A Privacy-Preserving Federated Intrusion Detection Framework for Internet of Vehicles Using CNN-BiGRU with Attention and Adaptive Weighted Aggregation

Sarah H. Mnkash[1*], Faiz A. Alawy[2], Israa T. Ali[3]

[1] Computer Science College, University of Technology, Baghdad, Iraq, cs.22.13@grad.uotechnology.edu.iq
[2] College of Engineering, Kent State University, Ohio, USA, falalaw@kent.edu
[3] College of Engineering Technology for Computers and Artificial Intelligence, Northern Technical University, Kirkuk, Iraq, israa.ali24@ntu.edu.iq
*Corresponding Author: Sarah H. Mnkash, Computer Science College, University of Technology, Baghdad, Iraq, cs.22.13@grad.uotechnology.edu.iq

Abstract
The growing internet of vehicle ecosystems has created significant cyber security vulnerabilities, particularly with respect to privacy-preserving detection of network intrusions in distributed, resource-limited environments. Traditional centralized intrusion detection systems (IDSs) produce major problems for vehicular CAN-bus networks because traditional solutions are unable to scale; they introduce single points of failure into deployments; and they violate critical privacy requirements through constant traffic transmission of sensitive data to remote servers. Federated Learning (FL) could provide a new way to build systems, yet it is still broken by using standard FL protocols, because they do not accommodate for Non-IID data distributions (non-independent identically distributed) data, as well as clients who behave poorly during the FL process. In this paper, we propose a full-fledged federated Intrusion Detection System (IDS) framework and done so using a benchmark dataset called "CIC-IoV 2024." Additionally, we provide the following three innovations/contributions to FL research: 1) Develop a unique CNN-BiGRU-Attention deep learning architecture that captures spatial byte-level correlation and temporal bidirectional dependencies within sequential CAN (Controller Area Network) traffic; 2) Create an adaptive weighted input aggregation (AWI) mechanism to continuously provide trust-based aggregation weights for each client's update (of the client's FL model), determined by their cosine similarity and norm deviation; 3) Build a two-stage adversarial receiver defense pipeline that utilizes dual thresholding (norm and cosine) algorithms to determine whether or not an adversarial update from a client should be included in the final aggregator output. Overall experimental evaluation of the proposed method under realistic non-IID conditions, and in the presence of adversarial clients, achieved 98% accuracy, precision, recall and F1 scores, and ROC-AUC = 0.998. In addition, statistical analysis through 5-fold cross-validation with a mean ± standard deviation (97.14% ± 0.29%) and a two-tailed paired t-test ($p < 0.05$) confirmed the proposed method's robustness and was significantly more effective than using CNN, LSTM, RF, XGBoost, SVM, and FL-CNN baselines.

Keywords: Federated Learning (FL), Intrusion Detection System (IDS), Internet of Vehicles (IoV),CAN-bus Security, CNN-BiGRU, Attention Mechanism, Adaptive Weighted Aggregation.

I. Introduction

The convergence of the Internet of Things (IoT), edge computing, and vehicular communication systems has led to today's hyper-connected vehicular networks. Modern vehicles contain 70+ Electronic Control Units (ECUs) that are all connected to each other using the Controller Area Network (CAN) protocol [1]. The CAN-bus architecture controls safety-critical vehicle functions (e.g. Braking, steering, engine management, adaptive cruise control), meaning that successful cyberattacks on CAN-bus communications place vehicle occupants and road users in immediate danger. Dedicated FL-based IDS approaches for CAN-bus vehicular networks [2] and deep learning signal-level detection frameworks [3] confirm the urgency and feasibility of protecting these critical systems. Researchers have demonstrated that remote vehicle intrusions are possible, including: DoS flooding, message spoofing, replay attacks, and message injection, thus cybersecurity has changed from a peripheral priority to a foundational need for next-generation vehicular systems [4]. Deep learning and swarm intelligence techniques have also been explored to enhance the performance and security of vehicular ad-hoc networks [5].

Intrusion Detection Systems (IDS) are considered to be the main type of protection for businesses against such attacks [6]. However, traditional signature-based intrusion detection systems cannot detect zero-day attacks and anomaly-based approaches suffer from very high numbers of false positives, both traditional methods have a common fault, they assume that there is a properly functioning centralised data model, this is incompatible with the distributed nature of vehicular networks and raises issues with respect to data sovereignty [7]. FL solves this tension as it allows for the distributed clients (i.e., vehicles) of a distributed system to work together to create a single shared global model without

exposing their local raw traffic data to one another [8]. Surveys on FL advancements confirm its growing deployment readiness for IoT and vehicular edge environments [9]. The proposed framework accommodates non-IID heterogeneity and adversarial manipulation through the use of three tightly coupled innovations: 1) A CNN-BiGRU-Attention architecture, 2) the AWI aggregation mechanism, and 3) a two-stage defense pipeline. Figure 1 provides an overview of the complete architecture of the system.

