Federated Trustworthy AI Architecture for Smart Cities

Sapdo Utomo
Grad. Institute of Ambient Intelligence and Smart Systems
National Chung Cheng University
Chiayi County, Taiwan R.O.C.
sapdo.utomo@gmail.com

John A.
Dept. of Computer Science and Information Engineering
National Chung Cheng University
Chiayi County, Taiwan R.O.C.
johnmtech@gmail.com

Adarsh Rouniyar
Dept. of Computer Science and Information Engineering
National Chung Cheng University
Chiayi County, Taiwan R.O.C.
adarsh@csie.io

Hsiu-Chun Hsu
Dept. of Information Management
National Chung Cheng University
Chiayi County, Taiwan R.O.C.
ellenj1022@gmail.com

Pao-Ann Hsiung
Dept. of Computer Science and Information Engineering
National Chung Cheng University
Chiayi County, Taiwan R.O.C.
pahsiung@cs.ccu.edu.tw

***Abstract*— As the number of smart citizens or users grows, several governments are eager to develop smart cities. All stakeholders, including intelligent users, demand an AI system that is trustworthy. To be considered trustworthy, an AI system must meet seven key criteria: (1) human agency and oversight, (2) robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, nondiscrimination, and fairness, (6) societal and environmental well-being, and (7) accountability. By merging the existing trustworthy AI framework from KServe and Federated Learning with a more reliable model aggregation protocol than earlier studies, we introduced the federated trustworthy Artificial Intelligence (FTAI) architecture. With the integration of FedPSO and AIF360, we updated the FedCS technique. The proposed architecture meets all seven of the essential requirements to a high degree of satisfaction. This paper also includes a detailed explanation of this claim. The scenario demonstrates that the new global model comes with clear measures of fairness metrics, ensuring that the model is devoid of bias.**

Keywords—trustworthy AI, federated learning, smart city

I. Introduction

Current smart city applications mainly collect, analyzes, predict and controls the information based on huge amount of data collected form the edge computing devices. In this work, the focus is on an integrated architecture design for smart cities using Federated Learning and Trustworthy AI (TAI). This architecture includes powerful computing and communication capabilities with the guarantee of secure and trustworthy and controlling resource efficiency in a robust and reliable high availability manner. In this work federated learning and TAI play a vital role for decentralized local data training and sharing without data and trust established in each stage of smart city applications.

The idea of TAI is used to establish a full potential foundation of trust in AI and to build sustainable applications using ethical principles globally. The AI-HLEP (AI- High-Level Expert Group) of the European Commission has established specific rules pertaining to the making of trustworthy AI [1], [2]. Accordingly, the main principles of TAI demand that the purpose of the AI should revolve around the augmentation of human life and living intelligence, the insights and data should belong to the creator; and the technology must be explainable and transparent. As per the AI HLEP, seven basic ethical principles have been laid out for the development of TAI: human agency, robustness and safety, data privacy and governance, transparency, diversity and fairness, social well-being and accountability [2]–[4]. The main objective behind a protocol for trustworthy AI is to instill trust and confidence in AI-based automation systems. In a similar context, Scott Thiebes et al., [5] described what engenders trustworthiness in terms of information technology and people. There is no doubt that one way of building credibility around information technology characteristics and automation systems is by incorporating trusted artificial intelligence into them. Federated learning is a machine learning method that trains the data across multiple samples without any local information exchange [6], [7]. The characteristics of federated learning that stand out are universality in terms of scenarios, massively independent distribution, decentralization, and node equality [8]. The design of federated learning is an important issue, and as far as the architecture of a federated learning system is concerned, specific considerations regarding the topology, protocol, device, decentralized server, analysis, privacy, workflow application, operational profile, etc. have to be made [9], [10].

The basic goal of smart cities is to maximize the advantages for all stakeholders by optimizing the use of available resources (e.g. residents, governments, service providers, and businesses). To accomplish this goal, stakeholders must first adapt to and trust smart city services and solutions. Many factors influence the widespread adoption of smart city services, the most important of which are service dependability and security [11]. Furthermore, the long-term planning of today's smart cities requires the sustainable use of smart city resources. The availability of infrastructure that will support efficient service offerings to customers is another important challenge for smart cities. This infrastructure will need to include powerful computer and communication capabilities with the guarantee of secure and trustworthy. Typically, such computer capabilities act as the brain, supervising and controlling resource efficiency in a stable (robust) and reliable (high availability) manner. Communication capabilities, on the other hand, are seen as the primary means of contact with users and other stakeholders. In the propose architecture provide powerful computing and communication capabilities with the guarantee of secure and trustworthy and controlling resource efficiency in a robust and reliable high availability manner.

This work was supported partially by Research Project Grants from the Ministry of Science and Technology, Taiwan MOST110-2634-F-194-006 and MOST110-2927-I-194-001.

