticulously aligned with regulatory requirements such as the GPDA and CCPA, ensuring compliance and safeguarding client data. [3].

### B. Limitations

While current systems and methods have been implemented with care, they exhibit certain significant shortcomings. The practice of centralized data storage introduces the risk of single-point failures, leaving these systems more susceptible to complex cyber-attacks. This centralization necessitates the transfer of sensitive data from its original location, thereby heightening the risk of data privacy violations.

In the realm of machine learning, conventional techniques may not fully leverage the available vast data resources, primarily due to privacy concerns. This leads to the development of models that fall short in terms of accuracy and robustness. Moreover, these models often fail to adapt quickly to the continuously evolving tactics in financial fraud, as they are not updated frequently enough to keep up with these changes. [4]

Therefore, it becomes evident that adopting this approach not only fortifies data security and privacy but also markedly improves the accuracy and efficiency of fraud detection mechanisms within the banking and financial sectors.

### C. Gradient Boosting in a Federated Environment

Implementing gradient boosting within a federated learning framework marks a significant shift from conventional approaches. Federated learning distinguishes itself by enabling the development of machine learning models using multiple, geographically distributed datasets, while crucially maintaining data on local servers. This approach is particularly beneficial in handling sensitive financial information, as it avoids the need for data sharing or transferring, thus greatly minimizing the risks associated with centralized data storage and transfer. [5].

Decentralization of data processing through federated learning helps prevent single-point failures and large-scale data breaches. It adheres to strict data privacy norms, ensuring that the confidentiality of individual users is not compromised, which is especially important in light of rigorous regulatory standards. Additionally, by harnessing the collective power of distributed data, gradient boosting models in a federated learning environment can be trained more efficiently and updated in real-time [6]. This results in a system that is both dynamic and agile, with an enhanced capability to adapt quickly to emerging fraud patterns.

## III. DATASET

### A. Acquisition

In our research on detecting anomalies in bank transactions, we used a dataset from PaySim [7], a financial simulator designed by a leading finance company to replicate real-world transaction patterns. PaySim generates datasets from actual

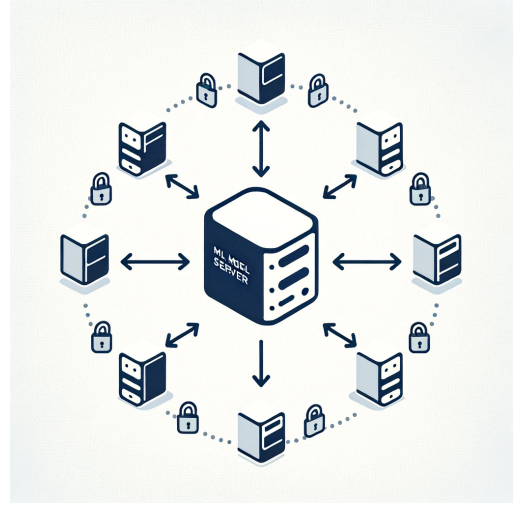


Fig. 1. Depiction of a Decentralized Model which offers Data Privacy

transaction data, capturing the intricate details and behaviours typical of genuine banking transactions. This method provided us with a rich and detailed dataset containing over 6 million records, effectively mirroring the complexities of banking activities, including both legitimate and fraudulent transactions. This ensures that our research findings are relevant and applicable to real-world scenarios. However, a significant challenge we faced was the severe imbalance in the dataset, where the majority of transactions were classified as non-fraudulent. This is because in real-world data, the detection and subsequent classification of transaction as fraud is relatively rare. Consequently, our model had to undergo significant adjustments to effectively address this issue.

### B. Attributes

The following are the attributes of the dataset:

- **step:** Maps a unit of time in the real world. In this case 1 step is 1 hour of time.
- **type:** CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
- **Amount:** Amount of the transaction in local currency.
- **nameOrig:** Customer who started the transaction.
- **oldbalanceOrig:** Initial balance before the transaction.
- **newbalanceOrig:** Customer's balance after the transaction.
- **nameDest:** Recipient ID of the transaction.
- **oldbalanceDest:** Initial recipient balance before the transaction.
- **newbalanceDest:** Recipient's balance after the transaction.
- **isFraud:** identifies a fraudulent transaction (1) and non fraudulent (0).

