Federated Learning-based Intrusion Detection System for Internet of Autonomous Vehicles against Poisoning Attacks*

Ulysses Lam[1], Jin-Hee Cho[2], Hyuk Lim[3], Terrence Moore[4], Frederica Free-Nelson[4], Hyunjae Kang[1], and Dan Dongseong Kim[1] †

[1] The University of Queensland, Brisbane, Queensland, Australia
{u.lam, hyunjae.kang, dan.kim}@uq.edu.au
[2] Virginia Tech, Falls Church, VA, US
jicho@vt.edu
[3] Korea Institute of Energy Technology, Naju-si, South Korea
hlim@kentech.ac.kr
[4] US DEVCOM Army Research Laboratory, Adelphi, MD, US
{terrence.j.moore.civ, frederica.f.nelson.civ}@army.mil

Abstract

Autonomous vehicles are becoming a new trend in transportation these days. They enable self-driving with the assistance of multiple types of electronic sensors to make every driving decision. Not surprisingly, these vehicles are potentially more vulnerable to cyber-attacks compared to traditional human-driven ones. Cybersecurity for autonomous vehicles will be crucial in the near future. However, intrusion detection systems (IDSes) for vehicles are still in the early stages. Many IDS models that claim to work for vehicles are actually built with traditional Internet datasets rather than those with real vehicle data, which is impractical in reality. In this paper, we develop IDS models with Federated Learning (FL) with datasets obtained from real vehicles, achieving high performance in attack detection. Furthermore, our IDS models are robust against poisoning attacks to local clients, which is tested in different scenarios.

Keywords: Federated Learning · Intrusion Detection · CAN bus · VANET · Poisoning Attack

1 Introduction

In recent years, the number of autonomous vehicles has grown rapidly around the world. In countries such as the United States and China, self-driving taxis have become a revolutionary means of transportation for passengers in major cities. Although autonomous vehicles have brought great convenience to transportation, statistical reports show that the accident rates of autonomous vehicles are generally higher than those of human-driven cars. In addition, since autonomous vehicles make driving decisions based on information obtained from various electronic devices, such as the Global Positioning System (GPS), Light Detection and Ranging (LiDAR), cameras, and radars, they can be more vulnerable to cyber-attacks compared to traditional human-driven vehicles. Therefore, as autonomous vehicles become the future trend of transportation, their safety issues will become the main concern of all road users at the same time.

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 5, December 16-18, 2025, Sapporo , Japan.

†Corresponding author

To protect autonomous vehicles from cyber-attacks, a potential solution is to develop new Intrusion Detection Systems (IDSes) for vehicles. Firstly, IDS models should be built with datasets obtained from real vehicle data so that they are more practical to detect attacks in vehicular networks. Moreover, since autonomous vehicles have very limited computational resources, which differ from those of computers, Federated Learning (FL) can be an ideal solution, as local clients share the training process with smaller amounts of data. FL has been studied in various domains, but its effectiveness and robustness for vehicular intrusion detection remain insufficiently explored.

In this paper, we analyse the feasibility of deploying FL-based IDSes on Vehicular Ad-hoc Networks (VANET), process data from two datasets obtained from real vehicles, and develop reliable FL-based IDS models for vehicles with high detection rates. In addition, the robustness of FL models is also an important measure, as local clients can be compromised by poisoning attacks, which can affect IDS performance. Therefore, we also develop FL models in different scenarios involving compromised clients and contaminated local data, demonstrating the performance and robustness of the models.

2 Background

2.1 Vehicular Ad-hoc Network

Vehicular Ad-hoc Networks (VANET) are decentralised communication networks that connect vehicles with each other and with roadside infrastructure. By enabling continuous data exchange across a large number of vehicles, VANET provides a rich source of traffic and security-related information that can be leveraged for collaborative learning [9]. This distributed nature makes VANET particularly suitable for FL, where vehicles act as clients that train local intrusion detection models on their own data and share only model updates with a central server. Such an approach preserves data privacy, reduces communication overhead, and allows intrusion detection systems to benefit from diverse and dynamic vehicular data without requiring direct data sharing.

