Secure and Privacy-Preserving Federated Learning for Connected and Automated Vehicles

Final Report

by

M. Hadi Amini
Florida International University

Mansoureh Jeihani, Morgan State University
Kemal Akkaya, Selcuk Uluagac, Farhad Shirani, Florida International University

July 2025



NATIONAL CENTER FOR TRANSPORTATION CYBERSECURITY AND RESILIENCY (TraCR)



DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by the National Center for Transportation Cybersecurity and Resiliency (TraCR) under Grant No. 69A3552344812 and 69A3552348317 which is headquartered at Clemson University, South Carolina, USA, from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

Non-exclusive rights are retained by the U.S. DOT.

CONTACTS

For more information:

M. Hadi Amini
11200 SW 8th Street, CASE 354
Miami, FL 33199
(305) 348-9936
amini@cs.fiu.edu

TraCR
Clemson University
One Research Dr
Greenville, SC 29607
tracr@clemson.edu

ACKNOWLEDGMENT

This work is based upon the work supported by the National Center for Transportation Cybersecurity and Resiliency (TraCR) (a U.S. Department of Transportation National University Transportation Center) headquartered at Clemson University, Clemson, South Carolina, USA. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of TraCR, and the U.S. Government assumes no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No. 3	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Secure and Privacy-Preserving Federated Learning for Connected and Automated Vehicles		**5. Report Date:** July 2025 **6. Performing Organization Code:** N/A
7. Author(s) M. Hadi Amini, Ph.D.; https://orcid.org/0000-0002-2768-3601 Mansoureh Jeihani, Ph.D.; https://orcid.org/0000-0001-8052-6931 Kemal Akkaya, Ph.D.; https://orcid.org/0000-0002-7103-4545 Selcuk Uluagac, Ph.D.; https://orcid.org/0000-0002-9823-3464 Farhad Shirani, Ph.D.; https://orcid.org/0000-0003-1316-3899		**8. Performing Organization Report No.** 3
9. Performing Organization Name and Address National Center for Transportation Cybersecurity and Resiliency (TraCR), Clemson University, 414 A One Research Dr, Greenville, SC 29607 Florida International University, 11200 SW 8th Street, Case 354, Miami, FL 33199		**10. Work Unit No.** N/A **11. Contract or Grant No.** 69A3552344812 and 69A3552348317
12. Sponsoring Agency Name and Address U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology, 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Report, 01/01/2024 - 05/31/2025 **14. Sponsoring Agency Code** OST-R
15. Supplementary Notes Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.		

16. Abstract

Federated Learning (FL) is a decentralized machine learning technique that creates a global model by aggregating local models from multiple edge devices without accessing their local data . However, due to the distributed nature of FL, there is a larger attack surface, making cyber-attack detection and defense challenging. Although prior work has developed various defense strategies to address security issues in FL settings, most approaches fail to mitigate cyber-attacks due to ~~the diverse~~ characteristics of the attack, edge devices, and data distribution. To address this issue, this paper develops a hybrid privacy-preserving algorithm that integrates central differential privacy (CDP) with a Preaggregation Similarity Measurement (PA-SM) technique to safeguard FL methods against malicious attacks in Intelligent Transportation Systems, considering object detection as a downstream machine learning task. This algorithm enables edge devices, such as autonomous vehicles (AVs) and roadside units to collaboratively train their models while maintaining the privacy of their respective data. By combining general DP and PA-SM, the algorithm provides robust security against data poisoning-based model replacement and inference attacks throughout the training phase. We evaluated our model using the CIFAR-10 and LISA Traffic Light datasets , demonstrating its ability to mitigate malicious attacks with minimal impact on the performance of main tasks. Building on this foundation, we extend the concept of general DP to introduce CorBin-FL, a novel privacy-aware mechanism that employs correlated binary stochastic quantization to achieve DP while maintaining model accuracy. CorBin-FL leverages secure multi-party computation to enable clients to quantize local model updates while safeguarding individual privacy. This approach optimizes the privacy-utility trade-off, achieving parameter-level local differential privacy (PLDP) with theoretical guarantees for privacy and utility measures. Additionally, we propose AugCorBin-FL, an extension of CorBin-FL that ensures user-level and sample-level central differential privacy. Extensive experiments on the MNIST and CIFAR-10 datasets validate that CorBin-FL and AugCorBin-FL outperform traditional DP mechanisms, such as Gaussian and Laplacian noise addition, in terms of accuracy and privacy preservation, making them robust solutions for enhancing security in FL frameworks.

To address the challenges identified, our project is divided into two key deliverables: Task 1: Preaggregation Similarity Measurement (PA-SM) for attack isolation, and Task 2: Privacy-aware approach leveraging CorBin-FL and AugCorBin-FL.

These methods will be comprehensively analyzed and implemented to enhance the robustness and privacy of FL frameworks.

