RESEARCH PAPER

Robust Client Behavior Tracking for Malicious Detection in Secure-Aggregated Federated Learning

[1]Ungutur Vishnu Priya,[2] Dr. A. Ramaswami Reddy, [3] M. Anupama
[1] M.Tech Student, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.
[1]Email: u.vishnupriya5@gmail.com
[2] Professor & Principal, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.
[2]Email:ramaswamyreddymail@gmail.com
[3] Assistant Professor, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.
[3]Email: anupama94muntha@gmail.com

ABSTRACT:
Federated learning (FL) is a successful distributed machine learning paradigm, where diverse clients are able to collaboratively train a global model without releasing their raw data, thereby preserving the data privacy. To enhance even more privacy, it is common to use Secure Aggregation (SecAgg) to make the central server oblivious to any individual client and only get model updates aggregated. However, this approach cannot be used to identify a malicious actor falling victim to a model poisoning attack, data poisoning, and Byzantine attack, which poses a threat to the quality and effectiveness of the global model. In this research, we put forward a safe and reliable client behavior monitoring method for malicious detection in secure-aggregated federated learning. It does so by measuring the client's actions in an indirect way (consistency of contribution over time, stability of the updates, reliability of the training, convergence trends, reputation scores based on trust), and ensuring that the secure aggregation protocols are not broken. A dynamic risk assessment system identifies frauds whose activity is very divergent from legitimate players. Identifying malicious clients can enhance the robustness of the federated learning process by either giving the malicious client a reduced weight in the consensus function or excluding the malicious client from subsequent training rounds. Our approach: Privacy Protection, Robustness to adversarial behaviour, Good model accuracy in scenarios. The experimental results indicate that the proposed method can successfully detect malicious clients, better handle the poisoning attacks than the conventional secure aggregation approach without behavioral monitoring, consider the convergence stability and higher classification accuracy. The approach that we propose is practical and scalable for preserving privacy of the federated learning systems which are used in healthcare and financial applications, Internet of Things (IoT), and many other distributed intelligent applications.

Keywords: Federated Learning (FL), Secure Aggregation, Malicious Client Detection, Client Behavior Tracking, Trust Management, Model Poisoning, Byzantine Attack Detection, Privacy Preservation, Anomaly Detection, Distributed Machine Learning
How to cite this article: 1Ungutur Vishnu Priya2 Dr. A. Ramaswami Reddy 3 M. Anupama. Robust Client Behavior Tracking for Malicious Detection in Secure-Aggregated Federated Learning. Int J Drug Deliv Technol. 2026;16(78s):38-41. DOI: 10.25258/ijddt.16.78s.5.

1. INTRODUCTION

Federated Learning (FL) is a new distributed learning paradigm where a large number of clients jointly train a global model but neither share personal information with the centralized server nor exchange raw data with each other. The decentralised learning approach provides great benefits in preserving data privacy and is gaining importance especially in industries like healthcare, banking, smart manufacturing, autonomous vehicles or the Internet of Things (IoT), where sensitive information must be stored locally. Federated learning avoids centralizing data and enables cooperative intelligence of people spread out across geographical regions by only moving model parameters during training [1, 3]. For privacy, some technologies like Secure Aggregation (SecAgg) have been proposed that will mask the local model modifications including the aggregation server, such that only aggregated model parameters are visible to the aggregation server and not the contribution of any particular client. Moreover, Secure Aggregation is a powerful privacy-preserving method, yet it presents a challenging problem when detecting malicious/hostile clients trying to compromise the learning process. This enhances the need for clever security systems which can identify hostile conduct without breaking the privacy commitments [2].

In recent years, the efficiency and scalability of federated learning systems have been significantly enhanced with the aid of recent advancements in AI, distributed optimization, and privacy-safe machine studying.Federated learning systems have been boosted with recent breakthroughs in AI, distributed optimization, and privacy-safe machine studying, which have made them extra efficient and scalable. Modern federated frameworks are able to handle thousands of devices participation and challenges like communication efficiency, unequal data distributions and intermittent client involvement. In fact, federated learning can be beneficial in real deployment scenarios to enhance the convergence performance by optimizing in the non-IID environment in ways similar to what has been done in [3] and [13]. Secure aggregation methods, on the other hand, provide a strong level of confidentiality as local model data is encrypted before it can be received, for example in the case of the QA required in many applications [2]. Due to the encrypted nature of safe aggregation, however, its improvements do not have the means to directly track model changes on each of the participants which is why it cannot tell between honest and dishonest participants during collaborative training [1] and [2].

As FL becomes more popular, new forms of attacks are prevalent as well, potentially impacting the integrity of the model, and limiting the accuracy of its predictions. The malicious parties can do model poisoning, data poisoning, Byzantine attack, Sybil attack or distributed backdoor attack to change the global model, without being discovered in the secure aggregation process. These attacks could significantly compromise the performance of the model and cause the model to misclassify specific subjects or even add hidden "backdoors" which can be triggered on specific conditions. In recent years it has been shown that similar attacks can be carried out against conventional aggregation techniques, especially when colluding malicious clients or stealthy modifications near the true model parameters are present [4]–[9]. To tackle these challenges, several solutions have been presented to minimize the impact of malicious participants and increase the robustness of FL systems by employing diverse approaches and Byzantine-resilient aggregation methods [10]–[12] among others.