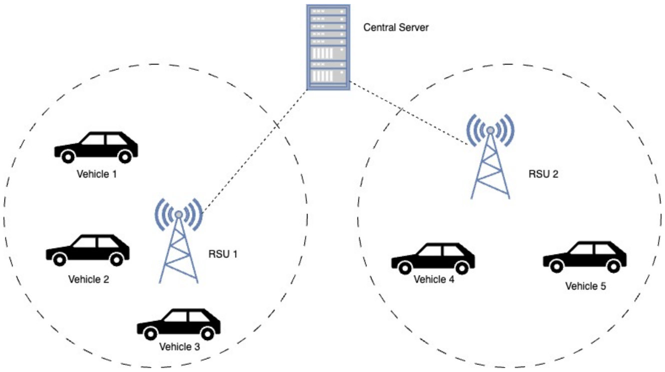

Figure 1: Vehicle-to-Infrastructure in VANET

In this context, VANET can be realised through two communication scenarios: Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V). In this paper, we focus only on the V2I scenario as shown in Figure 1. The V2I scenario consists of a central server, multiple Roadside Units (RSUs) and multiple vehicles. It enables communication between vehicles and RSUs, and subsequently between RSUs and the central server. Therefore, vehicles can share data with the central server and vice versa. V2I is often used for managing live traffic and sending emergency messages.

2.2 Federated Learning

FL is a decentralised machine learning technique that enables multiple clients to collaboratively train a shared global model without exchanging raw data. Each client trains the model locally on its own dataset and transmits only the model updates to a central server, where aggregation is performed to refine the global model. This approach enhances data privacy, reduces communication costs, and allows learning from diverse, decentralised data sources [12].

Given these properties, integrating FL with VANET offers several advantages. First, since clients upload only model updates rather than raw data collected from vehicle sensors, FL enables vehicles to contribute to training without exposing sensitive information such as location (GPS data) or driving behaviour. Second, by avoiding direct data transmission, FL also reduces communication overhead between vehicles and the central server. Third, FL supports vehicular edge computing by leveraging the computational resources of edge nodes, thereby alleviating the training burden on the central server that would otherwise need to process data from numerous vehicles [18].

However, security issues have emerged due to FL's distributed and decentralised nature. One major threat is poisoning attacks [16, 5], in which malicious participants corrupt the global model either by manipulating local training data (data poisoning) or by directly altering model updates (model poisoning). Data poisoning, such as label flipping, can bias the model toward targeted misclassifications, while model poisoning injects adversarial gradients to degrade overall performance. These threats highlight the need for further research on the impact of poisoning attacks, particularly in safety-critical domains such as vehicular intrusion detection.

3 Related Work

Recent studies have shown growing interest in applying FL to intrusion detection in vehicular networks. Chen et al. [4] proposed VAN-FED-IDS for VANETs, where RSUs serve as FL clients to aggregate models trained with packet- and physics-based features, demonstrating that FL can preserve privacy while maintaining detection accuracy. Gurjar et al. [7] developed an FL-based misbehaviour classification system for VANETs using RNN and LSTM variants under different aggregation strategies, highlighting FL's scalability in dynamic traffic environments. Huang et al. [8] introduced FED-IoV for the Internet of Vehicles (IoV), applying MobileNet-Tiny within an FL framework and validating its effectiveness on both in-vehicle network and computer network intrusion datasets. Similarly, Althunayyan et al. [1] applied hierarchical FL to IoV using an in-vehicle network intrusion dataset, while Bhavsar et al. [3] proposed FL-IDS, which leverages vehicular edge devices in transportation IoT to reduce communication costs and distribute computation. Collectively, these works establish FL as a promising paradigm for intrusion detection in vehicular networks.

Beyond accuracy and efficiency, a few studies have addressed the security of FL in vehicular contexts. Mansouri et al. [11] integrated FL with blockchain to enhance trust and aggregation integrity in VANETs, while Ullah et al. [17] proposed SPBFL-IoV, a blockchain-based framework incorporating homomorphic encryption and filtering to mitigate poisoning attacks in general IoV environments. Although these studies advanced privacy and robustness in FL,