17. Keywords Federated Learning, Intelligent Transportation Systems (ITS), Adversarial Attacks, Preaggregation Similarity Measurement		**18. Distribution Statement** No restrictions.	
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 22	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

DISCLAIMER 2
CONTACTS 2
ACKNOWLEDGMENT 3
EXECUTIVE SUMMARY 7
CHAPTER 1 8
Introduction 8
1.1 Secure FL Technique 8
1.2 Privacy-aware Approach 9
CHAPTER 2 10
Literature Review 10
CHAPTER 3 12
Preliminaries 12
3.1 Federated Learning 12
3.2 Threat Model 12
CHAPTER 4 14
Methodology 14
4.1 Task 1: Hybrid Method 14
4.2 Task 2: Privacy-aware approach leveraging CorBin-FL and AugCorBin-FL 15
CHAPTER 5 16
Experiments and Result Analysis 16
5.1 Experimental Setup 16
5.2 Backdoor Sample 16
5.3 Result Analysis 17
CHAPTER 6 19
Conclusions 19
REFERENCES 20

List of Figures

Figure 1. Overview of the proposed hybrid privacy-preserving algorithm to protect FL against adversarial attack in ITS ----------14

Figure 2. Backdoor samples ----------17

Figure 3. Performance of our proposed hybrid algorithm----------17

Figure 4. Comparison of different privacy mechanisms----------18

Figure 5. Experimental results for (a) number of shared common random bits, (b) dropout probability and (c) number of clients ----------18

EXECUTIVE SUMMARY

This work aims to enhance the security and privacy of FL in Intelligent Transportation Systems (ITS) by addressing key challenges related to attack isolation and privacy preservation. The first approach introduces a hybrid privacy-preserving algorithm that integrates PA-SM with CDP to safeguard FL against backdoor and inference attacks. PA-SM measures the similarity between local model updates before aggregation to identify and isolate malicious contributions, while CDP ensures that the aggregation process preserves data privacy. These methods are evaluated on CIFAR10 and LISA traffic light datasets, showcasing their ability to mitigate adversarial attacks such as data poisoning-based model replacement with minimal performance degradation. The second approach extends privacy preservation through CorBin-FL and AugCorBin-FL, novel mechanisms leveraging correlated binary stochastic quantization for achieving parameter-level, user-level, and sample-level differential privacy (DP). CorBin-FL employs secure multi-party computation to quantize model updates, optimizing the trade-off between privacy and utility, while AugCorBin-FL expands privacy guarantees to centralized settings. Extensive experiments on MNIST and CIFAR10 datasets reveal that these methods outperform traditional DP techniques like Gaussian and Laplacian noise in maintaining high model accuracy under equal privacy budgets. Together, these approaches address critical gaps in FL frameworks, offering robust solutions for attack isolation and privacy-aware learning in ITS applications.

CHAPTER 1

Introduction

1.1 Secure FL Technique

ITS increasingly leverage advanced artificial intelligence (AI) and machine learning (ML) methods to enhance safety, efficiency, and automation in applications such as object detection. However, adversarial attacks during the training and inference phases pose significant challenges, with potential consequences including public safety risks, communication vulnerabilities, and authentication concerns [Yamany et al., 2021], [Qayyum et al., 2020], [Greenberg et al., 2015]. The distributed nature of ITS applications, such as AVs, highlights the need for robust security and privacy-preserving frameworks [Sun et al., 2021]. FL has emerged as a decentralized approach, enabling edge devices to collaboratively train global models without sharing local data, thereby reducing privacy risks. Despite these advantages, FL remains vulnerable to targeted and untargeted cyber-attacks, including data poisoning, model poisoning, and inference attacks, which compromise model integrity and user privacy [McMahan et al, 2017], [Bagdasaryan et al., 2020], [Jere et al., 2020], [Nasr et al., 2019], [Zhang et al., 2020].

To address these challenges, this study introduces a hybrid privacy-preserving FL algorithm that integrates **PA-SM** and **CDP**. PA-SM measures the similarity between local updates to identify and isolate backdoor updates during training, effectively mitigating data poisoning-based model replacement attacks [Lu et al., 2022]. Meanwhile, CDP ensures privacy by adding noise at the aggregation level, enhancing resilience against inference attacks [Abadi et al., 2016]. Unlike Local Differential Privacy (LDP), which can negatively impact performance, CDP balances privacy and model accuracy. Evaluations on the CIFAR10 and LISA traffic light datasets validate this hybrid approach, demonstrating robust defenses against adversarial attacks with minimal performance degradation [McMahan et al., 2017], [Erlingsson et al., 2019]. This method provides a comprehensive solution for securing FL in ITS applications while addressing critical privacy and security concerns.

The primary contributions of this research are as follows: we have developed and implemented a hybrid privacy-preserving algorithm that integrates an anomaly detection technique, Pre-Aggregation Similarity Measurement (PA-SM) [Lu et al., 2022], and CDP [Abadi et al., 2016] to defend against data poisoning-based model replacement attacks and inference attacks, all while maintaining performance during the training phase. The proposed algorithm offers enhanced security by mitigating the risk of data leakage in FL environments. Additionally, we have conducted an in-depth analysis of the influential factors and behavioral characteristics of backdoor attacks over multiple communication rounds, laying a solid foundation for future research in securing ITS.

In the initial study, we used CDP to protect the global model from inference attacks. However, to enhance resilience against such attacks, we have developed a novel DP method, which will be introduced in the next section.

1.2 Privacy-aware Approach

Distributed stochastic gradient descent and FL are essential components of modern machine learning, facilitating collaborative model training across distributed clients without sharing raw data. FL involves a central server coordinating with multiple clients, each equipped with local datasets, to train a global model over several communication rounds. Clients update the global model locally and send gradients to the server, which aggregates them to refine the model further. This decentralized approach inherently provides a degree of privacy by eliminating the need to share raw data, making it particularly valuable in sensitive applications such as healthcare [Roy et al., 2019], [Xu et al., 2021]. The privacy of FL systems is further enhanced through secure aggregation techniques, which enable clients to transmit updates without revealing individual values [Erlingsson et al.,2019]. However, recent studies have highlighted significant vulnerabilities in FL systems, including the potential for aggregated updates and final model weights to leak sensitive information about the training data. Such weaknesses can lead to privacy breaches, enabling the reconstruction of sensitive data [Papernot et al., 2016], [Bhowmick at al. 201] [Carlini et al., 2019], [Zhu et al., 2019]. To address these risks, DP has been adopted as a rigorous framework to quantify and mitigate data leakage. DP ensures that the outcome of an aggregation mechanism is insensitive to the inclusion or exclusion of any individual's data, making it a cornerstone for secure FL [Dwork et al., 2006].