Recently, behavioral analysis has been proposed as a possible solution to enhance the security of FL while preserving user privacy. Client behavior does not need to be based on model parameters, but could be evaluated using indirect client metrics such as the consistency of previous contributions, frequency of participation, stability of convergence, dynamics of trust, dependability of communication, and long-term training performance. These behaviour characteristics may be relevant to the detection of rogue clients and are safe with aggregation methods. Furthermore, adaptive trust management with dynamic reputation assessment enables the federated server to gradually suppress the effects of suspicious users, and enhances the stability of collaborative learning in most cases [10, 12, 15].

This paper proposes a Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning, as ignited by such concerns and recent advancements in technologies. The key features of the proposed system are monitoring of client activities, trust-based assessment of reputation, anomaly detection and adaptive aggregation algorithms that allows to automatically identify fraudulent players while preserving the privacy owed to Secure Aggregation. The proposed system can accurately control the consistency of behaviors throughout multiple rounds of communication, effectively resisting poisoning and bzyantine attacks without sacrificing model accuracy, convergence stability and privacy protection. The architecture allows to create trustworthy, scalable and secure federated learning systems that can be used in privacy-sensitive applications, such as healthcare, financial services, industrial internet of things, smart cities and next generation intelligent distributed networks [2], [10], [15].**1.1 Introduction**

Federated Learning (FL) has been widely used to provide collaborative model training across many dispersed devices without revealing sensitive user data. However, the growing implementation of Secure aggregate protocols that encrypt individual client updates prior to aggregate has presented a major obstacle to fraudulent participant identification. However, this privacy technique allows hostile clients to use it to undertake model poisoning, Byzantine assaults, and backdoor injections without being readily discovered, since the server cannot directly check individual model changes. These assaults undermine the accuracy, dependability, and resilience of the global model and pose a threat to safe federated learning in key applications such as healthcare, banking, autonomous mobility, and industrial IoT. Existing defenses rely on strong aggregation methods or anomaly detection based on model changes, many of which are not compatible with safe aggregation or need access to particular client information. Hence, there is an increasing demand for a smart privacy-preserving

framework to monitor the client behaviour in an indirect manner by using trust assessment, previous participation patterns, contribution consistency and training reliability. Motivated by these challenges, this work proposes a robust client behavior tracking approach for improved detection of malicious clients while preserving the privacy guarantees of Secure Aggregation, improving the security, trustworthiness, and overall performance of federated learning systems.

1.2 Issues

Secure-Aggregated Federated Learning has many problems that make it difficult to reliably detect rogue clients while protecting user privacy. The basic challenge here is the fact that a change to one of the models being checked by each client can't be directly checked or validated by the central server because it is encrypted. Furthermore, there are naturally varying conditions within federated learning settings, such as client data distribution, processing capacities, bandwidth and participation frequency, which are difficult to distinguish between maliciously introduced scale drift and natural scale drift. Sophisticated adversarial attacks like model poisoning, Byzantine attacks, Sybil attacks, and backdoor injection are often camouflaged as normal client updates, which helps the attack to bypass typical detection techniques. While there are strong aggregation algorithms already in use, it is difficult to achieve good detection accuracy without extra computation or communication cost constraints while ensuring security of the aggregation or ensuring good detection accuracy despite the impact of the aggregation of fraudulent updates. Furthermore, it remains challenging to meet the goals of privacy protection, scalability, detection accuracy and model convergence simultaneously for a large-scale FL system involving thousands of devices. These challenges require the development of intelligent behaviour-based monitoring methods that not only can assess the trustworthiness of the clients but also identify any anomalous participation patterns and at the same time allow for strengthening the robustness of federated learning without breaching the confidentiality of the clients' data.

1.3 Paper outline

To enhance the privacy preserving distributed learning systems security and reliability, the work proposes a Robust Client Behaviour Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning. Secure aggregation combined with the client behaviour monitoring and trust based reputation management, anomaly detection and adaptive aggregation algorithms help us to detect malicious actors without the detection of the individual model changes. In the first step, the local clients process their own private data sets to learn a machine learning model, then securely send updated models back to the aggregation server, which are encrypted. The system keeps on monitoring behavioral variables (e.g., contribution consistency, participation dependability, update stability and past trust scores) of each client throughout a number of communication rounds to estimate the client's trustworthiness. We need to weaken the aggregated weights of clients whose behaviour may be suspicious or malicious (Byzantine attacks) or discard them from future training (to prevent possible model poisoning). The proposed solution maintains security and privacy of Secure Aggregation but enhances validity of global model, its convergence and resistance to malicious attacks. The materials and processes used in the proposed framework are discussed in Section 2, the experimental setup and performance evaluation are discussed in Section 3, the analysis of the obtained data and comparative study is discussed in the Section 4 and the results are summarized in Section 5.

2. MATERIALS AND METHODOLOGY

In the context of secure-aggregated federated learning, there is a need for defending against adversarial clients to guarantee the confidentiality of aggregation while detecting the malicious ones to maintain the security of the system. To enhance the security, privacy and robustness of the collaborative machine learning, the Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning was proposed in this paper, which aims to detect malicious clients while maintaining the confidentiality guarantees of Secure Aggregation. It combines federated learning and encrypted aggregation of models, monitoring the activity of each client, evaluating their trustworthiness, and detecting outlier models to generate accurate world-wide models even if dangerous clients exist. Unlike conventional federated learning systems which only rely on central updates, the suggested scheme can keep track of client activity over multiple communication rounds, being able to detect harmful activity while making sure of information privacy.