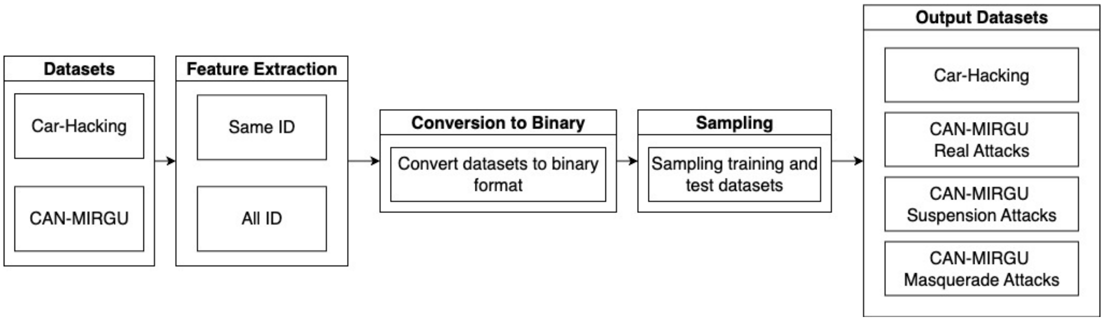

Figure 2: Data Processing of Datasets

none specifically investigated poisoning attacks in the context of intrusion detection for vehicular networks using FL. To the best of our knowledge, this paper is the first to address this gap by systematically evaluating poisoning attacks against FL-based intrusion detection in vehicular environments.

4 Data Processing

We have selected the Car-Hacking [14] and CAN-MIRGU [13] datasets to build FL-based IDS models and compare their performance. The procedures of data processing are shown in Figure 2. They are multi-class, with each attack type saved in an individual file. To simplify the detection, we convert them into binary classes: attack and benign. Both datasets consist of Controller Area Network (CAN) bus messages, the de facto communication protocol for in-vehicle networks. Each message includes basic information such as timestamp, CAN ID, and data.

The Car-Hacking dataset [14] is published with real vehicle data by the Hacking and Countermeasure Research Lab (HCRL). It contains four attack types: DoS Attack, Fuzzy Attack, Spoofing the drive gear, and Spoofing the RPM gauge, as well as attack-free data. The dataset is unbalanced with labels for attacks and benign data.

The CAN-MIRGU dataset [13] is a comprehensive dataset published in 2024 that comprises 36 attack types divided into three main groups: real attacks, suspension attacks, and masquerade attacks. The attacks were injected while the vehicle was driven on roads rather than on a dynamometer, which is the same as practical cases. There is no IDS built with the CAN-MIRGU dataset that has been published yet, as it is relatively new.

4.1 Feature Extraction

Since the datasets contain only basic information about CAN bus messages, e.g. timestamp, CAN ID and data, we extract message counts in three time windows: 1 second, 0.5 second and 0.25 second, based on the same CAN IDs and all CAN IDs. Therefore, six features are extracted from the data. Lastly, the extracted features are normalised with Min-Max Normalisation.

Table 1: Summary of Sampled Datasets.

Dataset	Sampled Dataset	Total Records	Attack Records	Benign Records
Car-Hacking	Training	400,000	98,997	301,003
	Test	40,000	9,816	30,184
CAN-MIRGU	Training	400,000	131,938	268,062
(Real Attacks)	Test	40,000	14,674	25,326
CAN-MIRGU	Training	400,000	131,287	268,713
(Suspension Attacks)	Test	40,000	13,173	26,827
CAN-MIRGU	Training	8,8500	22,730	65,770
(Masquerade Attacks)	Test	9,761	2,453	7,308

4.2 Conversion to Binary

The six normalised features, as well as CAN ID and data, are converted to binary format. Since the features are normalised, their values are multiplied by 1000, then the integer parts are converted to binary. These eight columns of data will be the input data for training FL models.

4.3 Sampling

Since the datasets have tens of millions of records, it is infeasible to include all the data to build FL models. Therefore, sampling a small amount of data is necessary before developing the IDS models. Both datasets contain two types of sub-datasets: attack datasets and benign datasets. The attack datasets contain both benign data and one type of attack data, while the benign datasets contain only benign data. In each attack dataset, the number of benign data is already overwhelming compared to the number of attack data. Therefore, we do not include benign datasets in data processing to make the sampled datasets more balanced. Moreover, to train FL models, each client should have their own local training dataset, so the training dataset is divided into 100 sub-datasets, as we will have 100 clients in the FL model. Lastly, each sampled dataset contains a test dataset and 100 local training datasets.

