Privacy-Preserving Anomaly Detection in Smart Cities Leveraging Federated Learning

Boyun Eom
Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, Rep of Korea
eby@etri.re.kr

Dong-Hwan Park
Industrial Energy Convergence Research Department
Electronics and Telecommunications Research Institute
Daejeon, Rep of Korea
dhpark@etri.re.kr

***Abstract*—To address growing privacy concerns in data collection and sharing, federated learning (FL) has gained increasing attention as a decentralized machine learning paradigm. In particular, FL has emerged as a viable solution for smart cities, where the safety is paramount and privacy must be preserved. In this paper, we introduce our study on adapting FL to a smart city use case involving anomaly detection in surveillance systems. To enhance model robustness under non-IID conditions, we design and evaluate two novel aggregation strategies built upon the baseline FedAvg algorithm. Our approach is validated using a CCTV dataset featuring abnormal behaviors relevant to smart city environments, demonstrating the potential of FL for privacy-preserving video analytics in urban infrastructure.**

Keywords— machine learning, federated learning, human motion detection, smart city

I. INTRODUCTION

The Smart City, which has enhanced urban efficiency and security intelligence, relies on the integration of advanced technologies such as the Internet of Things. Building on this foundation, the rapid growth of artificial intelligence (AI) has enabled extensive analysis and prediction using the abundant IoT-generated data [1-2]. One compelling application in this context is anomaly detection in public spaces (e.g., parks, streets, transit hubs) using closed-circuit television (CCTV) systems.

However, the widespread deployment of CCTV cameras and other surveillance technologies raises substantial privacy concerns. CCTV systems typically transmit raw video footage to a central server for processing, which risks exposing sensitive personal information. This centralization not only makes the data vulnerable to security breaches but also raises ethical questions regarding individuals' right to privacy.

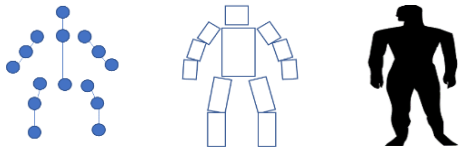
In fact, in countries with strict privacy regulations-such as South Korea's Personal Information Protection Act-such centralized data practices are often legally restricted or outright prohibited. As a result, the increasing deployment of AI-powered surveillance systems in smart cities necessitates the adoption of decentralized and privacy-preserving AI methods. Motivated by this challenge, we propose a privacy-preserving approach that leverages federated learning (FL) to train AI models directly at CCTV sites.

The remainder of this paper is structured as follows. Section 2 provides a review of related work in human action recognition and federated learning. Section 3 describes our methodology, including data preparation and aggregation algorithms used in the federated learning scheme. Section 4 presents and discusses the experimental results and their implications. Finally, Section 5 concludes the paper with a summary.

II. RELATED WORK

A. Human Action Recognition

In the field of computer vision, pose estimation-based human action recognition is a fundamental task with wide-ranging applications, including abnormal behavior detection. There are three traditional approaches to human pose estimation: skeleton-based, contour-based, and volume-based [3]. Among these, skeleton-based techniques have gained significant attention due to their efficiency in representing human body motion through key joint coordinates. These models represent human poses as a set of connected joints forming a skeletal structure, which effectively captures body motion. The extracted features are then fed into temporal models such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), or Graph Convolutional Networks (GCNs) for action classification.



(a) Skeleton (b) Contour (c) Volume

Fig. 1. Conventional human pose estimation methods

Popular frameworks for extracting skeletal features from video frames include OpenPose [4] and the High-Resolution Network (HRNet) [5], both of which are convolutional neural network (CNN)-based architectures that have demonstrated state-of-the-art performance in terms of accuracy and speed.

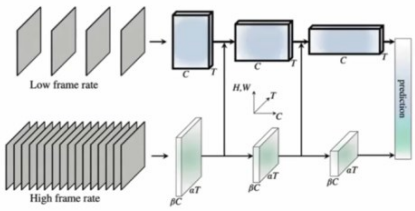


Fig. 2. Slowfast Network [6]

By contrast, RGB-based video architectures like the SlowFast Network [6] leverage a dual-pathway design to capture multi-scale temporal dynamics. It is particularly capable of modeling varying temporal dynamics present in video data. The network operates with two parallel pathways as shown in Fig.2 : the Slow pathway processes inputs at a low frame rate to capture features with minimal temporal change, emphasizing long-term spatiotemporal relationships, while the Fast pathway processes inputs at a high frame rate to capture rapidly changing features, focusing on short-term temporal details. This dual-path design enables the model to effectively

capture both coarse and fine-grained temporal structures in videos, resulting in superior performance in tasks such as action recognition, video classification, and other video analysis applications. By integrating the outputs of these two pathways, the SlowFast Network achieves more accurate and temporally robust predictions.