2022 IEEE International Smart Cities Conference (ISC2) | 978-1-6654-8561-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISC255366.2022.9922069

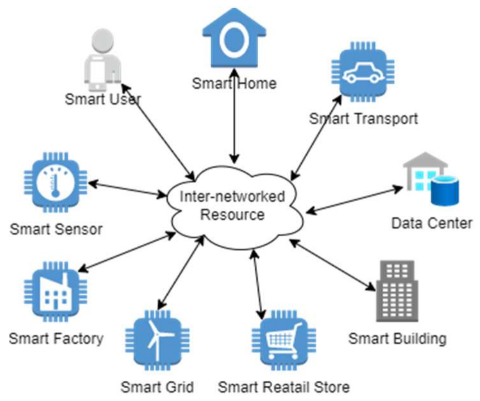

Fig. 1. A typical smart city ecosystem [11]–[13]

For example, the smart city framework and its stakeholders are depicted in Figure 1. A smart city often has a significant number of smart residents or smart users, as well as smart buildings, smart transportation, various sensing capabilities, and advanced computer and communication infrastructures. A substantial amount of user data is predicted to be generated by smart buildings and sensor networks. In most circumstances, quick processing capabilities will be required to generate vital time-sensitive decisions. Another issue is the security and privacy of user data. Smart city services should be built to support the sustainability of existing resources and trustworthiness in order to address the aforementioned issues.

The rest of the paper is organized as follows. Related works in terms of smart city framework design, issues and evaluation metrics are discussed in section 2. The functional requirement of Architecture has been described section 3. The detailed design of the Federated Trustworthy AI Architecture for Smart Cities has been presented in section 4 and finally, the conclusion is presented in section 5 of the paper.

II. Related Work

Smart city architecture and framework design are important aspects of smart city applications. Many researchers have described various issues and parameters required for designing smart city applications. The authors of [12] described the federated learning challenges and opportunities in a smart city environment. This work summarized the privacy, security, and data collection issues with regard to federated learning. The authors of [14] presented initial design aspects of federated learning using unlabeled data in smart cities. In this work, a semi-supervised edge learning method called FedSem, used in unlabeled real-time data, showed an improvement in accuracy by 8% with the use of labeled data.

Further consideration of data for increased accuracy was made by the author of [15], describing how built-in sensors are used in smartphones for collecting data from transportation networks, health services, and emergency services using smart city infrastructure. The main challenges of this work device were that it was self-activated and generated all the available data. Therefore, the size of the data was mounted in a haphazard structure. The authors of [16] proposed federated learning for real time traffic prediction using roadside units. Local data was used and distributed on edge with high learning ranges and further recommendations were made for investigating different models using different dynamic parameters. The authors of [17] proposed an aggregated approach using federated learning for vehicle traffic predictions and forecasts. In this work, decentralized local data was used for traffic prediction and the trained data was securely shared to the server, and for the implementation, only spatial parameters were used for the training of local data.

The authors of [18] proposed federated learning and Block chain used for prediction of traffic flow and to avoid data poisoning attacks. Using this proposed framework, the privacy of data was enhanced while prediction of traffic flow was improved. The authors [19] proposed a framework for traffic prediction with enhanced data privacy using FedGRU model, and this method was found to reduce communication overhead issues during the transmission of data and utilized ensemble-based clustering for location-based prediction of data. For the implementation, a real time dataset was used, and a 90.96% accuracy achieved, which was better than many advanced deep learning techniques. Furthermore, only spatial features were used and recommended in this work using Graph Convolutional Network with federated learning for better spatial- temporal features.

The authors of [20] described decentralized and distributed federated learning with UAV sensors for monitoring and made and future predictions on air quality and pollution. In this work, UAV sensors were used to collect the air quality data and LSTM used to process the data. The effectiveness of the model was evaluated using other machine learning models and data tested was collected from the capital of India. The authors of [21] similarly proposed a model for air pollution prediction using a federated learning paradigm and designed a distributed framework used cooperative training from various spatial areas. Each area CRNN was trained locally and predicted local warning levels for all regions. Using this method, newly portioned areas were locally trained and transferred to the global model. Both air pollution models [20], [21] considered spatial features for prediction and forecasting of data. The authors of [22] proposed federated sensing model using real time sensing data for water quality prediction. In this work, distributed observations with geologically used local models and parameters were considered for local and global predictions.

The authors of [23] summarized various federated learning frameworks and their applicability. The authors evaluated the development, deployment, accuracy, capability, and performance as well as the applicability in IoT models. Three datasets comprising of two signal datasets and one image dataset were used for the implementation. In this work, TFF (TensorFlow Federated) [24], FATE [25], PFL from Baidu [26], PySyft from OpenMinded [27], and FL&DP from Sherpa [28] models were investigated using the signal and image datasets. It was observed that the TF (TensorFlow) produced better accuracy while using the image datasets, and the TFF produced better results while using the signal dataset.