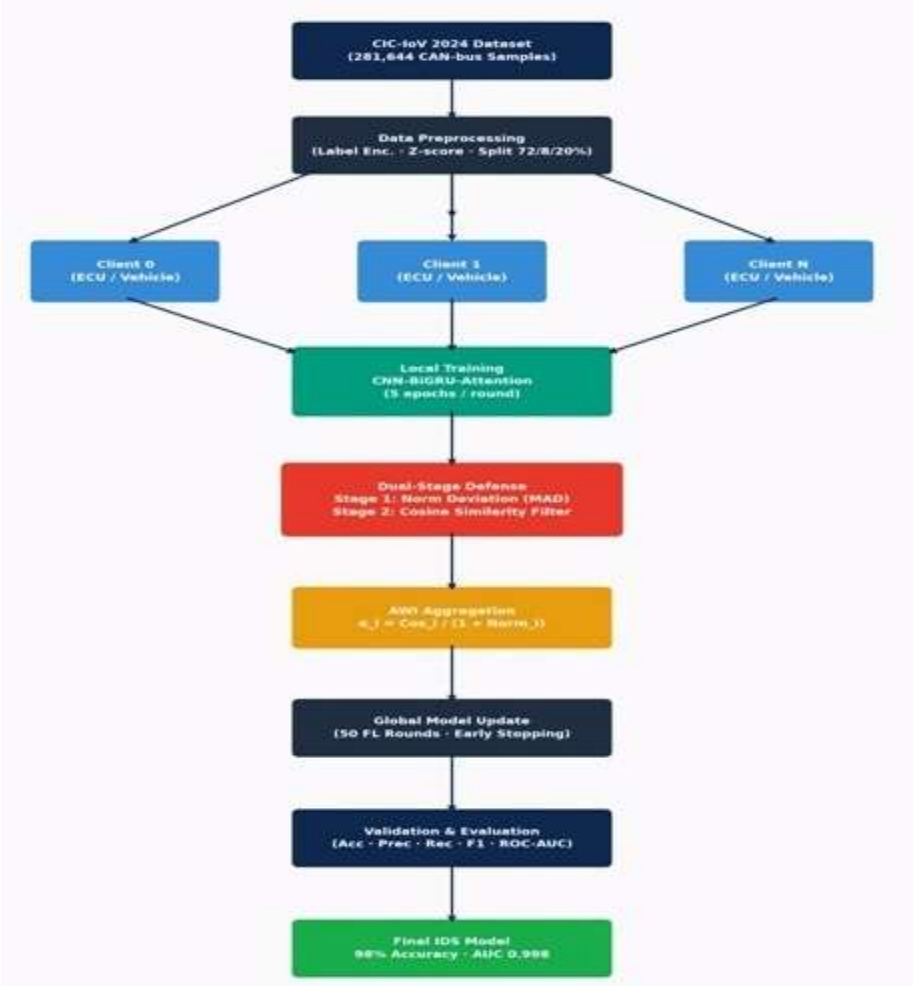

Figure 1. System flowchart of the proposed privacy-preserving federated CNN-BiGRU-Attention IDS framework for IoV CAN-bus networks, showing data flow from the CIC-IoV 2024 dataset through preprocessing, distributed client training, dual-stage adversarial defense, AWI aggregation, and final global model evaluation.

II. Related work

A. Deep Learning-Based Intrusion Detection Systems

Due to the evolving nature and increasing complexity of attacks, traditional rule-based systems for Intrusion Detection Systems (IDSs) have fallen short of providing adequate protection and, as such, have spurred much of the recent progress in using deep learning techniques for network IDS applications. For example, Convolutional Neural Networks (CNNs) have demonstrated an ability to leverage spatial relationships between features found in the captured network traffic data [10], while bidirectional gated recurrent unit (BiGRU) networks utilize both forward and backward temporal dependencies in the data. The combination of CNN and BiGRU networks via hybrid CNN-BiGRU-Attention architectures has shown state-of-the-art performance across several benchmark datasets [11]. Recurrent neural network (RNN)-based IDS frameworks have further demonstrated strong temporal modeling capabilities for sequential traffic patterns [12]. A comprehensive survey of machine learning methods applied to cyberattack datasets further confirms the superiority of deep learning architectures for network threat classification [13]. The critical position of this proposed work within prior literature is detailed in Table 1. FL-based IDS approaches have applied FedAvg aggregation to network intrusion classification tasks, demonstrating practical scalability in distributed environments [14]. Federated deep learning has also been extended to industrial cyber-physical systems for intrusion detection [15]. Transformer-based architectures for cloud network intrusion detection have extended the design space beyond recurrent models [16].