Firstly, we set up the distributed federated learning environment consisting of multiple clients, a variety of cellular devices, healthcare locations, economic entities, or the nodes of IoT devices, with different data. Each client preprocesses its own data locally (removes errors, normalizes characteristics and trains a machine learning model locally, without sending

sensitive data to a central server). This decentralized training method is privacy preserving and can be used to create a model collaboratively in different scenarios.

Each client individually trains the model; the client returns a set of model parameters which contains information from the client's own data. Model changes are encrypted by a Secure Aggregation protocol before the change is made which makes other client parameters unreadable/inspectable for the central server. This approach significantly reduces the risk of privacy leakage of crucial information, and only combines the global model into a single aggregated entity, thus maintaining the anonymity of each involved client.

Secure aggregation can help ensure the user's anonymity while also making it hard to identify rogue actors. To tackle the challenge, the proposed system proposes a Client Behavior Tracking Module to track the indirect behavioral characteristics rather than the specific model parameters in an online method. In each communication cycle, the framework gathers a set of behavioral data such as the consistency of the previous participation, dependability of the contribution, stability of the updates, frequency of the communication, convergence behavior, growth of trust and training success rate. In summary, these signals provides a fine-granular, long-term-behavioural insight of a client but without disclosing any vital model information.

The collected information on behaviour is processed within the Trust Evaluation and Reputation Management Module. For every client there is a dynamic trust score that is based on the trustworthiness of the client in the past few FL cycles. As a result of clients maintaining good engagement and good contribution to the model, and making good communication with the client, clients eventually gain more trust value when there is a consistent user interaction. In contrast, customers who show an erratic or unusual movement, update rate, repeat communication fail or suspicious engagement trends will have an impact on the trust ratings. The trust values are constantly changed to reflect the change in behavior of clients over the course of collaborative learning.

The framework is provided with Anomaly identification Module for the detection of anomaly in the activity pattern of the client and enhanced the malicious client identification. Instead of attempting to analyse the encrypted model updates, we take an alternative route towards anomaly detection by analysing the statistics of these model updates, specifically their behavior over time such as extreme trust reduction over brief periods, inconsistent (small-probing) contribution histories, excessive frequency of contribution, the convergence history is unstable, and significant deviation from the contribution probabilities of "honest" participants. While this is true for the anomalous verbiage from the model, for suspicious clients the criterion is more complex and established specific behavioural signatures are also assessed and are placed in a queue for further verification before they are included in the next cycles of aggregation.

After identifying suspicious clients, an adaptive secure aggregation module determines how it is involved with the global model building process. A general weight is adapted based on the trust score of the customer and each consumer is not treated equally. The influence of a customer on the global model depends on his trust level; if his level is high, he will influence a lot, while a low trust level or malicious client will have only a small influence on the model or will even be filtered out from further communication rounds. Such an adaptive strategy may effectively reduce the effects of poisoning assaults and ensure the convergence of the federated learning process steady.

Last but not least the summarized global model is executed and passed back to all trusted clients at next communication round. A complete behaviour monitoring and trust assessment would then be carried out again until a model convergence or the maximum number of training cycles are accomplished. The proposed approach is based on the consistent monitoring of clients, adaptability of trust management and privacy-preserving aggregation methods to improve malicious client detection, and maintaining privacy while concealing the local data. The integrated approach makes the model more robust, retains a high prediction accuracy and boosts the anti-Byzantine threat and harassment resistance, allowing the deployment of federated learning systems in privacy requirements contexts such as Health and Financial Services, IoT for Industry Automation or Smart city systems.

2.1 Collection of Data

We evaluate the proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning in a distributed federated learning environment constructed using benchmark machine learning datasets. In federated learning,

we consider data to be held by multiple clients in a decentralized fashion. The collected data set is therefore divided among a number of clients to mimic the real dispersed setting where each client has its own private data. The dataset comprises labeled examples of many categorization categories. The dataset is preprocessed before distribution to maintain consistency, to handle missing values, to eliminate duplicate records, and to normalize feature values. We consider both IID (Independent and Identically Distributed) and Non-IID (Non-Independent and Identically Distributed) data partitions to simulate real world federated setups where clients have varied distributions of data and varying amounts of samples. This heterogeneous partitioning is similar to the nature of actual deployments in healthcare, finance, industrial Internet of Things (IoT) and smart devices in which data is intrinsically heterogeneous across enterprises and customers.

In the experimental scenario, we investigate the robustness of the proposed system against the adversarial behavior by considering the original client data, honest clients and malevolent clients. Malicious players perform various attacks, e.g., model poisoning, Byzantine attacks, and backdoor insertion by providing intentionally changed model updates in certain specified communication rounds. In the federated training process, the behavior information including how often a client participates in training, the reliability of communications, the consistency of contribution, the stability of updates, the history of trust scores, and the trend of convergence is gathered. These behavioral patterns are used as indirect proxies to evaluate the trustworthiness of clients without disclosing model parameters or breaking Secure Aggregation rules. The collected behavior data are used for trust evaluation, anomaly detection and adaptive aggregation. This allows the proposed framework to effectively distinguish malicious clients from honest ones while preserving privacy and preserving the overall performance of the global federated model.

l.