### C. Preprocessing

This dataset required to be preprocessed as XGBoost cannot deal with any values other than *int* and *float*. Hence, the

following processes were carried out on the dataset to ensure interoperability with XGBoost Algorithm.

TABLE I  
BEFORE PREPROCESSING

type	nameOrig	nameDest
PAYMENT	C1231006815	M1979787155
PAYMENT	C1666544295	M2044282225
TRANSFER	C1305486145	C553264065
CASH_OUT	C840083671	C38997010
DEBIT	C712410124	C195600860

- **Null Productions:** Removing all records where no actual customer or transaction takes place.
- **One Hot Encoding:** This is done on the attributes of *type* to convert the transaction type to respective class integers.
- **Tokenisation:** This is done on the attributes of *nameOrig* and *nameDest* to tokenise their unique name to an integer.

TABLE II  
AFTER PREPROCESSING

type	nameOrig	nameDest
0	170136	160296
0	21249	19384
1	181	21182
2	41554	299885
3	41720	10845

#### IV. GRADIENT BOOSTING

##### A. History and Benefits

Gradient boosting algorithms, a form of machine learning techniques, have been a focal point of research since their inception in the late 1990s. Pioneered by scholars like Jerome H. Friedman, gradient boosting evolved from the concept of boosting weak learners into a strong one. The fundamental principle is to iteratively improve predictions by focusing on the errors of previous models, refining the overall predictive accuracy. These algorithms have gained immense popularity due to their flexibility and effectiveness in handling various types of data, including non-linear, complex and imbalanced datasets [8] [9]. They are particularly renowned for their high predictive power and efficiency in classification and regression tasks, making them suitable for diverse applications, including the financial sector.

##### B. Advantages over Traditional Neural Networks

Gradient boosting algorithms and traditional neural networks are both highly effective in the machine learning domain, yet gradient boosting holds some notable benefits. Unlike neural networks, gradient boosting often demands less data preprocessing like normalization or scaling. This attribute makes them more adaptable in managing various data types, as they're less affected by the scale of input features. Particularly in structured and tabular data scenarios, common

#### Algorithm 1 XGBoost Algorithm

---

**Require:** Training data set  $D = \{(x_i, y_i)\}$ , a differentiable loss function  $l(y, \hat{y})$ , number of iterations  $N$ , learning rate  $\eta$

**Ensure:** A strong learner  $F(x)$

Initialize model with a constant value:  $F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n l(y_i, \gamma)$

**for**  $m = 1$  to  $N$  **do**

    Compute gradients:  $g_i = \partial_{\hat{y}_i} l(y_i, \hat{y}_i)$  for  $i = 1, \dots, n$

    Fit a new model  $h_m(x)$  to predict the gradients  $g_i$

    Update the model:  $F_m(x) = F_{m-1}(x) + \eta h_m(x)$

**end for**

$F(x) = F_N(x)$

---

in banking datasets, gradient boosting algorithms frequently surpass the performance of neural networks. Additionally, they offer greater interpretability, shedding light on which features are most significant and how decisions are made [10].

##### C. XGBoost and its Superiority

XGBoost (eXtreme Gradient Boosting), is known to be very robust in managing imbalanced datasets, which are commonly seen in sectors like banking fraud detection. This is largely due to its comprehensive hyperparameter tuning capabilities. For instance, it employs parameters such as *scale\_pos\_weight* to shift the algorithm's attention towards the minority classes, which is essential when pinpointing infrequent fraudulent activities [11]. Furthermore, adjusting the *max\_depth* is a strategic move to strike a balance between model complexity and the risk of overfitting. Meanwhile, fine-tuning the learning rate promotes steady and substantial learning. XGBoost also incorporates built-in regularisation terms, *lambda* and *alpha*, to curb overfitting which is a frequent challenge with imbalanced datasets. Through the meticulous adjustment of these hyperparameters, XGBoost proves to be exceptionally adept at unraveling the complexities of imbalanced data. This makes it a prime choice for uncovering nuanced patterns and irregularities in financial transactions.