Car-Hacking Dataset

The numbers of the four attack types are generally balanced, but benign data are six times more than attack data. Before sampling, 50% of the benign data are removed randomly. The numbers of records are shown in Table 1.

CAN-MIRGU Dataset

The CAN-MIRGU dataset comprises 36 attack types, with the number of records for each type varying from around one hundred to hundreds of thousands. To make the data more balanced and less complicated, we divided the CAN-MIRGU dataset into three sub-datasets based on the three main groups: real attacks, suspension attacks, and masquerade attacks.

Real Attacks

Real attacks in CAN-MIRGU contain 26 attack types with the number of records ranging from 118 to 40,759, which is extremely unbalanced. Therefore, the numbers of attack records are sampled using Algorithm 1. Attack types with more records are sampled with a smaller proportion so that their numbers are more balanced. In addition, the total number of benign records is nearly 65 times that of attacks, so only one-thirtieth of the benign data is randomly chosen before sampling training and test datasets. The numbers of records are shown in Table 1.

Algorithm 1 Sampling Attack Records for Real Attacks

```
Number of Attacks: $N$;
if $N \geq 30,000$ then
    $N = N \cdot 0.25$
else
    if $N \geq 10,000$ then
        $N = N \cdot 0.7$
    end if
end if
```

Suspension Attacks

Suspension attacks in CAN-MIRGU comprise 5 attack types, with approximately 20,000 records for each type, and the ratio between benign and attack records is approximately 2:1. The training and test datasets are sampled directly, as they are generally balanced, as shown in Table 1.

Masquerade Attacks

Masquerade attacks in CAN-MIRGU contain 5 attack types with numbers of records ranging from 118 to 10,936, which is also unbalanced. The ratio of benign to attack records is approximately 87:1. Since the total number of attack records is significantly smaller than that of other datasets, all attacks are included in the sampled dataset, while only one-thirtieth of benign data is selected, as shown in Table 1.

5 Federated Learning Model

The FL model usually consists of a central server and multiple fixed clients. However, in VANET, clients are equivalent to vehicles, and they are moving all the time. Therefore, the clients connected to the central server are supposed to change from time to time. To simulate practical cases, our FL models consist of 100 clients with individual local training datasets, and 10 clients are randomly selected in each round of training.

The phases of one round of training are as follows:

1. **Server-to-Client Broadcast:** The central server updates selected clients with the weight of the global model.

2. **Local Clients Update:** Each selected client trains their local model with local training datasets.

3. **Client-to-Server Upload:** Selected clients upload weights of local models to the central server.

4. **Server Update:** The central server updates the global model by aggregating local model weights with an aggregation algorithm.

In our models, the training process of FL models involves five rounds of these phases. In the **Local Clients Update** phase, local models are built as Convolutional Neural Network (CNN) [10] for binary classification. In the **Server Update** phase, Federated Averaging (FedAvg) [15] is applied. FedAvg is a simple aggregation algorithm to obtain a new global model by calculating the average values of local model weights.

5.1 Simulation Cases with Contaminated Datasets and Compromised Clients

In addition to normal datasets, different FL models are also trained by changing the number of compromised clients and the percentage of data contamination, so that the robustness of FL models is tested. To contaminate local training datasets, a certain percentage of labels are flipped each time, so attacks are labelled as benign and vice versa. The numbers of compromised clients are set to be 20, 40, 60, 80 and 100, while the percentages of flipped labels in their local training datasets are set to be 20%, 40%, 60%, 80% and 100%. Hence, there are a total of 26 cases: 25 cases with abnormal data and 1 case with normal data.

6 Results

Multiple FL models are built in the 26 cases mentioned, and for each case, the model is trained and tested 100 times. In this section, we show and compare the performance and robustness of our FL-based IDS models with different datasets.

6.1 Car-Hacking Dataset

For the Car-Hacking dataset, our FL model achieves an accuracy of over 96% when no client is compromised, as shown in Figure 3. When 20 out of 100 clients are compromised, the accuracy is hardly affected. Similarly, when only 20% of the labels are flipped, the accuracy is maintained above 96%. The model has high robustness, so it achieves accuracy above 80% when at least 60% of the clients or data are normal. The accuracy starts to drop rapidly as the majority of clients are compromised, with a higher percentage of data labels being flipped.