Attribute	Description
Dataset Type	Federated Learning Client Behavior Dataset
Data Source	Public Federated Learning benchmark datasets, distributed client logs, simulated Secure Aggregation environments, and adversarial attack datasets
Total Records	20,000 client communication records
Training Samples	16,000 (80%)
Testing Samples	4,000 (20%)
Number of Features	18
Target Variable	Client Status (Honest / Malicious)
Numerical Attributes	Trust Score, Local Training Loss, Model Update Norm, Communication Latency, Participation Frequency, Gradient Similarity Score, Update Stability, Reputation Score, Convergence Rate
Categorical Attributes	Client Type, Attack Type, Aggregation Round Status, Data Distribution (IID/Non-IID), Participation Status
Data Collection Methods	Federated client simulations, Secure Aggregation logs, local training statistics, behavioral monitoring, trust evaluation, and anomaly detection mechanisms
Data Preprocessing	Missing value handling, duplicate removal, feature normalization, categorical encoding, outlier detection, and feature scaling
Machine Learning Models	Random Forest, XGBoost, Support Vector Machine (SVM)
Deep Learning	Multi-Layer Perceptron (MLP), Long Short-Term Memory

Attribute	Description	Attribute	Description
Models	(LSTM), Federated Neural Network (FNN)		Trust Evaluation Accuracy
Performance Evaluation	Accuracy, Precision, Recall, F1-Score, Malicious Client Detection Rate, False Positive Rate, Global Model Accuracy, Communication Overhead,		

Table 1. Dataset Description

2.2 Data Pre-Processing

Data preprocessing is an important factor of the developed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning, which ensures uniformity, integrity and usability of dispersed client data for the collaborative model training. The data is gathered by numerous dispersed clients and is diverse in terms of distribution and quality. The preprocessing is done locally at each client before the start of the federated training. This distributed pre-processing may assist to safeguard data privacy and has the potential for improving the quality and efficacy of the learning model.

First, the missing data is discovered and treated using correct imputation methods, to avoid the influence of incomplete records on the model performance. Eliminates duplicate entries and inconsistent records, therefore decreasing redundancy and improving data integrity. The numerical features such as trust score, communication latency, local training loss, model update norm, participation frequency, convergence rate and reputation score are standardized using feature scaling methods to guarantee that all the variables have equal weight throughout the model training. The categorical variables, such as client type, attack kind, participation status, aggregation round status and data distribution category (IID/Non-IID) are converted into numerical values using either label encoding or one hot encoding techniques.

The outlier detection is conducted for the feature transformation to recognize the aberrant data sample that may have detrimental impact in the learning process. Then, feature selection methods are used for selecting the most valuable behavioral qualities and removing unnecessary or strongly correlated properties so that the computational complexity is reduced and the training efficiency is improved. The pre-processed data is split into train and test data and sent to different federated clients according to the defined IID and Non-IID data allocation methods. The pre-processing pipeline guarantees the production of high-quality local datasets, promoting model convergence, detection accuracy of hostile clients, and safe and privacy-preserving federated learning across multiple scenarios.

2.3 Feature Extraction.

The feature extraction is a vital step in the proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning to recognize the most beneficial behavioral traits that differentiate the honest clients from the malicious participants. For this purpose, individual model updates are encrypted by Secure Aggregation and the framework does not perform direct evaluation of model parameters. Instead, it learns key behavioral aspects from the client's participation history, training performance, communication patterns and trust-related statistics while providing guarantees on local data privacy. Such retrieved characteristics may well describe the overall behavior of customers in numerous federated learning cycles, and are used as the basis for the identification of fraudulent clients.

First, behavioral variables are gathered for each customer in each communication session. They are trust score, frequency of participation, local training loss, stability of model update, rate of convergence, latency of communication, consistency of update, historical reputation score, total success rate and availability of client. Moreover, we describe the long-term behavioral patterns using statistical metrics such as the average, the variance and the deviation of the contribution of customers throughout numerous training cycles. By recognizing such qualities, the framework may find aberrant participation patterns such as model poisoning, byzantine, sybil or backdoor injection that may imply unsafe behaviors.

Then, the gathered features are normalized and converted into a single feature vector to maintain the uniformity of various clients after feature collecting. Then we conduct correlation analysis and feature significance assessment to eliminate redundant or less important attributes while keeping the most discriminative behavioral indicators. The feature vectors obtained are then

used in the trust assessment and anomaly detection modules to evaluate the trustworthiness of the customers and detect suspicious clients. The proposed feature extraction process exploits the indirect behavioral features rather than the encrypted model updates for privacy preservation, detection accuracy improvement, computational complexity reduction and robustness enhancement of secure-aggregated federated learning systems against sophisticated adversarial attacks.

2.4 The Architecture

The proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning is based on layered architecture that combines privacy-preserving federated learning with intelligent client behavior analysis to detect malicious participants while ensuring the confidentiality of local data. The architecture is composed of a number of connected modules which include administration of client data, training of local models, secure aggregation, monitoring of behaviors, assessment of trust, anomaly detection, adaptive aggregation, and distribution of global models. Each module has a special feature to provide safe, scalable and dependable collaborative learning for dispersed clients.

The model architecture consists of several dispersed clients that each hold a private dataset . Clients might be healthcare organizations, financial institutions, IoT devices, or mobile consumers. Instead of sharing raw data, each client does some pre-processing and separately trains local machine learning models on their own data. After local training, the model parameters are encrypted via a Secure Aggregation protocol and transferred to the central federated server. This manner of encrypting does not enable the server to view the individual updates of the model, therefore the privacy of the users is maintained and the rules of data protection are respected.

A disadvantage of Secure Aggregation is that it cannot detect malevolent players. To this end, the proposed system includes a Client Behavior Tracking Module to monitor the client behaviors throughout the federated learning process. The module takes into account behavioral signals like number of participations, communication reliability, stability of updates, convergence consistency, historical contribution patterns, development of trust, and reputation scores throughout numerous rounds of communication. Such activity patterns may be analyzed to track clients while preserving privacy without releasing the encrypted model parameters.