B. *Aggregation Algorithm in Federated Learning*

In federated learning, multiple clients collaborate to train a global model without exposing their raw data[7]. This approach is especially beneficial in scenarios where sensitive data is distributed across multiple devices or organizations. A critical component influencing federated learning performance is the server-side aggregation algorithm.

FedAvg [8] is the original aggregation algorithm in FL and it constructs a global model by computing a weighted average of the clients' local model parameters, where each client's contribution is proportional to the number of data samples. Although there has been extensive study on aggregation method in FL, FedAvg is still favored due to balance between performance and communication cost.

Several well-known aggregation algorithms, including FedAvg, FedProx, FedMedian and q-FedAvg across different scenarios are well compared in [9].

III. Methodology And Experimental Setup

A. *FL Framework*

Flower, an open-source framework designed for federated learning, was used to implement our experiments[10]. A single FL server and two clients were deployed with CUDA support to compare the performance of different aggregation algorithms under consistent conditions.

Although federated learning typically involves a large number of clients, we conducted our experiment with only two clients to facilitate controlled analysis of aggregation performance under a deliberately constructed Non-IID(non-independent and identically distributed) data distribution. This setup allows for a focused examination of convergence dynamics and algorithmic differences in a simplified setting.

B. *Data Preparation*

To train local models on each client, we employed the Abnormal Behavior CCTV Footage dataset provided by AI-hub [11]. It was constructed to support training AI models to detect abnormal behaviors which are classified into 12 types, including assault, burglary, kidnap, etc. This public dataset contains approximately 700 hours (8,400 videos) of CCTV footage captured by public safety cameras and its correspondent annotation files.

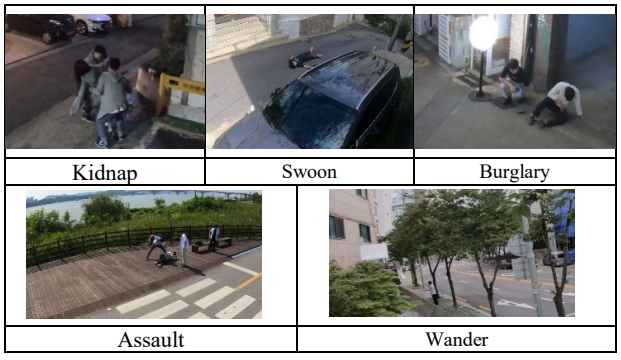


Fig. 3. Abnormal behavior CCTV footage dataset

To tailor the dataset to the smart park use case and facilitate manageable experimentation, we extracted a subset of five abnormal behavior categories from the full dataset. Fig. 3 presents representative frames from video footage corresponding to these selected classes. These video files were segmented into short clips using the provided annotation files to capture individual behavior instances. These preprocessed clips were then deliberately distributed across two clients with imbalanced class distributions, simulating a realistic Non-IID scenario for local model training.

TABLE I. Number of Video Clips

	Kidnap	Swoon	Burglary	Assault	Wander	Total
Client1	28	28	47	223	73	403
Client2	129	102	44	179	400	854

Since each client collects data under different circumstances, a well-known challenge in the federated learning scheme is the Non-IID nature of the data [7]. In our setting, two types of data skew are configured: label distribution skew and quantity skew. The classes and the number of video clips assigned to each client in the experiments are shown in Table 1 and Fig. 4 illustrates the imbalanced data distribution across those two clients.

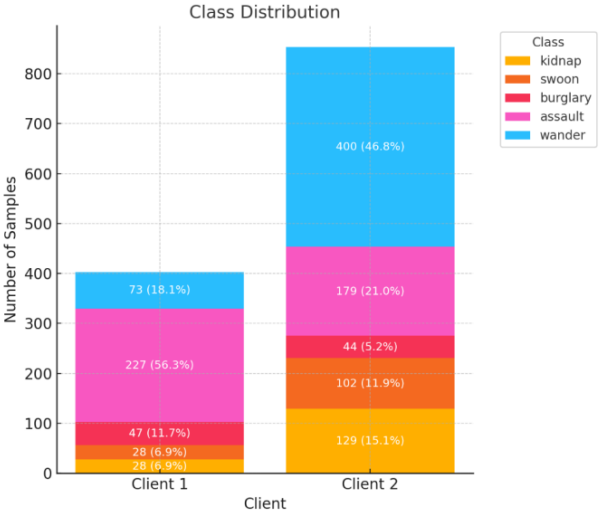


Fig. 4. Imbalanced data distribution

C. *Client-Side Training and Server-Side Aggregation*

For local training on each client, we fine-tuned SlowFast network to perform detection of abnormal behaviors. We employed the Adam optimizer along with a categorical cross-entropy loss function. Table 2 shows the hyperparameters used during our experiments.