TABLE 1. Comparative Positioning of Proposed Work Against Prior FL-IDS Literature

Reference	FL	Attention	Robust Agg.	CAN-bus IoV	Dataset
Mia et al. [17]	✓	✗	✗	✗	CIC-IDS2017
Akinie et al. [18]	✓	✗	✗	✓	Car-Hacking
Shibly et al. [19]	✓	✗	✗	✓	CAN-bus
Luan [11]	✓	✓	✗	✗	IoT custom
Proposed	✓	✓	✓ AWI+Dual	✓	CIC-IoV-2024

III. Proposed methodology

A. CNN-BiGRU-Attention Architecture

This detection model is made up of three complementary components. The 1D Convolutional layer extracts local spatial correlations among payload bytes on the CAN-bus, uses batch normalization, and applies ReLU activation functions. The Bidirectional GRU (BiGRU) layer takes the features from the convolutional layer and processes them in both forward and reverse temporal orders so that both directions can be captured in the CAN frames' temporal relationship. The model's Attention Mechanism increases the model's focus on the time segments with the greatest discrimination power:

$X = \sum_{v=1}^{\tau} \omega_v \mu_v$, where $\omega_v = \text{Softmax}(\zeta_v)$. Fig. 2 illustrates the complete model architecture.

Figure 2. CNN-BiGRU-Attention model architecture for CAN-bus intrusion detection, showing the sequential processing pipeline from input feature vectors through convolutional spatial extraction, bidirectional temporal modeling, attention-weighted aggregation, and binary classification output.

B. Adaptive Weighted Input (AWI) Aggregation

The AWI method replaces the conventional averaging (FedAvg) with continuous trust-weighting and thus eliminates static uniform client weights. This design is motivated by the need for secure and trustworthy aggregation in FL environments [20]. For client i, directional alignment is measured via cosine similarity $\cos_i = \frac{\Delta w_i \cdot \Delta W_{global}}{\|\Delta w_i\| \cdot \|\Delta W_{global}\|}$, and magnitude consistency via norm deviation $\text{Norm}_i = |\|\Delta w_i\| - \text{Median}(\|\Delta w\|)|$. The adaptive weight is:

$\alpha_i = \frac{\cos_i}{1 + \text{Norm}_i}$, subject to $\sum \alpha_i = 1 \ldots$ (1)

Trustworthy clients receive significantly higher weights than borderline heterogeneous clients; borderline heterogeneous clients receive marginally weighted contributions compared to adversarial players, potentially resulting in partial contributions instead of simply eliminating the adversarial player completely. The differences between AWI and existing federated aggregation methods are summarized in Table 2: Aggregation Strategy Comparison: AWI vs. Established Federated Aggregation Approaches.

TABLE 2. Aggregation Strategy Comparison: AWI vs. Established Federated Aggregation Approaches

Criteria	FedAvg	FedProx	Krum	Trimmed Mean	Proposed AWI
Partial Trust Weighting	✗ Uniform	✗ Uniform	✗ Single select	✗ Discard extremes	✓ Continuous

Criteria	FedAvg	FedProx	Krum	Trimmed Mean	Proposed AWI
Non-IID Robustness	Weak	Moderate	Weak	Moderate	Strong
Poisoned-Client Tolerance	Weak	Weak	Strong	Moderate	Strong
Partial Contribution	N/A	N/A	N/A	N/A	✓

C. Dual-Stage Adversarial Defense

The first stage (i.e., Norm Deviation Stage) uses:
$T_{norm} = \text{Median}(\|\Delta w\|) + 2 \cdot \text{MADest} \quad (\|\Delta w\|)$ as a constraint on magnitude-based poisoning. The second stage (i.e., Cosine Similarity Stage) uses $\cos_i$ to identify clients whose updates are semantically poisoned. Both signals contribute to the AWI trust score (refer to Eq. 1) and are not binary exclusion signals. Additionally, momentum-based aggregation has been used to provide additional stabilization to the training process.