Compare the following two cases:

1. 60 clients are compromised, and 80% of labels are flipped

2. 80 clients are compromised, and 60% of labels are flipped

Both cases indicate that 48% of the overall data is contaminated among all clients. However, Case 2 has an obviously higher accuracy than that of Case 1, which is 73.4% and 58.2%, respectively. Our models for the CAN-MIRGU dataset also show the same results as in Figure 5. Therefore, increasing the number of compromised clients may have a bigger impact on the robustness of models than increasing the percentage of flipped labels, given that the overall percentage of contaminated data is equal.

The graphs have shown an abnormal trend, in which the accuracy increases slightly as the number of clients increases. However, the increment in accuracy is less than 0.7%, which can be neglected. It can be a minor fluctuation because the clients are reselected in each round of training.

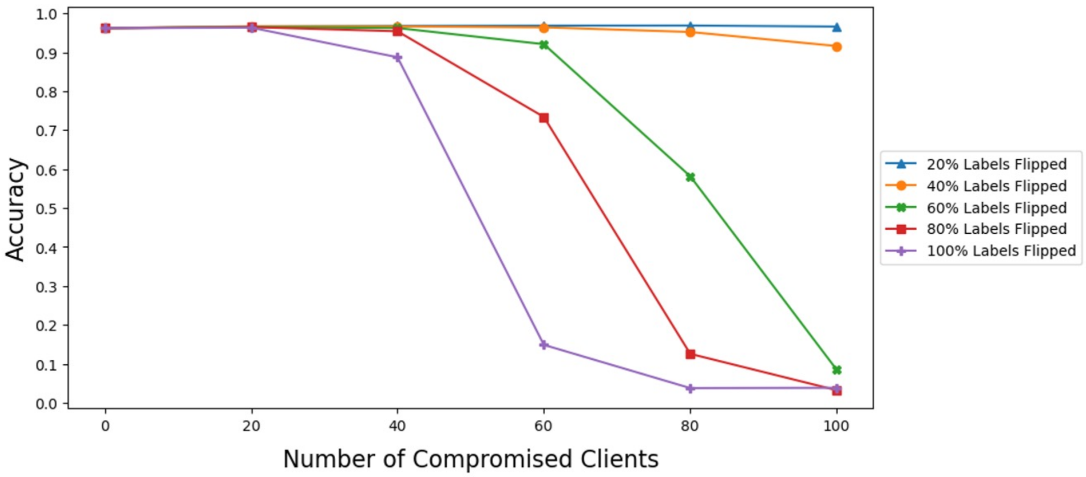

Figure 3: Accuracy vs Compromised Clients and Flipped Labels for Car-Hacking

Since the datasets are unbalanced, we also look at the F1-score of the model, as it is more reliable in balancing the performance of precision and recall [2]. Since the FL models are trained and tested 100 times for each case, with different selected clients in each training round, the average F1-score should be calculated. A more accurate way to obtain the average F1-score is to calculate it with the total numbers of true positives (TP), false positives (FP) and false negatives (FN). However, since the values of TP, FP, and FN are not available, our average F1-score is calculated with the average values of precision and recall:

$$F1 = \frac{2 \cdot (precision \cdot recall)}{(precision + recall)}$$

Since the datasets have positive classes of more than 10%, which is not highly unbalanced, the bias of the calculated F1-score is minor [6].

The FL model also has a high F1-score greater than 0.91, as shown in Figure 4. The F1-score is similar to the accuracy, and it remains high when the majority of clients or data are not compromised or contaminated.

6.2 CAN-MIRGU Dataset

For the CAN-MIRGU dataset, our FL models have also demonstrated high robustness to compromised clients and contaminated data, as shown in Figures 5 and 6. Both the accuracy and the F1-score decrease slightly when less than half of the clients and local data are affected by flipped labels. The FL models for both real attacks and masquerade attacks achieve very high performance, comparable to that of the Car-Hacking dataset. However, the accuracy and F1-score are significantly lower for suspension attacks.