#### V. FEDERATED LEARNING ENVIRONMENT

In conducting our research, we chose the Flower framework to set up our federated learning system due to its flexible nature and its ability to work with various machine learning algorithms. Flower is specifically designed to aid in creating and implementing federated learning systems. It provides a user-friendly platform for handling the distributed training process efficiently. A significant benefit of Flower is its wide compatibility with numerous machine learning frameworks and algorithms. This feature was particularly valuable in our project, as it facilitated the use of a gradient boosted federated learning model.

The framework's non-restrictive approach to machine learning algorithms meant we could easily incorporate the XGBoost algorithm. This integration did not require major changes or extensive customization in our federated learning setup.

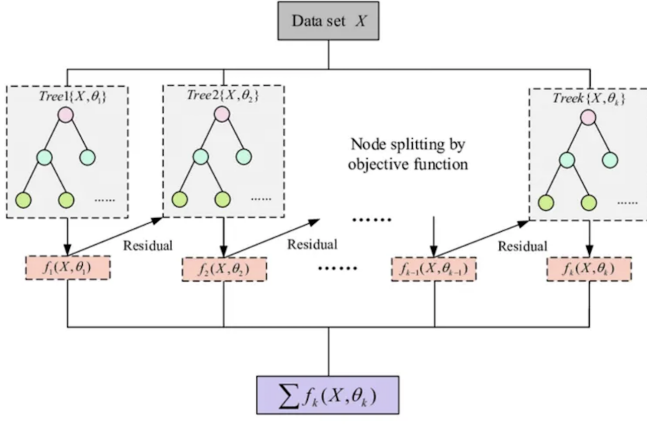


Fig. 2. Decision Tree convergence within XGBoost Algorithm [12].

Thanks to Flower's compatibility, integrating our chosen gradient boosting method was straightforward, ensuring effective management of the complex tasks involved in training and aggregating the model across multiple nodes [13].

By utilizing the flexible and adaptable capabilities of Flower, we could concentrate on enhancing our model's performance and precision in detecting anomalies in banking. This focus was possible without being hampered by the technical constraints typically associated with federated learning frameworks.

#### A. Setup of Environment

Federated Learning (FL) represents a groundbreaking method that prioritizes data privacy and security by distributing algorithm training across numerous independent nodes or devices. When establishing an FL system, we typically work with two kinds of nodes: the central server node and multiple client nodes. Each client node holds a portion of the overall data, ensuring that no single location contains the entire dataset. This division is essential for safeguarding data privacy because it reduces the risk of exposing sensitive information. The server node plays a pivotal role in orchestrating the learning process among these scattered datasets. It kick-starts the learning model and sends its parameters to the client nodes to initiate the training. This decentralized approach is fundamental in upholding data privacy, as it negates the necessity of moving sensitive data across the network [14].

#### B. Sever Node

The server node is crucial in the federated learning process. It kickstarts the cycle with a pre-trained model, dispatching the initial parameters to the client nodes. After the client nodes have fine-tuned their models using their unique datasets, they transmit their enhanced parameters back to the server. Here, the server undertakes the vital task of federated aggregation. This involves merging the parameters gathered from all the client nodes to refine the overall model. Subsequently, the enhanced model is sent back to the client nodes for additional

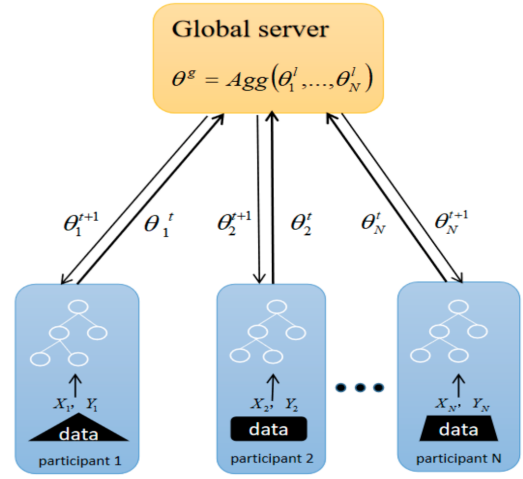


Fig. 3. Pictorial representation of Horizontal Federated Learning with XGBoost [6].

training. Through this repetitive process, the model undergoes continual enhancement and fine-tuning. Importantly, throughout this procedure, the actual data remains safely within the confines of the client nodes.