This study introduces **CorBin-FL**, a privacy-preserving mechanism that employs correlated binary stochastic quantization to achieve multiple notions of DP, including user-level central DP (UCDP), sample-level central DP (SCDP), and parameter-level local DP (PLDP). Using the Diffie-Hellman (DH) key exchange protocol, clients securely exchange private keys to establish shared randomness, which facilitates correlated quantization without exposing local model updates. This approach ensures unbiased quantization while minimizing mean square error (MSE) during aggregation by canceling correlated noise. Unlike prior methods requiring large amounts of shared randomness, CorBin-FL achieves near-optimal MSE with minimal shared bits per parameter [Diffie et al., 2022]. Additionally, the study introduces **AugCorBin-FL**, an extension where a fraction of clients performs correlated quantization, while others quantize individually, enabling scalability and flexibility. These mechanisms advance the state of privacy in FL by balancing robust DP guarantees, communication efficiency, and aggregation accuracy, making them well-suited for practical applications [Witsenhausen et al., 1975].

The key contributions of this work are as follows: we introduce a novel privacy-utility tradeoff framework that balances mean squared error (MSE) as the utility metric and parameter-level local differential privacy (PLDP) as the privacy measure and identify an optimal class of binary stochastic quantizers to achieve this balance. This formulation is further extended to include correlated stochastic quantizer pairs, enabling more efficient and accurate quantization under privacy constraints. We propose the CorBin-FL mechanism, which leverages these quantization strategies to provide strong privacy guarantees while preserving utility. Additionally, we introduce AugCorBin-FL, a hybrid approach that enhances privacy by offering guarantees at both the parameter level and the user level, while maintaining high model performance. Extensive empirical evaluations on the MNIST and CIFAR-10 datasets demonstrate that our approach consistently achieves superior model accuracy compared to existing privacy mechanisms, including Gaussian, Laplacian, and traditional LDP methods, under fixed privacy budgets.

CHAPTER 2

Literature Review

FL has gained significant attention in ITS due to its ability to enable distributed model training across heterogeneous devices while maintaining data privacy. However, the exponential growth of data, diverse device capabilities, and an increasing number of cyber-attacks pose significant challenges in ITS applications. Adversarial attacks such as backdoor attacks in lane detection [Han et al., 2022], model manipulation in traffic control systems [Wang et al.,2021], and Targeted Universal Adversarial Perturbation (TUAP) [Zhang et al., 2021] highlight the vulnerabilities of FL-based ITS systems. Among these, the most prevalent is the model replacement attack, where a malicious client alters updates to compromise the global model [Bagdasaryan et al., 2020]. Various aggregation techniques, such as FedAvg [Moore et al., 2017], FLTrust [Cao et al., 2020], and Trim Mean [Yin et al., 2018], have been introduced to defend against these attacks. However, these methods primarily focus on mitigating backdoor attacks, often neglecting inference attacks.
To strengthen security, PA-SM [Lu et al., 2022] has emerged as a promising approach to isolate backdoor updates by measuring similarity metrics during the pre-aggregation phase. Furthermore, multi-party computation (MPC) has been employed to secure FL by enabling indistinguishable client data, as seen in frameworks like SAFEFL [Gehlhar et al., 2023]. However, MPC-based methods often incur higher computational costs. DP offers an alternative mechanism to secure FL against adversarial attacks by adding random noise to model updates, thereby preserving user privacy. Notable studies include Sun et al. [Sun et al., 2019], who utilized client-level DP with norm clipping to defend against backdoor attacks, and Olowononi et al. [Olowononi et al., 2021], who applied DP to enhance the robustness of vehicular cyber-physical systems (VCPS). Despite advancements in cryptographic and DP-based techniques, existing methods often focus on either inference or poisoning attacks, leaving gaps in comprehensive defense mechanisms for ITS applications.

In this study, we propose a novel hybrid algorithm that integrates PA-SM [Lu et al., 2022] and CDP [Abadi et al., 2016] to simultaneously defend against data poisoning-based model replacement and inference attacks in FL. Unlike traditional methods that rely heavily on cryptography or optimization techniques, our approach ensures robust protection for both internal and external clients in ITS environments. Evaluated on benchmark datasets, the proposed method demonstrates enhanced resilience and efficiency, providing a scalable and secure FL framework tailored to the unique demands of ITS.