The FL model performs poorly for suspension attacks due to its low recall, as shown in Figure 7. The recall values are always around 0.5, which indicates that around 50% of attacks are correctly classified. In addition, contaminated data can barely affect recall values, even though all 100 clients are compromised with 100% of labels flipped. Since suspension attacks are launched by compromising ECUs for a period and preventing their messages from being

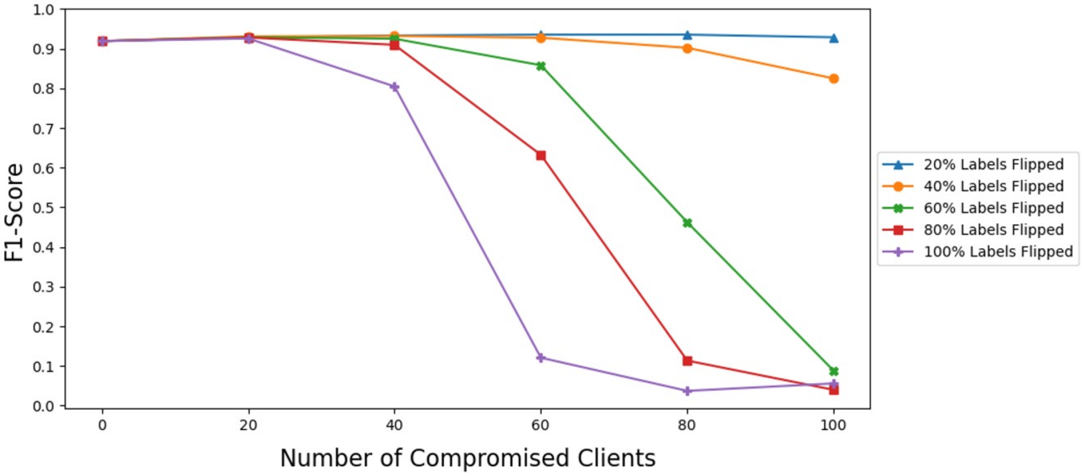

Figure 4: F1-Score vs Compromised Clients and Flipped Labels for Car-Hacking

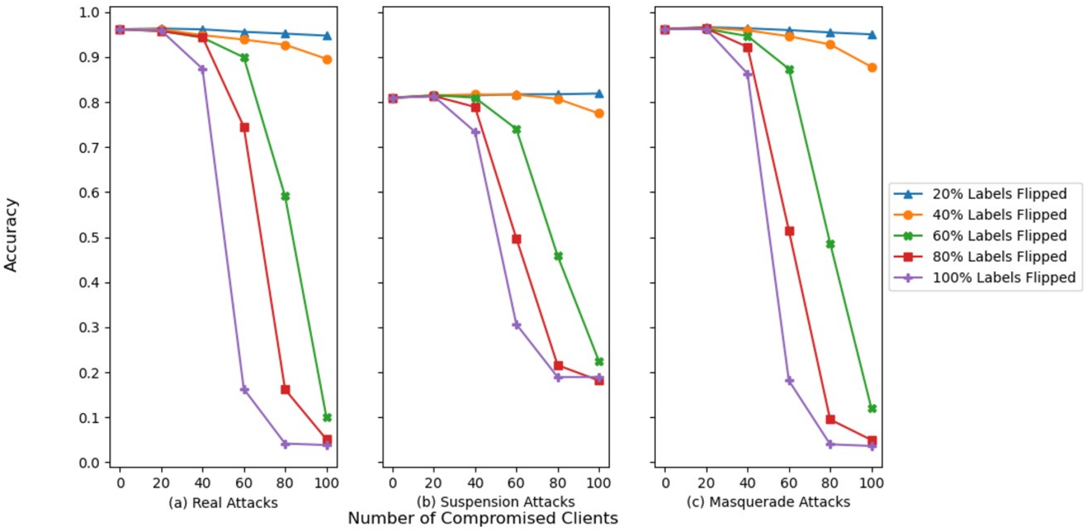

Figure 5: Accuracy vs Compromised Clients and Flipped Labels for CAN-MIRGU

sent, the attack messages may contain the same CAN IDs and data as the benign ones, only with different timestamps and labels. Therefore, flipping the labels does not affect the local training datasets. Moreover, since suspension attacks do not inject additional attack frames, unlike real attacks, which inject attacks in a fixed time interval, the extracted features of message counts cannot provide useful information for attack detection. Hence, the FL model has low detection rates for suspension attacks.