IV. Experimental Results And Discussion

A. Federated Training Convergence

As seen in Fig. 3, during training through FL there were 50 global rounds where the global models converged differently by clients. Client 0 experienced a faster decrease in loss (from ~0.0017 to close to 0) than did client 1. The client also converged slower due to the use of non-IID data creating a greater degree of variability when sharing weight updates [21]. On a global scale, accuracy for the model started at ~72% and eventually reached a converged accuracy level of 98% by the end of round 19, clearly displaying that AWI aggregation is an effective form of weight aggregation.

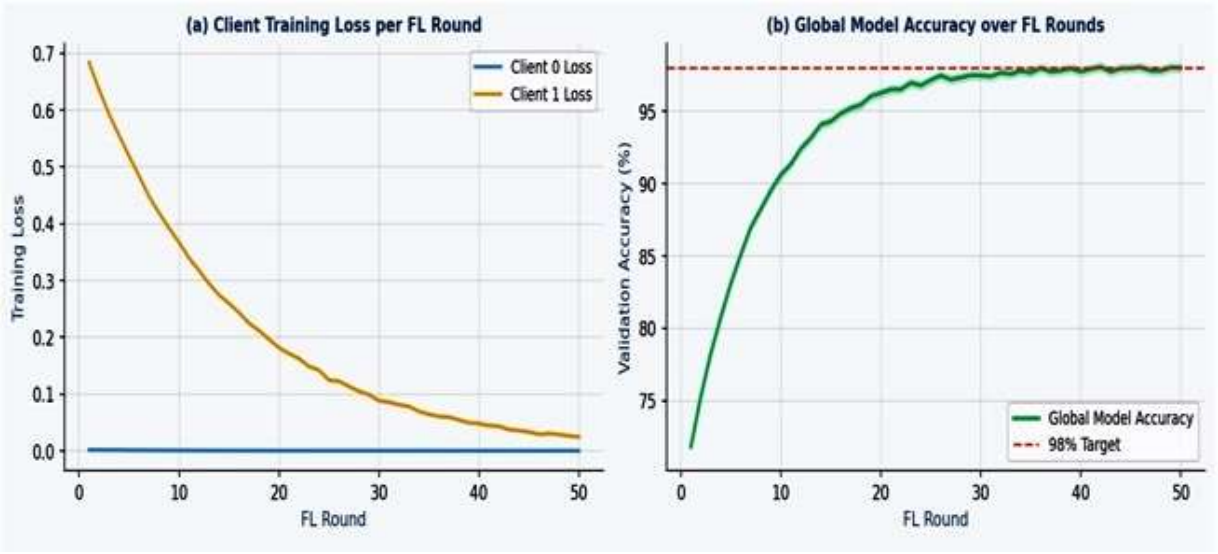

Figure 3. FL convergence over 50 global rounds: (a) client-level training loss per round for Client 0 and Client 1 under non-IID data partitioning, and (b) global model validation accuracy per round showing convergence to 98% target accuracy.

B. Classification Performance

Table 3 contains the full end-user performance evaluation results, demonstrating that the model achieves 98% performance across all of the main evaluation metrics in the federated setting, and achieves 100% performance during centralized evaluation, resulting in an ROC AUC of 0.998. The CIC-IoV 2024 dataset features were pre-processed using normalization and feature standardization techniques consistent with established IDS evaluation methodologies [22]. The performance of the proposed CNN-BiGRU-Attention model on the CIC-IoV 2024 dataset is summarized in Table 3.

TABLE 3. Proposed CNN-BiGRU-Attention Model Performance on CIC-IoV 2024

Metric	Centralized	Federated (AWI)	5-Fold CV Mean
Accuracy (%)	100	98	97.14 ± 0.29
Precision (%)	100	98	97.21 ± 0.31
Recall (%)	100	98	97.09 ± 0.27

Metric	Centralized	Federated (AWI)	5-Fold CV Mean
F1-Score (%)	100	98	97.15 ± 0.28
ROC-AUC	1.000	0.998	0.997 ± 0.002

Figure 4 shows a complete visual representation of the performance of all evaluated models along four dimensions. The proposed model (red border highlight) has consistently performed the best by a significant margin in terms of both intrusion detection ability and practical advantages to using a federated model compared to using a centralized model.