TABLE II. HyperParameters

	Learning Rate	**Dropout**	**Batch size**	**Epochs**
Value	0.0001	True	32	10

For the aggregation of models on the server, we developed two variants of FedAvg to improve upon its limitations. FedAvg aggregates the global model parameters, θ_{global} , with the summation of local model parameters , θ_i , from K clients.

$$\theta_{global} = \sum_{j=1}^{K} \lambda_i \times \theta_i \quad (1)$$

In (1), λ_i represents the contribution of client i, which is calculated as the proportion of its sample size, as shown in (2):

$$\lambda_i = \frac{n_i}{\sum_{j=1}^{K} n_j} \quad (2)$$

One major drawback of FedAvg is the limitation in handling Non-IID data. Since FedAvg does not consider performance metrics during each federated round, we have made modifications to improve its performance on Non-IID situation. In this study, we have developed two aggregation algorithms built upon FedAvg; Fed-Acc+Loss and Fed-Metrics. In each federated round, Fed-Acc+Loss computes the contribution, λ' , of each client using validation accuracy and training loss as well as number of samples as shown in (3).

$$\lambda'_i = [\frac{1}{loss+\epsilon} \times \alpha + (1-\alpha) \times accuracy\ ] \times \frac{n_i}{\sum_{j=1}^{K} n_j} \quad (3)$$

where, ϵ stabilizes division by zero and we used 0.7 for α.

Accuracy is an intuitive measure of a client's model performance, while Loss reflects how closely the model's output matches the ground truth in a fine-grained manner, which can help guide the model's convergence.

Meanwhile, for better performance on skewed data and fairer and more reliable global model assessment, Fed-Metrics utilizes performance metrics such as precision, recall and f1-score for the contribution of each client every federated round. Reflecting the harmonic mean of these metrics, the contribution of Fed-Metrics, λ'' , for each client is defined as in (4).

$$\lambda''_i = (\frac{\text{precision}_i + \text{recall}_i + \text{F1\_Score}_i}{3}) \times \frac{n_i}{\sum_{j=1}^{K} n_j} \quad (4)$$

IV. Experimetal Result and Discussion

In this section, we present the evaluation results and discuss the main insights derived from the experiments.

Table 3 presents the experimental outcomes, and Fig. 5 illustrates the performance evolution of the three aggregation approaches.

TABLE III. Experimental Result

	Fl_Round	1	2	3	4	5	6	7	8	9	10
Fed-AVG	Client 1	64.63	70.73	70.73	68.29	71.95	76.83	91.46	89.02	82.93	93.9
	Client 2	64.33	40.35	45.03	46.2	45.61	74.27	66.08	63.74	92.4	91.23
Fed-Accuracy+loss	Client 1	65.85	69.51	65.85	68.29	67.07	68.29	80.49	84.15	79.27	79.27
	Client 2	33.33	53.8	43.86	42.11	49.12	63.16	67.25	87.13	87.72	70.76
Fed-Metrics	Client 1	74.39	68.29	71.95	71.95	69.51	71.95	78.05	79.27	79.27	**96.34**
	Client 2	52.05	67.25	63.74	53.22	44.44	47.95	63.74	83.04	90.06	**92.98**

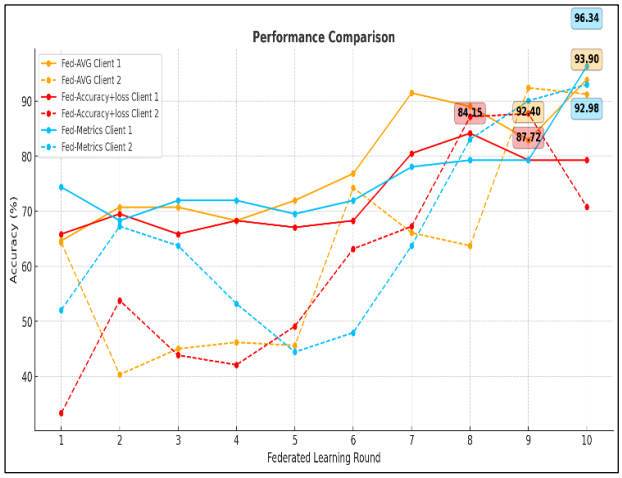
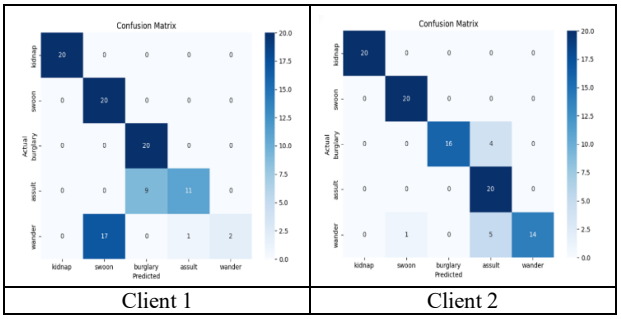