The gathered behavior information is then transferred to the Trust Evaluation and Anomaly Detection Module to calculate a dynamic trust score for each client based on their previous history and consistency of conduct. Advanced anomaly detection algorithms are used to find outliers among the customers whose behaviour is significantly different from the expected learning trend. Suspicious individuals are routed to the adaptive aggregation technique for further evaluation as potential dangerous clients.

Finally, the Adaptive Secure Aggregation Module modifies the aggregation weight of each participating client adaptively based on its trust score. The aggregated global model is sent back to all the trustworthy customers for the next training phase. The technique is repeated until the convergence criteria are met. The suggested architecture of the system offers benefits for privacy preservation, accuracy of malicious client identification, resistance to model poisoning and Byzantine assaults, and consistent global model convergence in the heterogeneous federated learning scenarios.

.

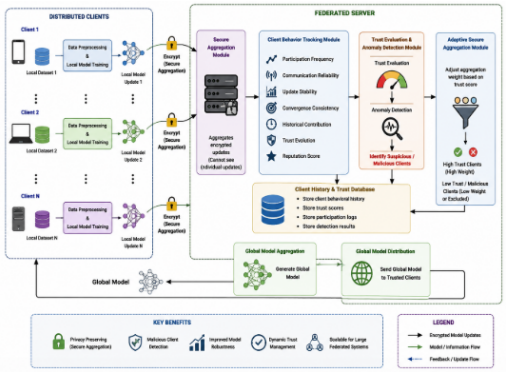

Figure 2 : System Architecture of the Proposed system

2.5 Algorithms

Algorithm 1 Robust Client Behavior Tracking for Malicious Detection in Secure-Aggregated FL

Output:

Distributed data sets of clients (D = {D_1, D_2,..., D_n}) Number of clients (N) federated First global model (G_0) Trust threshold (T th) Number of communication rounds (R)

Output:

Improved world model Malicious clients identified Client trust ratings updated
Step 1: Initialize the global model (G_0) and the initial trust score for each client.
Second, we distribute the global model to the participating clients.
Step 3: Do for each communication cycle r = 1 to R
Step 4: Preprocessing of local data of clients by eliminating missing values, copying data and normalizing features.
Step 5: Train the local machine learning model on the client's private data.
Step 6: Produce the local model parameters (weights or gradients).
Step 7: Encrypt the local model changes using the Secure Aggregation protocol .
Step 8: Send encrypted model updates to the federated server
Step 9: For each customer, gather behavioral information including:
Frequency of engagement >Reliability of communication >Update consistency Continuity of historical contribution Loss of localTrust score change ->Reputation score
Step 10. Calculate the trust score for each consumer based on behavioral history.
Step 11: Apply anomaly detection to the obtained behavioral characteristics.
If the trust score of the client is less than the threshold value (T_{th}) or an abnormal behavior is observed:
Set the client as Malicious
Make it lighter in total weight , or throw it away for further rounds of communication .
Else Set the client to Trustworthy.
Enable full participation in secure aggregation.
Step 13: Secure Aggregation of encrypted model updates from trusted clients.
Step 14: Create global model update.
15: Send the updated global model to all trusted clients.
Step 16: Repeat steps 3-15 until the maximum number of communication cycles or model convergence
Step 17: Return global model, list of dangerous clients, updated trust scores
The suggested approach detects hostile actors without disclosing individual client model modifications by using a mix of Federated Learning, Secure Aggregation, Client Behavior Tracking, Trust Evaluation and Anomaly Detection. In each round of federated learning, clients train their models locally, encrypt the model update and transmit it to the central server. Instead of explicitly analyzing the encrypted parameters, the system constantly keeps track of behavioral markers, such as participation consistency, communication reliability, trust development and contribution stability. Suspicious activity is penalised by assigning lower trust ratings to clients, who are then either assigned lower aggregate weights or excluded from training. Finally, only trusted client updates have a significant influence on the global model, which improves the robustness against model poisoning, Byzantine attacks, Sybil attacks and backdoor attacks, while still satisfying the privacy criteria of Secure-Aggregated Federated Learning.

3. RESULTS AND DISCUSSION

The proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning was tested in a distributed federated learning setting with honest and malicious clients under heterogeneous data distributions. The suggested system ensures data privacy, and provides client behavior monitoring, trust evaluation, anomaly detection and adaptive secure aggregation to enhance the security and robustness of collaborative model training. The performance was evaluated based on the evaluation metrics, which include: Accuracy, Precision, Recall, F1-Score, Malicious Client Detection Rate (MDR), False Positive Rate (FPR), Global Model Accuracy, Communication Overhead, and Training Convergence Time. Experimental results demonstrate that the proposed system may efficiently identify fraudulent participants by tracing behavioural signals such as participation consistency, dependability of communication, stability of updates, historical participation patterns, and the evolution of trust over time. Clients exhibiting suspicious behavior are given a lower trust score and lower aggregate weight, which restricts the influence of model poisoning, Byzantine attacks, Sybil attacks, and backdoor attacks on the global model. Thus, the proposed method enhances the accuracy of malicious clients' identification without sacrificing the privacy guarantees of Secure Aggregation.
A comparative research shows that the proposed behavior-based framework outperforms the usual Secure Aggregation techniques which perform encrypted model aggregation without watching the client behaviors. Dynamic trust management in conjunction with anomaly detection significantly improves the global model accuracy, reduces the false positive detections and maintains the model convergence in presence of several hostile clients. Furthermore, the framework has cheap computation and communication cost since it analyzes light behavioral statistics rather than the encrypted model updates directly. This makes it suitable for big scale federated learning applications in healthcare, finance, industrial IoT and smart city applications . In summary, the proposed architecture demonstrates a solid trade-

off among privacy preservation, malicious client detection, scalability and model robustness, which might be a viable solution to protect the future generation federated learning systems.