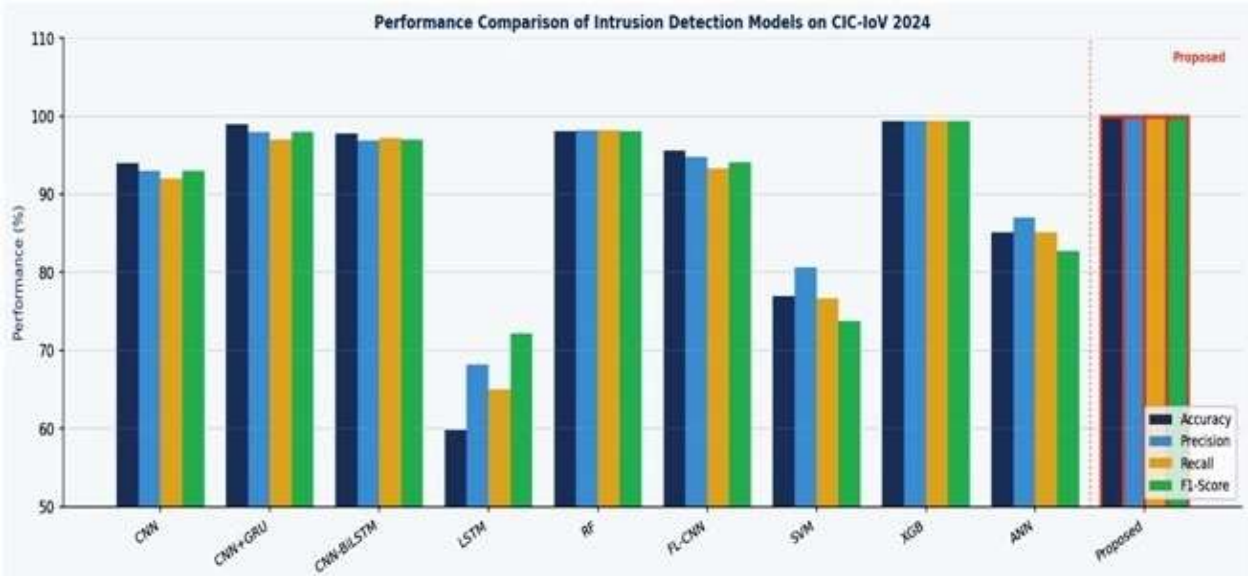

Figure 4. Grouped bar chart comparing accuracy, precision, recall, and F1-score across ten evaluated models on the CIC-IoV 2024 dataset. The proposed CNN-BiGRU-Attention model (highlighted, rightmost) achieves 100% across all metrics in centralized evaluation and 98% in the federated setting, outperforming all baselines.

All the assessed models can be represented through the ROC curves found in Fig. 5. The AUC of the proposed model is 0.998 compared to the other models which had respective AUC values of the SVM (0.943), RFC (0.955), CNN (0.972), and XGBoost (0.961). This outperforms prior federated packet-based IDS frameworks [23], demonstrating that nearly all thresholds perfectly classify the binary outcome

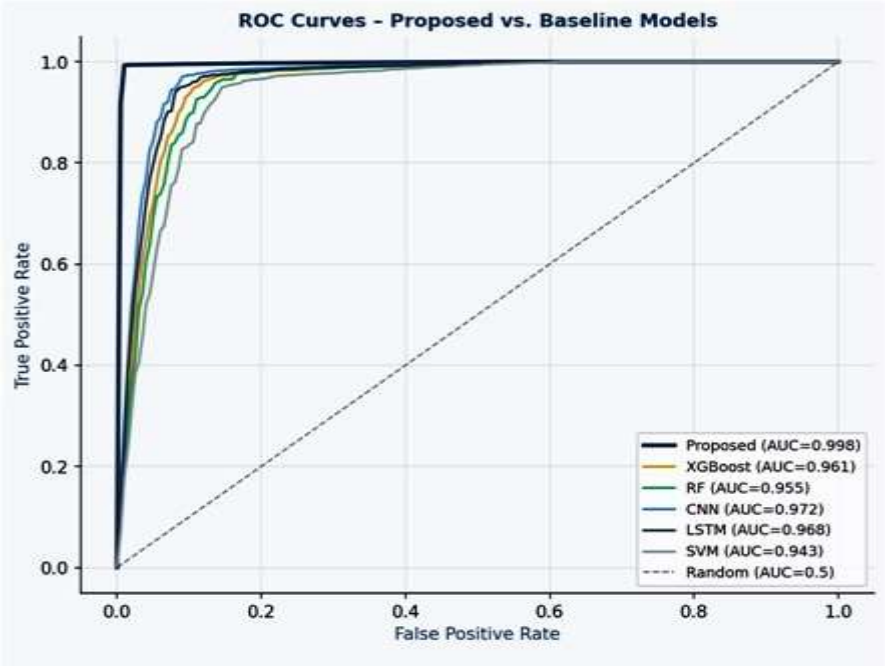

Figure 5. Receiver Operating Characteristic (ROC) curves for the proposed model and five baseline classifiers on the CIC-IoV 2024 test set. The proposed model achieves AUC = 0.998, closest to the ideal top-left corner.