Differential Privacy (DP), introduced by Dwork et al. [Dwork et al., 2006, has emerged as a critical framework for providing quantifiable guarantees against data leakage in FL. At its core, DP ensures that the outcome of an aggregation mechanism remains consistent regardless of whether a specific individual's data is included or excluded from the training dataset. In FL, different notions of DP have been explored, including User-level Central DP (UCDP), which protects against revealing client participation during training rounds [Dwork et al., 2014] [Abadi et al., 2016] [McMahan et al., 2017]; Sample-level Central DP (SCDP), which ensures that information about any single training sample is not disclosed [Zhao et al., 2024], [Ghazi et al., 2024]; and Parameter-level Local DP (PLDP), which safeguards individual model parameters during uplink communication [Sun et al., 2020]. Prior techniques to achieve DP have leveraged noise injection

mechanisms such as the Gaussian mechanism, Laplacian mechanism, and Discrete Gaussian mechanism [Abadi et al., 2016], [Bassily et al., 2014], [Canonne et al., 2020]. However, these methods often face challenges with communication overhead, particularly in mobile FL scenarios with limited bandwidth. To address this, gradient quantization methods like LDP-FL [Sun et al., 2020] have been introduced, combining one-bit quantization with PLDP to enhance privacy and communication efficiency. Recent advancements, such as correlated stochastic quantization in mechanisms like CorBin-FL and AugCorBin-FL, have further demonstrated the potential to achieve UCDP, SCDP, and PLDP while minimizing mean square error (MSE) during aggregation [Diffie et al., 2022], [Witsenhausen et al., 1975]. These innovations underline the evolving role of DP in enabling secure, privacy-preserving, and efficient FL systems.

CHAPTER 3

Preliminaries

In this section, we will explore the general concepts of FL and the associated threat model. As there is no specific implemented algorithm in ITS, we focus on object detection as a ML task in general.

3.1 Federated Learning

The FL system comprises two principal components: the central server and edge clients. Suppose we have the m number of clients who are communicating in an FL setting with $(D_{train})_{k=1}^{m}$ datasets. Edge client devices conduct local training on their datasets without sharing their data with central servers or other clients. Local clients communicate with the central server via sharing model weights $\triangle \mathrm{w}_k^p$ only. In every communication round, the central server selects *m* clients from a total of *C* clients and shares the global model with them. The local clients train their model using their private data $(D_{train})_{k=1}^{m}$. Furthermore, the central server collects all the local model updates $\triangle \mathrm{w}_k^p$ and aggregates them to build an updated global model w_{t+1}. The framework of general FL is presented in Figure 1. The aggregation method is an important part of FL to aggregate local client's updates. FedAvg is one of the most used algorithms proposed by McMahan et al. [McMahan et al., 2017]. The aggregation rule can be represented in Equation.

$$w_{t+1} = w_t + \eta \frac{\sum_{k \in (C_k)_{k=1}^{m}} n_k \triangle \mathrm{w}_k^p}{\sum_{k \in (C_k)_{k=1}^{m}} n_k}$$

where η is a learning rate, p is the total number of participants in FL. C_k is selected clients when k = 1· · · m, n_k represents the number of samples, and $\triangle \mathrm{w}_k^p$ is a local update.

3.2 Threat Model

This study considers a threat model based on data poisoning-based model replacement attacks, as outlined by Lu et al. [Lu et al., 2022] and evaluates its implications in ITS. In data poisoning attacks, backdoor attackers create poisoned training samples x_{poi} by embedding backdoor triggers into clean data. The poisoned data is generated using the equation:

$$x_{poi}(x; x_{trigger}) = (1 - \alpha) \otimes \mathrm{x} + \alpha \otimes x_{trigger}$$

Backdoor training can be presented using following equation.

$$w_{backdoor}^{p} = \min_{w} \mathrm{E}_{(x,y) \in (D_{benign} \cup D_{poisoned})} L_f((x, y); w)$$

where α is a trade-off hyper-parameter balancing the original sample x and the backdoor trigger $x_{trigger}$ and $\otimes$ represents element-wise multiplication. A malicious client combines benign and poisoned datasets to train a backdoor model that minimizes the expected loss L_f. The attacker's goal is to replace the global model w_{t+1} with the backdoor model $w_{backdoor}^{p}$, leveraging boosting

factors to amplify the attack's impact [Lu et al., 2022]. Backdoor triggers, such as pixel modifications, have been widely studied, with several types of triggers proposed to enhance the attack's effectiveness [Chen et al., 2017], [Gu et al., 2017], [Liu et al., 2020] [Li et al., 2021].

Inference attacks are another significant vulnerability in FL systems. Honest-but-curious or malicious clients exploit shared global models or updates transmitted through communication channels to infer sensitive information [Nasr et al., 2019], [Zhang et al., 2020]. These attackers may reconstruct private data by analyzing the aggregated updates, exposing critical vulnerabilities in the FL framework. The combined threats of data poisoning and inference attacks emphasize the need for robust, privacy-preserving mechanisms in FL, particularly in ITS applications.

CHAPTER 4

Methodology

Our proposed hybrid privacy-preserving algorithm simultaneously integrates two modules namely the PA-SM [Lu et al., 2022] and CDP [Abadi et al., 2016]. An overview of this algorithm is represented in Figure 3. Our algorithm uses FedAvg [McMahan et al., 2017] to aggregate local model weights. A description of the deployed algorithms, i.e., PA-SM and CDP, is given below