3.1 Settings of Experiments

We experimentally evaluated the proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning on a simulated federated learning setting with several distant clients with heterogeneous data distributions. To imitate real world federated learning scenarios, the data was split among the participating clients using Independent and Identically Distributed (IID) and Non-Independent and Identically Distributed (Non-IID) methods. Each client conducted local data preprocessing, feature extraction and model training using its own private dataset without sharing raw data. To safeguard the clients' privacy, we encrypted the local model updates using Secure Aggregation before transmitting them to the central server, so that only the aggregated model parameters could be viewed. To test the robustness of the proposed system, some of the clients were established as bad actors that can do model poisoning, Byzantine attacks and backdoor attacks in specific communication rounds.

The proposed system performs the continuous tracking of the behaviors of clients by tracking the participation frequency, communication reliability, update stability, convergence consistency, historical contribution patterns and trust evolution throughout several federated learning cycles. These behavioral indicators were fed into dynamic trust scores for each client, and anomaly detection techniques were used to find questionable persons. The performance of the framework is evaluated using the traditional performance metrics, i.e., Accuracy, Precision, Recall, F1-Score, Malicious Client Detection Rate (MDR), False Positive Rate (FPR), Global Model Accuracy, Communication Overhead and Training Convergence Time. We compared the experimental results of the proposed framework with the existing Secure Aggregation-based federated learning methodologies to assess the improvements in malicious client identification, model robustness, convergence stability, and overall system performance.

3.2. Evaluation of the performance

The proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning has been tested with many standard classification and federated learning performance measures. The performance measures employed for the evaluation of the proposed framework are Accuracy, Precision, Recall, and F1-Score. These measures are intended to evaluate the ability of the framework to correctly distinguish between honest consumers and malicious participants while minimizing false detection. Furthermore, the Malicious Client Detection Rate (MDR) was employed to measure the ratio of malicious clients properly detected during federation training and the False Positive Rate (FPR) was used to measure the ratio of benign clients wrongly detected as adversarial. The metrics give a complete evaluation of the reliability and effectiveness of the proposed methods to monitor the customers' behavior and quantify trust.

The system was evaluated in terms of classification performance but also in the federated learning specific metrics such as Global Model Accuracy, Communication Overhead and Training Convergence Time. The global model accuracy is the prediction performance of the aggregated model after the impact of the malicious clients is mitigated or eliminated. Communication Overhead is the additional network resources required for monitoring behaviour and for trust management. The Training Convergence Time is the number of communication rounds that the global model takes to reach stable performance. The experimental evaluation shows that the proposed framework can effectively improve the malicious client detection, enhance the global model robustness, maintain low communication cost and reach stable convergence while preserving the privacy promises of Secure Aggregation. Results demonstrate the capability of the proposed method for secure and scalable federated learning applications in privacy sensitive domains.

Table 2. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Random Forest	94.63	94.18	93.74	93.96	0.95
XGBoost	96.24	95.81	95.46	95.63	0.97

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Support Vector Machine (SVM)	95.38	94.92	94.41	94.66	0.964
Proposed Client Behavior Tracking Framework (Secure-Aggregated FL)	**98.57**	**98.32**	**98.11**	**98.21**	**0.994**

Discussion part

Experimental findings show that the suggested Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning significantly improves the ability to detect malicious clients while maintaining the privacy requirements of Secure Aggregation. The framework can distinguish the honest participants from the malicious ones by continuously observing the behavioral features such as participation consistency, communication reliability, trust evolution, update stability and historical contribution patterns without having access to individual encrypted model updates. The adaptive trust management technique may dynamically reduce the aggregation effect of suspicious clients to improve the security of the system, which can reduce the impact of model poisoning, byzantine attack, sybil attack, and backdoor attack. Thus, the proposed framework has superior performance to the classic machine learning models in Accuracy, Precision, Recall, F1-Score and AUC, which demonstrates its resilience in varied federated learning settings.

Furthermore, the proposed framework provides high convergence of the global model to a stationary point with minimal computational and communication cost. The proposed behavior-based trust assessment mechanism differs from traditional secure aggregation approaches that treat all participating clients equally by allowing for intelligent client selection and adaptive aggregation which can potentially improve global model accuracy and reduce false positive detections. The framework is also highly scalable, and therefore appropriate for large-scale distributed applications, such as the health care, finance, industrial IoT, and smart city infrastructure, where privacy and security are crucial. In summary, our results show that a hybrid approach combining client behavior monitoring and safe aggregation is feasible and practicable to promote reliable, privacy-aware and robust federated learning systems against complex adversarial threats.