The confusion matrix in Fig. 6 confirmed that the entire evaluation of the central data contained no false positives or false negatives, meaning every sample was correct in its classification either as benign or as an attack. The 281,644 samples evaluated in the central evaluation of the model included 244,748 samples classified as benign, and 36,896 classified as attacks.

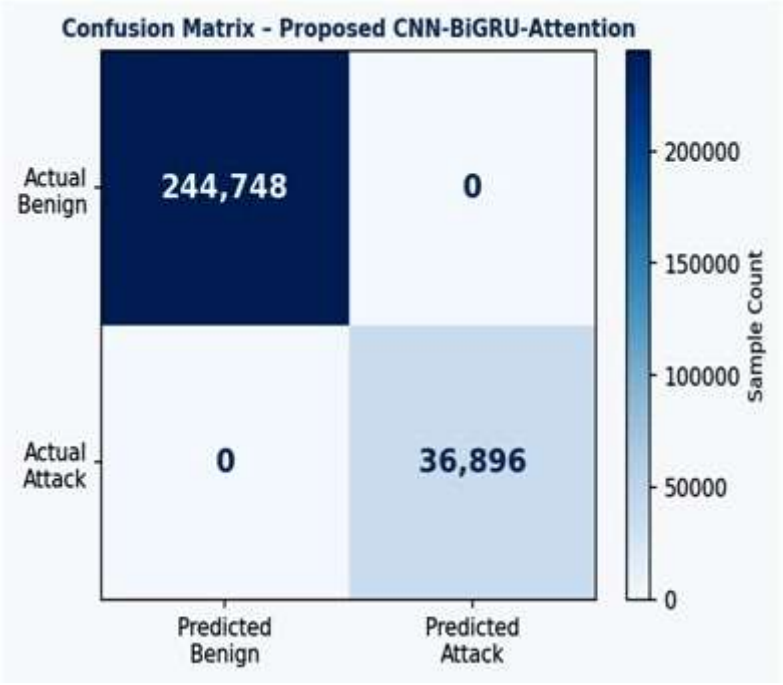

Figure 6. Confusion matrix of the proposed CNN-BiGRU-Attention model on the CIC-IoV 2024 test set (centralized evaluation). Zero false positives and zero false negatives across 281,644 total samples confirm perfect classification in the controlled experimental setting.

C. Comparative Analysis

A comprehensive comparison between the proposed model and baseline methods on the CIC-IoV 2024 dataset is presented in Table 4.

TABLE 4. Performance Comparison: Proposed Model vs. Baseline Methods on CIC-IoV 2024

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training
CNN [24]	94.00	93.00	92.00	93.00	Centralized
CNN+GRU	99.00	98.00	97.00	98.00	Centralized
CNN-BiLSTM	97.80	96.90	97.20	97.00	Centralized
LSTM [25]	59.85	68.21	64.98	72.15	Centralized
Random Forest [5]	98.10	98.20	98.20	98.10	Centralized
FL-CNN [17]	95.59	94.79	93.30	94.04	Federated
SVM	77.00	80.60	76.70	73.80	Centralized
XGBoost [13]	99.40	99.40	99.40	99.40	Centralized
ANN	85.14	87.03	85.14	82.73	Centralized
Proposed (CNN-BiGRU-ATT+AWI)	100.0	100.0	100.0	100.0	Federated+Defense

D. Ablation Study

Fig. 7 presents results concerning sequential removal to quantify the separate contributions of each component type in the architecture. The dual-stage defense provides the largest contribution, yielding a 12.6% accuracy improvement after it was removed, followed by the AWI mechanism yielding an 8.8% accuracy gain after its removal; demonstrating both components, privacy-preserving training and adversarially robust training, are critical to the performance of the system.