4.1 Task 1: Hybrid Method

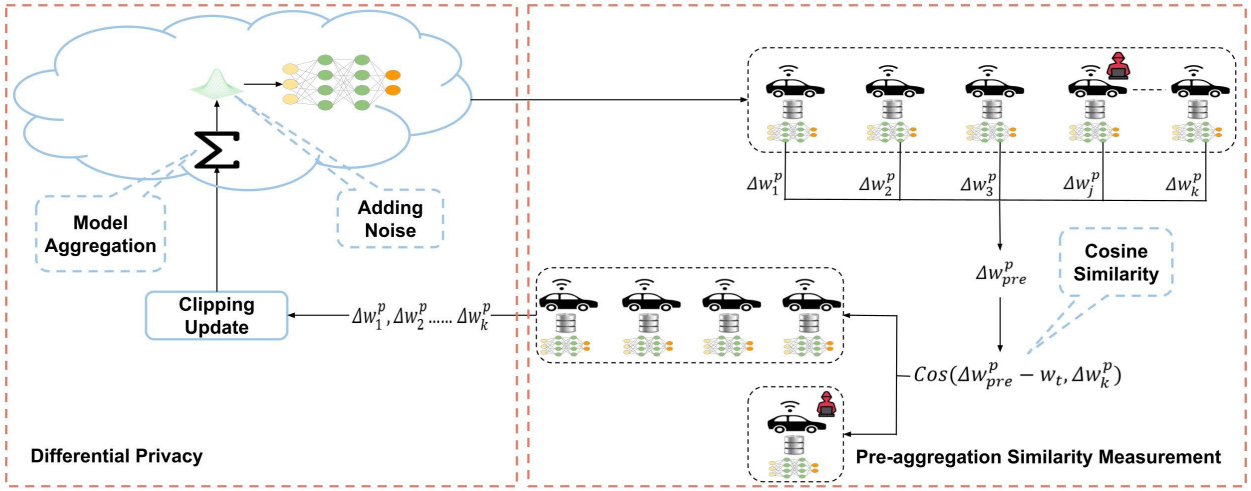


Figure 1. Overview of the proposed hybrid privacy-preserving algorithm to protect FL against adversarial attack in ITS

The proposed methodology introduces a hybrid privacy-preserving FL framework aimed at securing object detection tasks in ITS [mia et al., 2024] in Figure 1. The approach integrates two complementary modules—Pre-Aggregation Similarity Measurement (PA-SM) and CDP to protect against both data poisoning-based model replacement attacks and inference attacks. This hybrid mechanism is implemented within the standard FL process, where AVs act as edge clients and a central server coordinates model aggregation. The figure above illustrates the complete workflow, showing how these two defense strategies are systematically integrated into the FL pipeline.

In the PA-SM module (right side of the Figure 1), each client submits a local model update, which is then compared against a pre-aggregated global model using cosine similarity. Clients whose updates show unusually high similarity—indicative of backdoor behavior—are flagged and excluded from the aggregation process. This detection process ensures that only benign updates are incorporated, preventing attackers from injecting malicious patterns into the global model. The similarity is calculated between the difference of the previous and current model parameters and the expected update direction, which provides a statistically sound basis for identifying abnormal behavior during training.

Once the malicious updates are filtered out, the DP module (left side of the figure) is applied to the aggregation process. The benign model updates are first clipped to limit their sensitivity, then

aggregated, and finally perturbed by adding random noise drawn from a Laplace distribution. This makes it computationally infeasible for an adversary to infer any client-specific information from the global model. By combining anomaly detection with privacy-preserving noise addition, the proposed method offers a robust and scalable solution for securing FL in ITS applications, maintaining high model accuracy while protecting both data integrity and privacy.

4.2 Task 2: Privacy-aware approach leveraging CorBin-FL and AugCorBin-FL

The CorBin-FL [Salehi et al., 2024] mechanism is a privacy-preserving FL algorithm that combines correlated binary stochastic quantization with secure multiparty computation to achieve strong privacy guarantees while minimizing utility loss. The central idea is to allow pairs of clients to perform correlated quantization using a limited number of shared random bits, which significantly reduces the mean squared error (MSE) of aggregated model updates and satisfies parameter-level local differential privacy (PLDP).

At the beginning of each training round, clients establish pairwise secure communication using the Diffie-Hellman key exchange protocol. The server randomly pairs clients and shares with them the global model, the center and radius for clipping, and the privacy budget. Each client computes local model updates and uses the secure channel to generate and exchange random binary sequences with its partner. One client in the pair becomes the “leader” and generates a set of random bit sequences which are securely sent to the “follower.” Both clients then use these shared random bits to perform correlated quantization of their local model updates using the CORBINQ method. These quantized updates are designed such that their aggregation at the server remains unbiased and has significantly reduced variance compared to traditional local DP methods.

The server receives the obfuscated model updates from all clients and aggregates them to form the new global model. The mechanism guarantees PLDP for each model parameter, and the correlated nature of the quantization noise ensures that the aggregate remains accurate even with strong privacy constraints. To enhance privacy further, the authors introduce AugCorBin-FL, a hybrid mechanism where only a fraction of clients participate in correlated quantization while others use independent local DP quantization. This design provides user-level and sample-level central differential privacy (UCDP and SCDP) in addition to PLDP. Theoretical analysis confirms that both mechanisms achieve the desired privacy guarantees while significantly outperforming traditional Gaussian and Laplacian mechanisms in terms of accuracy and robustness to client dropout.