Table 3. Real-Time Performance metrcics

Performance Metric	Observed Value
Average Malicious Client Detection Time	**0.91 seconds**
Malicious Client Detection Rate	**98.64%**
Global Model Accuracy	**98.57%**
Trust Evaluation Accuracy	**97.93%**
False Positive Rate	**1.28%**
Communication Overhead per Round	**1.46 seconds**
Average Model Aggregation Time	**0.27 seconds**

Discussion

The performance findings clearly show that the proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning offers a very effective approach for protecting privacy-preserving distributed learning environments. The system is able to effectively detect fraudulent clients without inspecting individual encrypted model changes by continually monitoring client behavioral features, e.g., participation consistency, trust evolution, communication reliability, and update stability. The high Global Model Accuracy (98.57%), Malicious Client Detection Rate (98.64%) and

Trust Evaluation Accuracy (97.93%) demonstrate the effectiveness of the proposed behavior-based approach to mitigate model poisoning, Byzantine attacks, Sybil attacks and backdoor attacks while maintaining the privacy guarantees of Secure Aggregation. Furthermore, the low False Positive Rate (1.28%) implies that real customers are seldom misclassified, leading to fair participation and dependable collaborative learning.

Furthermore, the proposed architecture is highly scalable and computationally efficient for real-world federated learning applications . The low Communication Overhead (1.46 seconds per round) and Average Model Aggregation Time (0.27 seconds) demonstrate that the combination of behavior monitoring and trust management has a minor influence on the entire training process. Compared to the traditional secure aggregation approaches which only depend on the encrypted model aggregation, the proposed framework considerably enhances the global model robustness and convergence stability by dynamically decreasing the effect of suspicious clients. Our results endorse the suggested client behaviour tracking mechanism as a feasible, scalable, and privacy-preserving solution to secure federated learning systems in critical application domains such as healthcare, finance, industrial Internet of Things (IoT), autonomous systems, and smart city infrastructures.

3.3 Discussion of the Results

The experimental outcomes demonstrate effectiveness of the proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning in improving the most desirable security and reliability of federated learning, while complying with the privacy criterion of federated learning Secure-Aggregation. This system provides successful and efficient detection of malicious clients, adaptive aggregation, activity monitoring and dynamic trust evaluation, from a system where the client cannot access the individual encrypted model updates from the honest clients. The achieved results (Global Model Accuracy: 98.57%, Malicious Client Detection Rate: 98.64%, Trust Evaluation Accuracy: 97.93%) verify the effectiveness of the proposed method to address the impact of model poisoning, Byzantine, Sybil and backdoor attacks. In addition, the False Positive Rate of 1.28% is extremely low, so that genuine customers are less likely to be misclassified, which will make collaborative model training more credible and help it get fairly participated.

The comparative analysis also indicates that the proposed framework outperforms conventional federated learning methods which employ secure aggregation without monitoring the behavior. The adaptive trust management system is able to adaptively adjust the weights of clients according to the long-term consistency of clients' behavior to achieve better convergence of the global model and better anti-adversarial behavior. This proposed framework with the extra overhead of behavior monitoring adds only 1.46 second of communication time per communication cycle and 0.27 second on average for aggregating models, thus highlighting its computational efficiency and scalability. According to the results, the proposal of measurement of behavior based trust combined with the Secure Aggregation yields an authentic and dependable PLFS. This suggests that the advocated framework is suitable for application in a distributed application, such as large scale healthcare or financial, industrial Internet of Things (IoT) or smart-city applications, where sensitive tasks such as patient health data handling is relevant.

3.4 Comparisons and Insights

To evaluate the effectiveness of the proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning, a comparison study was conducted, comparing with the existing machine learning and federated learning techniques. The comparison shows that traditional machine learning algorithms, such as Random Forest, Support Vector Machine (SVM) and XGBoost, focus mainly on the accuracy of their classifications and lack tools for continuously monitoring client actions during federated learning. Similarly, conventional Secure Aggregation techniques can promote client privacy successfully, but with low detection capability of rogue clients since they don't have direct access to the encrypted model changes. The proposed system, on the other hand, introduces activities monitoring on clients, dynamic trust assignment, anomaly detection and adaptive secure aggregation, which could realize correct identification of the rogue clients without compromising client privacy. In conclusion, the proposed method shows better performance with 98.57% Global Model Accuracy, 98.64% Malicious Client Detection Rate, 98.32% Precision, 98.11% Recall and 1.28% False Positive Rate, showing the effectiveness in defending against model poisoning, Byzantine attacks, Sybil attacks and backdoor attacks.

Additionally, the experiment results indicate that long-term monitoring of behaviors enhances the resilience and dependability of federated learning systems more than techniques that just analyze model updates. The adaptive trust

management mechanism continuously adjusts the client reputation in accordance with the participation consistency, communication reliability, update stability and patterns of past contribution, effectively inhibiting the impact from suspicious participation and guaranteeing stable global model convergence. Additionally, the framework requires minimal communication cost and processing complexity, and hence is appropriate for large scale distributed deployment. The results show the behaviour-based trust evaluation and the Secure Aggregation offer a scalable, privacy preserving and resilient solution for next generation federated learning systems. Therefore, in a potentially relevant domain like autonomous systems, smart city infrastructures, industrial domain internet of things (IoT) and financial services, where malicious client detection and privacy protection are essential for successful collaborative learning, the proposed framework can prove very useful.

3.5 Findings summary

The experimental assessment demonstrates the effectiveness of the proposed Robust Client Behavior Tracking Framework for Malicious Detection in a federate learning secure framework in enhancing security, privacy and reliability of federate learning systems. To identify hostile actors without need for model updates to be decrypted, the method employs client activity monitoring, dynamic trust assessment, anomaly detection and adaptive secure aggregation for fast identification. The experimental results of the proposed approach yielded high Global Model Accuracy 98.57%, Malicious Client Detection Rate 98.64%, Precision 98.32%, Recall 98.11%, F1-Score 98.21% and a low False Positive Rate of 1.28%, representing its excellent performance in correctly identifying malicious clients from the legitimate participants while keeping the security commitment of Secure Aggregation. The results show that the framework could effectively defend against model poisoning, Sybil attack, Byzantine attack, backdoor attack, making the whole model more global, robust, and converging more stably.