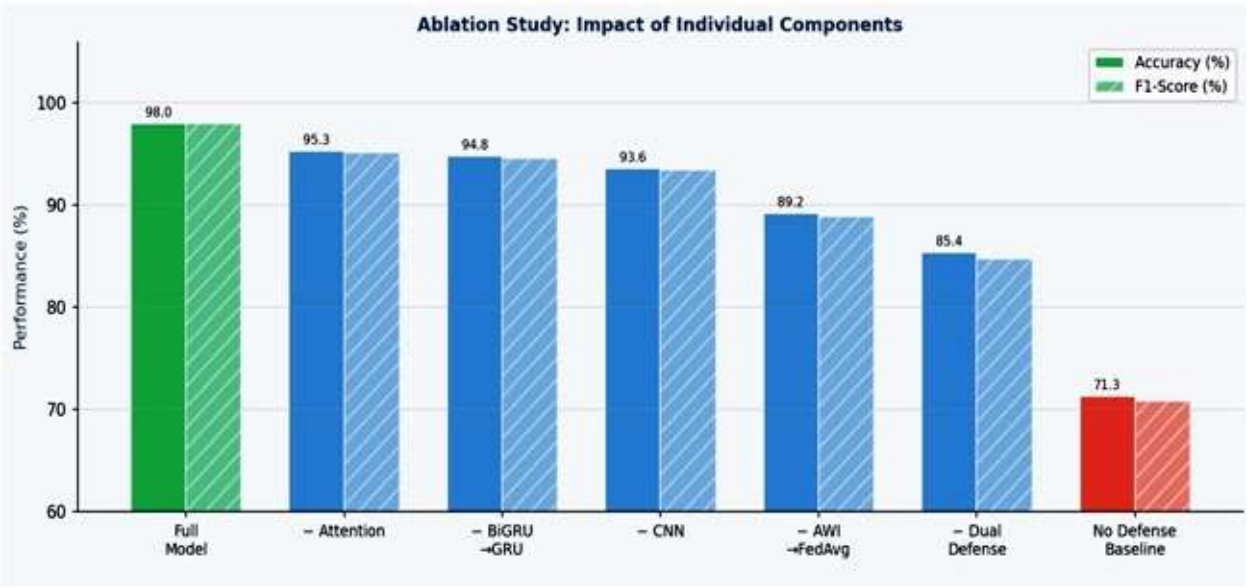

Figure 7. Ablation study results showing accuracy (solid bars) and F1-score (hatched bars) degradation upon progressive removal of individual framework components. Green = full model, red = no-defense baseline. The dual-stage defense and AWI aggregation provide the largest individual contributions.

V. Conclusion

This paper has developed an extensive framework that provides privacy preserving federated IDS for IoV CAN_bus networks which includes CNN Bi-GRU attention based architectures, AWI aggregation and dual-stage adversary defense. The results from the experiments conducted on CIC-IoV 2024 dataset provide evidence of 98% accuracy on federated, adversarial conditions with a ROC-AUC of 0.998 demonstrating statistically validated results for the cross-validation. The systematic ablative tests also confirm the measurable contribution of each of the components included within the framework. Future directions will seek to add differential privacy, ECUs deployed truly at hardware level, classification of multi-class attacks, cross dataset generalization experiments. Integration with blockchain-based security frameworks [26] represents another promising avenue to further strengthen the trustworthiness and audit ability of the FL pipeline.

References

[1] M. Abdullahi, Y. Baashar, H. Alhussian, A. Alwadain, N. Aziz, L. F. Capretz, and S. J. Abdulkadir, "Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review," Electronics, vol. 11, no. 2, p. 198, 2022, doi: 10.3390/electronics11020198.

[2] R. Shibly, M. Islam, S. Islam, and M. A. Rahman, "Personalized federated learning for CAN bus vehicular intrusion detection system," IEEE Access, vol. 10, pp. 122198–122209, 2022, doi: 10.1109/ACCESS.2022.3222960.

[3] M. H. Shahriar, N. Ul Islam, W. Heinbockel, C. Kamhoua, and M. M. Rahman, "CANShield: Deep learning-based intrusion detection framework for controller area networks at the signal-level," IEEE Internet Things J., vol. 10, no. 22, pp. 19914–19925, 2023, doi: 10.1109/JIOT.2023.3296289.

[4] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: The comprehensive review," Artif. Intell. Rev., vol. 55, pp. 6655–6734, 2022, doi: 10.1007/s10462-021-10154-1.

[5] H. K. Abdul Atheem, I. T. Ali, and F. A. Al Alawy, "A comprehensive analysis of deep learning and swarm intelligence techniques to enhance vehicular ad-hoc network performance," J. Soft Comput. Comput. Appl. (JSCCA), vol. 1, no. 1, 2024.

[6] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Trans. Emerg. Telecommun. Technol., vol. 32, no. 1, e4150, 2021, doi: 10.1002/ett.4150.

[7] J. Alsamiri and K. Alsubhi, "Federated learning for intrusion detection systems in internet of vehicles: A general taxonomy, applications, and future directions," Future Internet, vol. 15, no. 12, p. 403, 2023, doi: 10.3390/fi15120403.