#### C. Client Node

In a federated learning setup, client nodes are integral to the training of the model using private data. Once the server node sends out the initial parameters, each client node begins training the model using its own dataset. This step is essential in the federated learning process as it enables the model to gain insights from a diverse array of data points, all while upholding the strictest standards of data privacy. Once the training on these local datasets is completed, the client nodes relay their updated parameters - which essentially encapsulate their learned insights - back to the server. It's important to note that during this exchange, it's only these parameters or insights that are shared, not the actual data. This approach not only safeguards sensitive information but also facilitates a robust, collaborative effort in model building. By leveraging distributed data, this method ensures that the model is developed in adherence to the highest data security and privacy protocols.

## VI. RESULTS

Upon initialising our federated learning framework with the below parameters, we obtain the following exemplary results.

- **Client Nodes:** Five Client Nodes.
- **Federated Learning Rounds:** 5 Federated Learning Rounds.

## VII. CONCLUSION

In summary, our research paper introduces a revolutionary method for identifying anomalies in bank transactions by combining the power of gradient boosting algorithms in a federated learning setting. This novel approach meets the urgent need for

TABLE III  
CLIENT 1

Metric	Training	Testing
Accuracy	0.999695	0.999657
Precision	0.984390	0.975100
Recall	0.773773	0.749383
F1 Score	0.866466	0.847469

TABLE IV  
CLIENT 2

Metric	Training	Testing
Accuracy	0.999650	0.999651
Precision	0.982966	0.981178
Recall	0.743182	0.740123
F1 Score	0.846419	0.843772

TABLE V  
CLIENT 3

Metric	Training	Testing
Accuracy	0.999659	0.999635
Precision	0.984909	0.984887
Recall	0.746758	0.724074
F1 Score	0.849458	0.834578

TABLE VI  
CLIENT 4

Metric	Training	Testing
Accuracy	0.999681	0.999672
Precision	0.979710	0.976228
Recall	0.769347	0.760494
F1 Score	0.861878	0.854962

TABLE VII  
CLIENT 5

Metric	Training	Testing
Accuracy	0.999678	0.999664
Precision	0.983748	0.979116
Recall	0.767910	0.752469
F1 Score	0.862531	0.850960

improved financial security in banking operations, presenting a solid answer to the challenges of data privacy and security inherent in machine learning applications. Utilizing a synthetic dataset created by PaySim, we have successfully modeled and examined realistic transaction behaviors, encompassing both genuine and fraudulent activities. The effectiveness of gradient boosting algorithms, especially XGBoost, is evident in their capacity to process imbalanced datasets efficiently through meticulous tuning of hyperparameters. This is essential for detecting nuanced patterns that may suggest fraudulent activities.

Moreover, the federated learning model reinforces data privacy and security. It operates with a server node that oversees the learning process, while client nodes conduct local model training on their respective data sets. This structure not only secures data confidentiality but also leverages the collective knowledge from dispersed data, thereby increasing the model's precision. The fusion of these technologies signifies a substantial leap forward in detecting financial fraud. It offers banks a scalable, effective, and secure strategy to protect against fraud. Our research makes a significant contribution to the expanding domain of financial security. It provides a thorough and practicable solution that banks can adapt and apply in different contexts to uphold the integrity and trustworthiness of the financial system.

## VIII. FUTURE SCOPE AND POTENTIAL

Our study, focusing on horizontal federated learning, lays a foundation for breakthroughs in banking security and anomaly detection. Horizontal federated learning involves collaboration across datasets that share similar features but differ in samples. This method is particularly advantageous for banks, allowing them to share insights while ensuring customer data privacy remains intact.

Looking ahead, an exciting avenue for research is vertical federated learning [15]. This technique is perfect for partnerships between different entities holding various types of data on the same individuals, for example, banks and retail organizations. It provides a more complete picture of customer behavior, uncovering complex fraud patterns that might be overlooked with horizontal federated learning.

Moreover, the prospects of multimodal federated learning are particularly intriguing [16]. In this approach, client nodes are trained on multiple models, each focusing on distinct aspects of data. This method can significantly boost the robustness of detection systems, enabling them to cover a broader spectrum of fraudulent activities by analyzing various data dimensions, from transaction amounts to geographical trends.