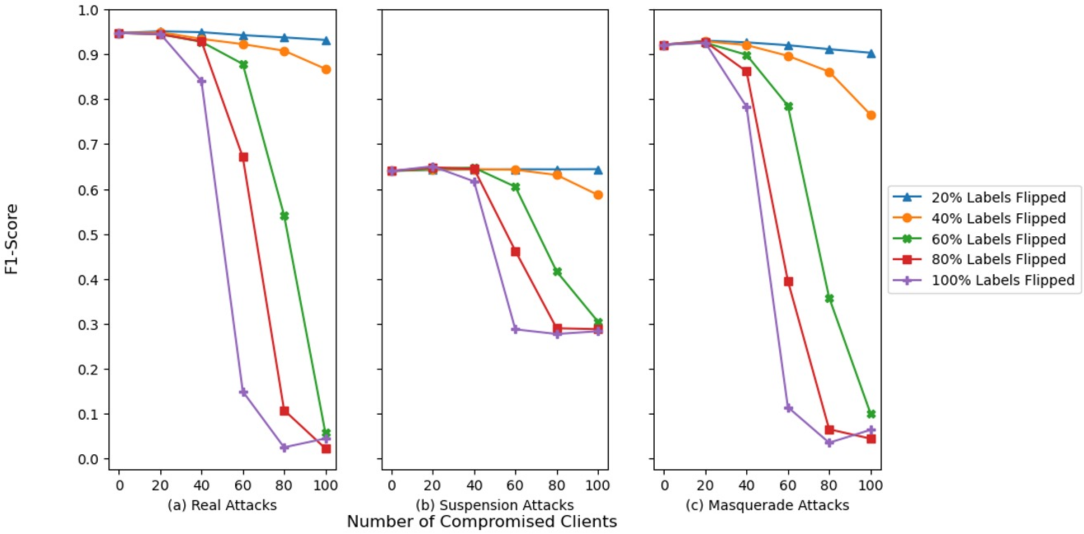

Figure 6: F1-Score vs Compromised Clients and Flipped Labels for CAN-MIRGU

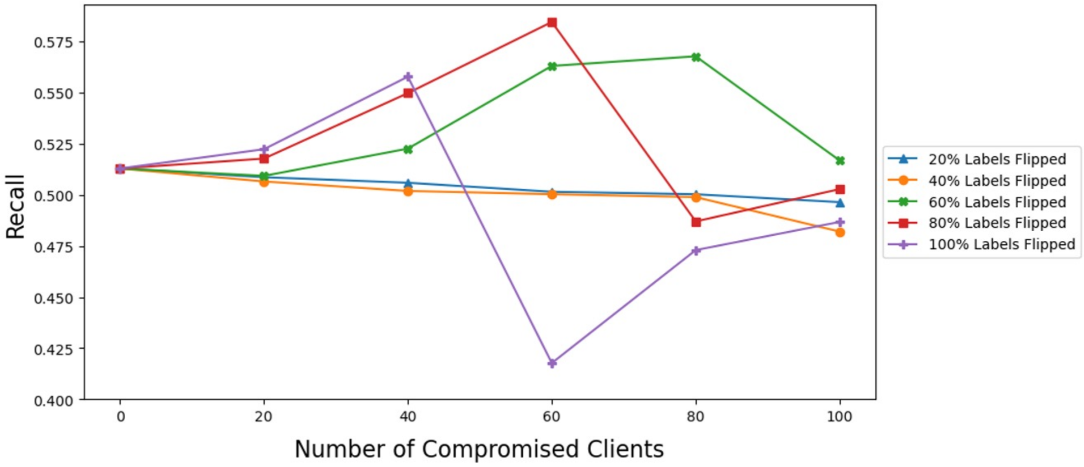

Figure 7: Recall vs Compromised Clients and Flipped Labels for Suspension Attacks

7 Conclusion

In this paper, we develop new FL-based IDS models for autonomous vehicles with real datasets containing CAN bus messages. The IDS models show the advantages of decentralisation, that local clients with limited computational resources can still detect intrusions with high accuracy by sharing the training process with a small amount of local data. Our models exhibit high performance in detecting most attack types, except suspension attacks, which have distinct characteristics. Our FL models have also shown strong robustness against data contamination.