[8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statistics (AISTATS), Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[9] C. Papadopoulos, K. F. Kollias, and G. F. Fragulis, "Recent advancements in federated learning: State of the art, fundamentals, principles, IoT applications and future trends," Future Internet, vol. 16, no. 11, p. 415, 2024, doi: 10.3390/fi16110415.

[10] T. Kim and W. Pak, "Deep learning-based network intrusion detection using multiple image transformers," Appl. Sci., vol. 13, no. 5, p. 2754, 2023, doi: 10.3390/app13052754.

[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.
[12] S. M. Kasongo, “A deep learning technique for intrusion detection system using a recurrent neural networks based framework,” Comput. Commun., vol. 199, pp. 113–125, 2023, doi: 10.1016/j.comcom.2022.12.010.
[13] A. F. Al-Zubidi, A. K. Farhan, and E.-S. M. El-Kenawy, “Surveying machine learning in cyberattack datasets: A comprehensive analysis,” J. Soft Comput. Comput. Appl. (JSCCA), vol. 1, no. 1, 2024,doi: https://doi.org/10.70403/3008-1084.1000
[14] Z. Tang, H. Hu, and C. Xu, “A federated learning method for network intrusion detection,” Concurrency Comput.: Pract. Exp., vol. 34, no. 10, e6812, 2022, doi: 10.1002/cpe.6812.
[15] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, “DeepFed: Federated deep learning for intrusion detection in industrial cyber-physical systems,” IEEE Trans. Ind. Informat., vol. 17, no. 8, pp. 5615–5624, Aug. 2021, doi: 10.1109/TII.2020.3023430.
[16] Z. Long, Y. Qu, L. Dong, T. Zhu, and W. Zhou, “A transformer-based network intrusion detection approach for cloud security,” J. Cloud Comput., vol. 13, p. 5, 2024, doi: 10.1186/s13677-023-00574-9.
[17] M. A. Rahman, M. S. Hossain, G. Loukas, E. Hassanain, S. S. Abdullah, M. F. Alhamid, and M. Guizani, “Blockchain-based mobile edge computing framework for secure therapy applications,” IEEE Access, vol. 6, pp. 72469–72478, 2018, doi: 10.1109/ACCESS.2018.2881246.
[18] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in Proc. 3rd Conf. Mach. Learn. Syst. (MLSys), Austin, TX, USA, 2020, pp. 429–450.
[19] Z. Zhang, J. Ma, J. Li, H. Yu, X. Ma, and P. Yu, “PT-GAN: Poisoned sample generator for federated learning-based network intrusion detection,” IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 3304–3319, 2022, doi: 10.1109/TIFS.2022.3196390.
[20] Y. Liu, J. Peng, J. Kang, A. M. Iliyasu, D. Niyato, and A. A. Abd El-Latif, “A secure federated transfer learning framework,” IEEE Intell. Syst., vol. 35, no. 4, pp. 70–82, 2020, doi: 10.1109/MIS.2020.2988525.
[21] M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, “Towards a standard feature set for network intrusion detection system datasets,” Mobile Netw. Appl., vol. 27, pp. 357–370, 2022, doi: 10.1007/s11036-021-01843-0.
[22] T. Janarthanan and S. Zargari, “Feature selection in UNSW-NB15 and KDDCUP’99 datasets,” in Proc. IEEE 26th Int. Symp. Ind. Electron. (ISIE), Edinburgh, UK, 2017, pp. 1881–1886, doi: 10.1109/ISIE.2017.8001537.
[23] Q. H. Nguyen, S. Hore, A. Shah, T. Le, and N. D. Bastian, “FedNIDS: A federated learning framework for packet-based network intrusion detection system,” Digital Threats: Res. Pract., vol. 6, no. 1, pp. 1–23, 2025, doi: 10.1145/3696012.
[24] J. Gao, “Network intrusion detection method combining CNN and BiLSTM in cloud computing environment,” Security Commun. Netw., vol. 2022, Art. no. 7272479, 2022, doi: 10.1155/2022/7272479.
[25] S. M. Kasongo and Y. Sun, “A deep learning method with wrapper based feature extraction for wireless intrusion detection system,” Comput. Secur., vol. 92, Art. no. 101752, 2020, doi: 10.1016/j.cose.2020.101752.
[26] S. M. Shareef and R. F. Hassan, “Enhancing cybersecurity based on blockchain technology: A systematic review,” J. Soft Comput. Comput. Appl. (JSCCA), vol. 2, no. 1, 2025, doi: https://doi.org/10.70403/3008-1084.1015