CHAPTER 5

Experiments and Result Analysis

Our proposed hybrid privacy-preserving algorithm simultaneously integrates two modules namely the PA-SM and DP. An overview of this algorithm is represented in Figure 3. Our algorithm uses FedAvg to aggregate local model weights. A description of the deployed algorithms, i.e., PA-SM and DP, is given below

5.1 Experimental Setup

In this study, we evaluate a privacy-preserving FL framework where AVs serve as edge clients and a central server—representing either a roadside unit (RSU) or a traffic management system—coordinates the training. For experimentation, we use the CIFAR-10 benchmark dataset with 10 object classes and the LISA traffic light dataset [Jensen et al., 2016] comprising three traffic light categories (red, yellow, green). Data is distributed among clients in both IID and Non-IID formats. In each communication round, 20 out of 100 AVs are randomly selected to perform local training using a ResNet18 model, with batch size 10 and 50 local epochs. Backdoor attack performance is assessed by injecting 10%, 20%, and 30% poisoned samples into the local datasets and evaluating the results every 5 epochs. The learning rate is fixed at 0.2, and we apply DP using privacy budgets $\varepsilon \in \{1.0, 2.0, 3.0, 4.0, 10.0\}$ and $\delta = 1e{-}5$. In parallel, we implement the CorBin-FL and AugCorBin-FL mechanisms on the CIFAR-10 and MNIST datasets, using ResNet18 and a two-layer VGG model, respectively. These experiments involve varying the number of shared random bits (e.g., $d = 5$), evaluating robustness under client dropout (up to $p = 0.5$), and testing across different client counts (from 50 to 500). A batch size of 64 is used by default, with 32 adopted for larger client settings. All experiments are conducted on an NVIDIA A100-PCIE-40GB GPU.

5.2 Backdoor Sample

An attacker requires a compromised AV to initiate a backdoor attack in FL settings. For experimental purposes, we have initiated backdoor attacks using five types of triggers: point, triangle, star, square, and circle. Among them, three have already been introduced in existing studies. In this study, we introduced two more triggers, circle and square, to assess the robustness of our proposed solution. In this attack scenario, the attacker changes the pixel color of other than targeted class. Furthermore, assigns the target-class label to the poisoned samples and concatenates them with the benign training data. This mixed dataset trains the backdoor model (e.g., compromised AVs local model) and sends back the learned knowledge or parameter to the central server for aggregation. Example backdoor samples for the LISA traffic light dataset [Jensen et al. 2016] are visualized in Figure 2. A dataset comprising 52,000 labeled traffic-sign images were obtained from the Mapillary dataset. These signs were extracted from real-world street-level imagery and span 400 distinct categories, including regulatory, warning, informational, and supplementary types. This dataset serves as a valuable resource for further evaluation of the hybrid algorithm.

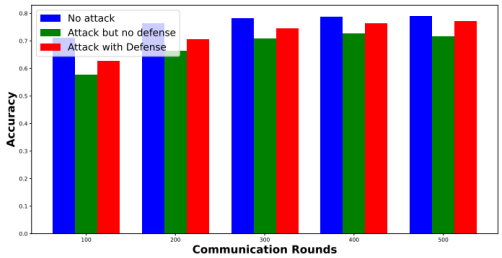


(a) Triangle (b) Circle (c) Point

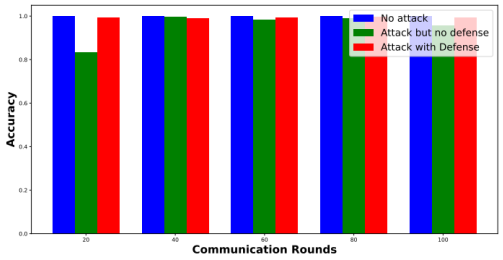
Figure 2. Backdoor samples

5.3 Result Analysis

5.3.1 Hybrid Algorithm



(a) Performance of our proposed hybrid algorithm at the selected round (100^{th}, 200^{th}, 300^{th}, 400^{th}, 500^{th}) of communication [privacy budget ϵ = 10, 30% of backdoor samples, Dataset: CIFAR-10].



(b) Performance of our proposed hybrid algorithm at the selected round (20^{th}, 40^{th}, 60^{th}, 80^{th}, 100^{th}) of communication [privacy budget ϵ = 10, 30% of backdoor samples, Dataset: LISA].

Figure 3. Performance of our proposed hybrid algorithm

Figure 3 analyzes the performance of our hybrid algorithm across three scenarios: no attack, attack without defense, and attack with defense, over several communication rounds. The bar chart (a) in Figure 10 demonstrates the comparative performance of the hybrid algorithm with different scenarios for the CIFAR-10 dataset. Without any attacks, the model consistently improves from 71.06% to 78.97% accuracy, showing stable learning in a secure setting. However, when attacked without defenses, the model's accuracy drops to 57.75% before recovering to 72.81%, highlighting its vulnerability and the difficulty of fully recovering without protective measures. With defenses, the model starts at 62.81% and climbs to 77.13%, demonstrating the effectiveness of defensive strategies in combating attacks and allowing the model to nearly reach the performance seen in the no-attack scenario. The bar chart (b) in Figure 10 demonstrates the comparative performance of the hybrid algorithm under different scenarios for the LISA traffic light dataset. In the no-attack scenario, the model maintains nearly perfect accuracy, consistently around 99.95%, indicating optimal performance in a secure environment. However, when attacked without defense, the model's accuracy initially drops to 83.27%, then fluctuates, peaking at 99.03% and finally settling at 95.58%, highlighting the vulnerability and recovery challenges without defenses. Conversely, the model under attack but with defense shows robust resilience, with accuracy initially at 99.47%, slightly decreasing and then peaking at 99.78%, before ending at 99.42%.

Overall, our proposed hybrid algorithm provides robust security and resilience against adversarial attackers in a FL environment.