Transitioning to vertical and multimodal federated learning, combined with the integration of diverse data sources, has the potential to transform fraud detection in the banking industry. This shift promises to create more advanced, nuanced systems capable of identifying intricate fraud schemes, while still adhering to stringent data privacy regulations. Our research not only paves the way for cutting-edge fraud detection

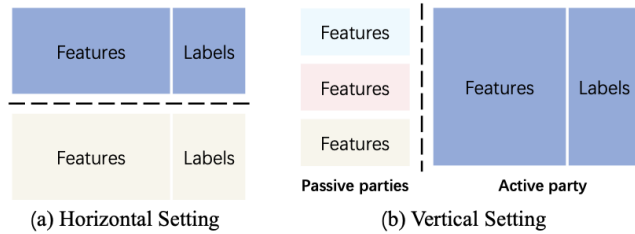


Fig. 4. Diagrammatic comparison of Horizontal and Vertical Federated Learning [13].

technologies but also underscores the immense potential of federated learning in tackling complex issues in the banking sector.

## REFERENCES

- [1] J. Konečný, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, 2015.
- [2] S. R. S. S. Ayachit, V. Patil, and A. Singh, "Competitive analysis of the top gradient boosting machine learning algorithms," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 191–196.
- [3] J. Serrado, R. F. Pereira, M. Mira da Silva, and I. Scalabrin Bianchi, "Information security frameworks for assisting gdpr compliance in banking industry," *Digital Policy, Regulation and Governance*, vol. 22, no. 3, pp. 227–244, 2020.
- [4] S. K. Lo, Q. Lu, H.-Y. Paik, and L. Zhu, "Flra: A reference architecture for federated learning systems," in *European Conference on Software Architecture*. Springer, 2021, pp. 83–98.
- [5] F. Yamamoto, S. Ozawa, and L. Wang, "efl-boost: Efficient federated learning for gradient boosting decision trees," *IEEE Access*, vol. 10, pp. 43 954–43 963, 2022.
- [6] X. Zhao, X. Li, S. Sun, and X. Jia, "Secure and efficient federated gradient boosting decision trees," *Applied Sciences*, vol. 13, no. 7, p. 4283, 2023.
- [7] K. Sengupta and P. K. Das, "Detection of financial fraud: comparisons of some tree-based machine learning approaches," *Journal of Data, Information and Management*, vol. 5, no. 1-2, pp. 23–37, 2023.
- [8] P. Zhang, Y. Jia, and Y. Shang, "Research and application of xgboost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, p. 15501329221106935, 2022.
- [9] C. Wang, C. Deng, and S. Wang, "Imbalance-xgboost: leveraging weighted and focal losses for binary label-imbalanced classification with xgboost," *Pattern Recognition Letters*, vol. 136, pp. 190–197, 2020.
- [10] F. Giannakas, C. Troussas, A. Krouska, C. Sgouropoulou, and I. Voyiatzis, "Xgboost and deep neural network comparison: The case of teams' performance," in *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*. Springer, 2021, pp. 343–349.
- [11] S. He, B. Li, H. Peng, J. Xin, and E. Zhang, "An effective cost-sensitive xgboost method for malicious urls detection in imbalanced dataset," *IEEE Access*, vol. 9, pp. 93 089–93 096, 2021.
- [12] R. Guo, Z. Zhao, T. Wang, G. Liu, J. Zhao, and D. Gao, "Degradation state recognition of piston pump based on iceemdan and xgboost," *Applied Sciences*, vol. 10, no. 18, p. 6593, 2020.
- [13] C. Ma, X. Qiu, D. Beutel, and N. Lane, "Gradient-less federated gradient boosting tree with learnable learning rates," in *Proceedings of the 3rd Workshop on Machine Learning and Systems*, 2023, pp. 56–63.
- [14] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão *et al.*, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.
- [15] T. Chen, X. Jin, Y. Sun, and W. Yin, "Vaf: a method of vertical asynchronous federated learning," *arXiv preprint arXiv:2007.06081*, 2020.
- [16] L. Che, J. Wang, Y. Zhou, and F. Ma, "Multimodal federated learning: A survey," *Sensors*, vol. 23, no. 15, p. 6986, 2023.