The models can maintain relatively high performance as long as more than half of the clients or local datasets are not under poisoning attacks. For future work, we will improve the FL models to detect more attack types, including suspension attacks. One potential solution is to extract more meaningful features from the limited information contained in CAN bus messages. Moreover, we will further enhance the robustness against poisoning attacks. This can be achieved by selecting an aggregation algorithm that differs from FedAvg.

8 Acknowledgments

This material is partly supported by the International Technology Center Indo-Pacific (ITC IPAC) under Contract No. FA520924C0027 and Army Research Office grant No. W911NF-24-2-0241.

References

[1] Muzun Althunayyan, Amir Javed, and Omer Rana. A robust multi-stage intrusion detection system for in-vehicle network security using hierarchical federated learning. *Vehicular Communications*, 49:100837, October 2024.

[2] Saptarshi Bej, Narek Davtyan, Markus Wolfien, Mariam Nassar, and Olaf Wolkenhauer. Loras: An oversampling approach for imbalanced datasets. *Machine Learning*, 110(2):279–301, 2021.

[3] Mansi H. Bhavsar, Yohannes B. Bekele, Kaushik Roy, John C. Kelly, and Daniel Limbrick. FL-IDS: Federated Learning-Based Intrusion Detection System Using Edge Devices for Transportation IoT. *IEEE Access*, 12:52215–52226, 2024.

[4] Xiuzhen Chen, Weicheng Qiu, Lixing Chen, Yinghua Ma, and Jin Ma. Fast and practical intrusion detection system based on federated learning for VANET. *Computers & Security*, 142:103881, July 2024.

[5] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to Byzantine-Robust federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1605–1622. USENIX Association, August 2020.

[6] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57, 2010.

[7] Dayanand Gurjar, Jyoti Grover, Vanisha Kheterpal, and Athanasios Vasilakos. Federated learning-based misbehavior classification system for VANET intrusion detection. *Journal of Intelligent Information Systems*, 63(3):807–830, June 2025.

[8] Kun Huang, Rundong Xian, Ming Xian, Huimei Wang, and Lin Ni. A comprehensive intrusion detection method for the internet of vehicles based on federated learning architecture. *Computers & Security*, 147:104067, December 2024.

[9] Muhammet Ali Karabulut, A. F. M. Shahen Shah, Haci Ilhan, Al-Sakib Khan Pathan, and Mohammed Atiquzzaman. Inspecting vanet with various critical aspects – a systematic review. *Ad Hoc Networks*, 150:103281, 2023.

[10] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2022.

[11] Fedwa Mansouri, Mounira Tarhouni, Bechir Alaya, and Salah Zidi. A distributed intrusion detection framework for vehicular Ad Hoc networks via federated learning and Blockchain. *Ad Hoc Networks*, 167:103677, February 2025.

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh

and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[13] Sampath Rajapaksha, Garikayi Madzudzo, Harsha Kumara Kalutarage, Andrei Petrovski, M.OmarAl-Kadri, Robert Gordon University, Horiba, and Mira Ltd. Can-mirgu: A comprehensive can bus attack dataset from moving vehicles for intrusion detection system evaluation. 2024.

[14] Hyun Min Song, Jiyoung Woo, and Huy Kang Kim. In-vehicle network intrusion detection using deep convolutional neural network. *Vehicular Communications*, 21:100198, 2020.

[15] Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4289–4301, 2023.

[16] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In Liqun Chen, Ninghui Li, Kaitai Liang, and Steve Schneider, editors, *Computer Security – ESORICS 2020*, pages 480–501, Cham, 2020. Springer International Publishing.

[17] Irshad Ullah, Xiaoheng Deng, Xinjun Pei, Husnain Mushtaq, Muhammad Uzair, and Shazib Qayyum. A Blockchain-Based Federated Learning Framework Against Poisoning Attacks in the Internet of Vehicles. *Computer Networks*, page 111705, September 2025.

[18] Xinran Zhang, Jingyuan Liu, Tao Hu, Zheng Chang, Yanru Zhang, and Geyong Min. Federated learning-assisted vehicular edge computing: Architecture and research directions. *IEEE Vehicular Technology Magazine*, 18(4):75–84, 2023.