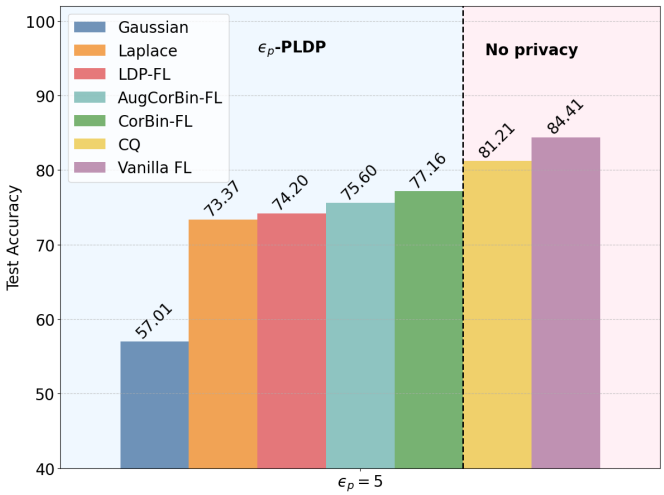
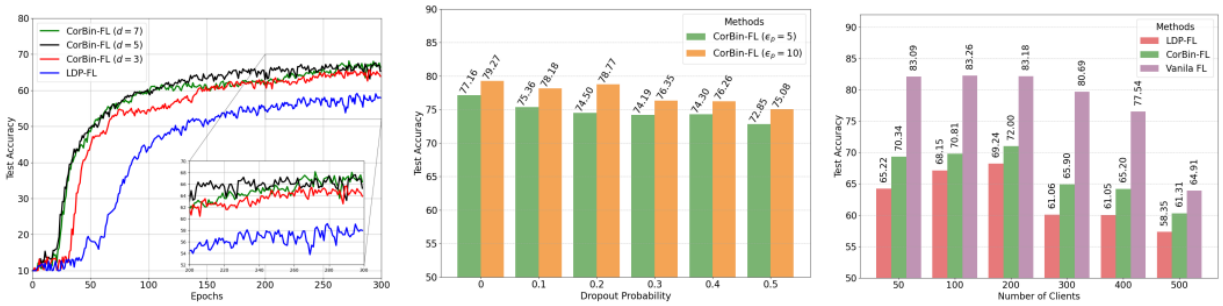


Figure 4. Comparison of different privacy mechanisms

We empirically evaluate the proposed privacy mechanisms through a series of experiments designed to assess model accuracy, robustness, and scalability. The results demonstrate that: (i) under a fixed privacy budget and communication rounds, the proposed mechanisms outperform LDP-FL, Gaussian, and Laplace mechanisms in terms of model accuracy (Figure 4); (ii) accuracy improves with an increasing number of shared random bits per model parameter, saturating at small values such as d=5(Figure 5a); (iii) the mechanisms remain robust even under high client dropout rates, maintaining accuracy for dropout probabilities up to p=0.5 (Figure 5b); and (iv) they consistently outperform LDP-FL across different client counts ranging from 50 to 500 (Figure 5c). For testing, 20% of the training samples are randomly selected for validation. All experiments assume an IID data distribution, with equal random sample allocation to clients. A batch size of 64 is used except in Figure 5c, where it is reduced to 32 due to smaller per-client datasets. Experiments utilize the ResNet18 model for CIFAR-10 and a small two-layer VGG model for MNIST. Unless stated otherwise, shared random bits per parameter are fixed at d=5, and the AugCorBin-FL parameter is set to γ=0.2. The center and radius values are determined by layer-wise min-max statistics and shared by the server each round. All experiments are conducted on a single NVIDIA A100-PCIE-40GB GPU.



(a) (b) (c)

Figure 5: Experimental results for (a) number of shared common random bits, (b) dropout probability and (c) number of clients

CHAPTER 6

Conclusions

This study addresses the pressing challenges of securing FL in ITS by introducing a hybrid privacy-preserving framework that combines general CDP with PA-SM. The proposed method enables robust detection and mitigation of data poisoning-based model replacement and inference attacks by isolating malicious model updates through similarity-based analysis while preserving the privacy of individual data contributions. Building on this foundation, we introduced CorBin-FL, a novel mechanism utilizing correlated binary stochastic quantization and secure multi-party computation to achieve PLDP with strong theoretical guarantees. Furthermore, AugCorBin-FL extends these guarantees to user-level and sample-level central DP, offering comprehensive privacy protection. Extensive experiments on CIFAR10, MNIST, and the LISA traffic light datasets demonstrate the effectiveness of both the hybrid PA-SM + CDP method and the CorBin-FL mechanisms in improving resilience against cyber-attacks while maintaining competitive model performance. These results highlight the potential of integrating tailored privacy and defense strategies in FL to enhance the security and trustworthiness of ITS applications. Future work will focus on expanding the robustness of these mechanisms to handle more dynamic and heterogeneous adversarial conditions.

REFERENCES

Yamany, W., Moustafa, N. and Turnbull, B., 2021. OQFL: An optimized quantum-based federated learning framework for defending against adversarial attacks in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, *24*(1), pp.893-903.

Qayyum, A., Usama, M., Qadir, J. and Al-Fuqaha, A., 2020. Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. *IEEE Communications Surveys & Tutorials*, *22*(2), pp.998-1026.

Greenberg, A., 2015. Hackers remotely kill a jeep on the highway—with me in it. *Wired*, *7*(2), pp.21-22.

Sun, X., Yu, F.R. and Zhang, P., 2021. A survey on cyber-security of connected and autonomous vehicles (CAVs). *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), pp.6240-6259.

McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A., 2017, April. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. and Shmatikov, V., 2020, June. How to backdoor federated learning. In *International conference on artificial intelligence and statistics* (pp. 2938-2948). PMLR.

Jere, M.S., Farnan, T. and Koushanfar, F., 2020. A taxonomy of attacks on federated learning. *IEEE Security & Privacy*, *19*(2), pp.20-28.

Nasr, M., Shokri, R. and Houmansadr, A., 2019, May. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)* (pp. 739-753). IEEE.

Zhang, J., Zhang, J., Chen, J. and Yu, S., 2020, June. Gan enhanced membership inference: A passive local attack in federated learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)* (pp. 1-6). IEEE.

Lu, S., Li, R., Liu, W. and Chen, X., 2022. Defense against backdoor attack in federated learning. *Computers & Security*, *121*, p.102819.

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016, October. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308-318).

Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K. and Thakurta, A., 2019. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 2468-2479). Society for Industrial and Applied Mathematics.

Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N. and Wachinger, C., 2019. Braintorrent: A peer-to-peer environment for decentralized federated learning. *arXiv preprint arXiv:1905.06731*.

Xu, X., Peng, H., Sun, L., Bhuiyan, M.Z.A., Liu, L. and He, L., 2021. Fedmood: Federated learning on mobile health data for mood detection. *arXiv preprint arXiv:2102.09342*.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. and Talwar, K., 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*.

Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G. and Rogers, R., 2018. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J. and Song, D., 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)* (pp. 267-284).

Zhu, L., Liu, Z. and Han, S., 2019. Deep leakage from gradients. *Advances in neural information processing systems*, *32*.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M., 2006. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25* (pp. 486-503). Springer Berlin Heidelberg.

Diffie, W. and Hellman, M.E., 2022. New directions in cryptography. In *Democratizing Cryptography: The Work of Whitfield Diffie and Martin Hellman* (pp. 365-390).

Witsenhausen, H.S., 1975. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, *28*(1), pp.100-113.

Han, X., Xu, G., Zhou, Y., Yang, X., Li, J. and Zhang, T., 2022, October. Physical backdoor attacks to lane detection systems in autonomous driving. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 2957-2968).

Wang, Y., Maniatakos, M. and Jabari, S.E., 2021, December. A trigger exploration method for backdoor attacks on deep learning-based traffic control systems. In *2021 60th IEEE Conference on Decision and Control (CDC)* (pp. 4394-4399). IEEE.

Zhang, Q., Ding, Y., Tian, Y., Guo, J., Yuan, M. and Jiang, Y., 2021, July. Advdoor: adversarial backdoor attack of deep learning system. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis* (pp. 127-138).

Cao, X., Fang, M., Liu, J. and Gong, N.Z., 2020. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*.

Yin, D., Chen, Y., Kannan, R. and Bartlett, P., 2018, July. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning* (pp. 5650-5659). Pmlr.

Gehlhar, T., Marx, F., Schneider, T., Suresh, A., Wehrle, T. and Yalame, H., 2023, May. SafeFL: MPC-friendly framework for private and robust federated learning. In *2023 IEEE Security and Privacy Workshops (SPW)* (pp. 69-76). IEEE.

Olowononi, F.O., Rawat, D.B. and Liu, C., 2021, January. Federated learning with differential privacy for resilient vehicular cyber physical systems. In *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)* (pp. 1-5). IEEE.

Sun, Z., Kairouz, P., Suresh, A.T. and McMahan, H.B., 2019. Can you really backdoor federated learning?. *arXiv preprint arXiv:1911.07963*.

Dwork, C. and Roth, A., 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), pp.211-407.

McMahan, H.B., Ramage, D., Talwar, K. and Zhang, L., 2017. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Zhao, P., Shen, L., Fan, R., Li, Q., Wu, H., Wu, J. and Liu, Z., 2024. Learning with User-Level Local Differential Privacy. *arXiv preprint arXiv:2405.17079*.

Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Meka, R. and Zhang, C., 2024. User-level differential privacy with few examples per user. *Advances in Neural Information Processing Systems*, *36*.

Sun, L., Qian, J. and Chen, X., 2020. LDP-FL: Practical private aggregation in federated learning with local differential privacy. *arXiv preprint arXiv:2007.15789*.

Bassily, R., Smith, A. and Thakurta, A., 2014, October. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science* (pp. 464-473). IEEE.

Canonne, C.L., Kamath, G. and Steinke, T., 2020. The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, *33*, pp.15676-15688.

Chen, X., Liu, C., Li, B., Lu, K. and Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Gu, T., Dolan-Gavitt, B. and Garg, S., 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Liu, Y., Mondal, A., Chakraborty, A., Zuzak, M., Jacobsen, N., Xing, D. and Srivastava, A., 2020, March. A survey on neural trojans. In *2020 21st International Symposium on Quality Electronic Design (ISQED)* (pp. 33-39). IEEE.

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16463-16472).

Jensen, M.B., Philipsen, M.P., Møgelmose, A., Moeslund, T.B. and Trivedi, M.M., 2016. Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE transactions on intelligent transportation systems*, *17*(7), pp.1800-1815.

Salehi, H.A., Mia, M.J., Pradhan, S.S., Amini, M.H. and Shirani, F., 2024. CorBin-FL: A Differentially Private Federated Learning Mechanism using Common Randomness. *arXiv preprint arXiv:2409.13133*.

Mia, M.J. and Amini, M.H., 2024. A secure object detection technique for intelligent transportation systems. *IEEE Open Journal of Intelligent Transportation Systems*.