Fig. 5. Accuracy comparision

As shown, Fed-Metrics outperforms the other methods for both Client 1 and Client 2. Although Client 2 exhibits noticeable performance fluctuations, particularly in the earlier rounds, both clients show consistent improvement after federated round 5, indicating strong convergence. This can be attributed to the integration of precision, recall, and F1-score, which enables the global model to weigh clients' contributions more holistically based on fine-grained performance metrics rather than solely on data quantity or simple accuracy.

In contrast, although FedAvg performs well in the end, Fig.5. Illustrates significant performance instability of the method throughout training. While Client 1 had a more skewed label distribution, larger fluctuations were observed in Client 2. This may be due to the presence of quantity skew and the lack of performance-aware adjustments in FedAvg, which can exacerbate inconsistencies during local training in non-IID settings.

These findings support the view that FedAvg may be less effective in scenarios with unbalanced or non-representative client data, as it ignores model quality and relies exclusively on data size, resulting in less stable convergence in practice.



Client 1 | Client 2

Fig. 6. Confusion matrix

Fig. 6 presents the confusion matrix of Fed-Metrics. In this setting, although Client 1 achieves slightly higher accuracy, the confusion matrix for Client 2 demonstrates that Fed-Metrics provides more balanced classification performance across all classes.

V. Conclusion

This study demonstrates that federated learning enables the analysis of sensitive video data without centralization, thereby preserving individual privacy. We developed two FedAvg-based aggregation algorithms and evaluated their effectiveness, along with a baseline approach, in the context of privacy-preserving analysis of CCTV video data. Fed-Metrics, which integrates precision, recall, and F1-score into the aggregation process, proved to be the most robust and effective strategy, delivering high accuracy and stable convergence, and being well-suited for imbalanced or non-IID data distribution. In scenarios with heterogeneous or skewed data, metric-aware strategies like Fed-Metrics may offer the best trade-off between fairness and performance.

The experimental results validate the effectiveness of the proposed framework in maintaining a balance between accuracy and privacy, making it a promising solution for real-world surveillance systems.

Acknowledgment

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant

funded by the Ministry of Land, Infrastructure and Transport (Grant: RS-2022-00155803).

REFERENCES

[1] L. Zhou,X. Li,Y. Wang and J. Chen, “Adoption of artificial intelligence in smart cities: A comprehensive review,” in *Commun. Comput. Inf. Sci.*, vol. 199, AI-IoT 2022, Springer, Cham, 2022, pp. 1–25.

[2] H. Shin, K.-I. Na, J. Chang, and T. Uhm, “Multimodal layer surveillance map based on anomaly detection using multi-agents for smart city security,” *ETRI Journal*, vol. 44, no. 2, pp. 183–193, Apr. 2022

[3] W. Choi, T. Choi, and S. Heo, “A comparative study of automated machine learning platforms for exercise anthropometry-based typology analysis: Performance evaluation of AWS SageMaker, GCP VertexAI, and MS Azure,” Bioengineering, vol. 10, no. 8, art. 891, Jul. 2023

[4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1302–1310

[5] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5693–5703

[6] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “SlowFast networks for video recognition,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 6201–6210

[7] H. Zhu, J. Xu, S. Liu and Y. Jin, “Federated learning on non-IID data: A survey, Neurocomputing, vol. 465,pp. 371-390, 2021.

[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. Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282

[9] V. Agarwal, C. J. Chandnani, S. C. Kulkarni, A. Aren, and K. Srinivasan, “A comparative analysis of aggregation methods in federated learning on MNIST,” in *Commun. Comput. Inf. Sci.*, vol. 2228, *AIKP*, Springer, Dec. 2024, pp. 225–238.

[10] Flower: A Friendly Federated Learning Framework

[11] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=171