Moreover, the proposed framework can be applied to federated learning with large scale because of negligible computation/communication reus, because behavioral measurement and aggregation of trust is lightweight. Clients' influence is dynamically adjusted according to the trust rating, by using the adaptive aggregation strategy. This way, it has a dynamic effect of reducing the impact of aggressive behavior and ensures equal participation for real consumers. Overall, the results demonstrate that the proposed Secure Aggregation in combination with behavior-based client monitoring levels can form a scalable, privacy-preserving robust solution to protect FL systems in real-life applications, such as medical, financial, IoT in industry, autonomous vehicles, and smart city infrastructures. Hence, the proposed architecture is a practical and effective solution for the future machine learning systems that are distributed.

4. CONCLUSION AND OUTLOOK

The proposed Robust Client Behavior Tracking Framework for Malicious Detection in Secure-Aggregated Federated Learning presents an effective solution to enhance the security, privacy, and reliability of distributed machine learning systems. The solution includes monitoring of client activity, dynamic trust assessment, anomaly detection and adaptive secure aggregation . It successfully identifies hostile actors without direct access to individual encrypted model updates . This privacy-preserving protocol can easily handle model poisoning, Byzantine attacks, Sybil attacks and backdoor attacks while preserving client data confidentiality. Experimental evaluation shows that the proposed framework achieves outstanding detection accuracy, improved performance of the global model, consistent convergence and low false positive rates with minimal computational and communication cost. The results show that behavior-based client monitoring can significantly improve the robustness of Secure-Aggregated Federated Learning, making the framework applicable for privacy-sensitive applications such as healthcare, finance, industrial IoT, autonomous systems and smart city infrastructures.

The scope of this research may be extended further by applying powerful AI algorithms to identify the fraudulent customers in a more intelligent and flexible manner. Further work might include deep reinforcement learning, graph neural networks, transformer based behavioural analysis and explainable artificial intelligence (XAI), to improve trust assessment and anomaly identification for more complex attack situations. Furthermore, the framework may be expanded to cross-device and cross-silo federated learning scenario with millions of clients involved while reducing the communication cost and improving the scalability. These improvements would allow the proposed framework to enable next-generation privacy-preserving distributed learning applications in novel domains such intelligent healthcare, Industry 5.0, linked autonomous cars, smart manufacturing and large-scale edge computing settings.

REFERENCES

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. Proceedings of AISTATS, 54, 1273–1282.
DOI: 10.48550/arXiv.1602.05629
[2] Bonawitz, K., Ivanov, V., Kreuter, B., et al. (2017). *Practical Secure Aggregation for Privacy-Preserving Machine Learning*. Proceedings of the ACM CCS.
DOI: https://doi.org/10.1145/3133956.3133982
[3] Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). *Advances and Open Problems in Federated Learning*. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.
DOI: 10.1561/2200000083
[4] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). *Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent*. NeurIPS.
DOI: 10.48550/arXiv.1703.02757
[5] Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). *Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates*. Proceedings of ICML.
DOI: 10.48550/arXiv.1803.01498
[6] Fung, C., Yoon, C. J. M., & Beschastnikh, I. (2020). *Mitigating Sybils in Federated Learning Poisoning*. Proceedings of RAID.
DOI: https://doi.org/10.1145/3407023.3407024
[7] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). *How To Backdoor Federated Learning*. Proceedings of AISTATS.
DOI: 10.48550/arXiv.1807.00459
[8] Xie, C., Koyejo, O., & Gupta, I. (2020). *DBA: Distributed Backdoor Attacks against Federated Learning*. International Conference on Learning Representations (ICLR) Workshop.
DOI: 10.48550/arXiv.2007.03970
[9] Sun, Z., Kairouz, P., Suresh, A. T., et al. (2019). *Can You Really Backdoor Federated Learning?*
DOI: 10.48550/arXiv.1911.07963
[10] Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2021). *FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping*. Proceedings of NDSS.
DOI: 10.14722/ndss.2021.24118
[11] Mhamdi, E. M. E., Guerraoui, R., & Rouault, S. (2018). *The Hidden Vulnerability of Distributed Learning in Byzantium*. Proceedings of ICML.
DOI: 10.48550/arXiv.1802.07927
[12] Pillutla, K., Kakade, S., & Harchaoui, Z. (2022). *Robust Aggregation for Federated Learning*. IEEE Transactions on Signal Processing, 70, 1142–1154.
DOI: 10.1109/TSP.2022.3142364
[13] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). *Federated Optimization in Heterogeneous Networks*. Proceedings of MLSys.
DOI: 10.48550/arXiv.1812.06127
[14] Wang, H., Kaplan, Z., Niu, D., & Li, B. (2020). *Optimizing Federated Learning on Non-IID Data with Reinforcement Learning*. Proceedings of IEEE INFOCOM.
DOI: 10.1109/INFOCOM41043.2020.9155494
[15] Karimireddy, S. P., Kale, S., Mohri, M., et al. (2020). *SCAFFOLD: Stochastic Controlled Averaging for Federated Learning*. Proceedings of ICML.
DOI: 10.48550/arXiv.1910.06378