The authors of [29] proposed a lossless data privacy federated framework called SecureBoost using tree-boosting. Using this method, multiple parties and samples of common user features were utilized for training and for privacy. The authors of [30] proposed an Iterative Federated Clustering Algorithm (IFCA) framework for clustering different user objectives using optimized parameters. The IFCA also utilized weight sharing techniques for multi-task learning. The authors [31] proposed a FedBCD algorithm to communicate with and to update multiple local parties. This multiple local update method has been used in varieties of datasets and tasks.

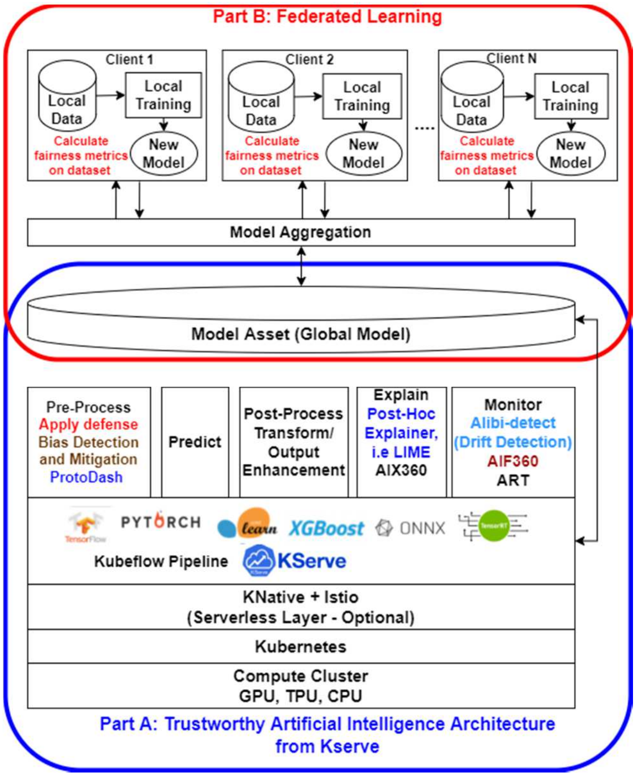

Fig. 2. The proposed federated trustworthy artificial intelligent (FTAI) architecture

The authors in [32] proposed some new metrics to evaluate the performance and fairness of federated learning. The fairness metrics were used to access the quality of federated learning and the performance metrics were used to measure the accuracy of the datasets. The trustworthy evaluation is based on the seven requirements of the European assessment of trustworthy AI. The authors of [33] proposed a theoretical trustworthy AI framework for a transferable learning model using a Bayesian model for analytical approximations. Using this work measured and analyzed the privacy leakage, interpretability and transferability. The authors of [34] proposed a toolkit for designing trusted AI systems. The authors of [35] proposed a framework for trustworthy AI implementation (TAII), and, using this work, the AI ethics process and supporting management are implemented inside and outside of the organization. Most of the federated learning approaches and trustworthy AI frameworks have been implemented using sample data, and the frameworks do not use real-time data. And also, when we consider the real-time data, the guarantee of secure communication from the best client model to the global model updates, aggregation of client results and better accuracy updations are not considered.

III. Functional Requirements of Architecture

Floridi [2], [36] summarized all applicable laws and regulations, as well as a set of requirements, should be respected by trustworthy AI (TAI); specialized evaluation lists attempt to help verify the application of each of the essential requirements: (1) Human agency and oversight; (2) Robustness and safety; (3) Privacy and data governance; (4) Transparency; (5) Diversity, non-discrimination and fairness; (6) Societal and environmental well-being; and (7) Accountability.

The Linux Foundation Artificial Intelligence (LF-AI) formulated the same thing in the Trusted AI project with 8 principles called (R)REPEATS an acronym for Reproducibility, Robustness, Equitability, Privacy, Explainability, Accountability, Transparency, and Security [37]. KServe [38], a well-known framework for machine learning model serving, generalizes TAI architecture as shown in Figure 2 part A with adopted three main components from LF-AI Trusted AI project called AI Explainability 360 (AIX360) [39], AI Fairness 360 (AIF360) [40], and Adversarial Robustness Toolbox (ART) [41] so then a TAI architecture with (R)REPEATS has been established.

IV. Federated Trustworthy AI Architecture

Our main concern is user data privacy, TAI architecture designed by KServe [38] still has the hole for data breaches. To cover that hole we added a federated learning protocol on top of KServe's TAI architecture as shown in Figure 2 part B. Federated Learning strongly believed could preserve the privacy issues since users or client keep their data on local storage and retrain the model on their own machine [6]–[8], [10], [12], [17]–[20], [23], [30], [42]–[44]. So, the risk of data leaked would be decreased dramatically. Figure 2 shows our proposed architecture which adapted KServe architecture with the featuring of a federated learning protocol. This is one of this paper's contributions. The other contribution is the model aggregation method which will be deeply explained below.

A. General Federated Learning and the Proposed Aggregation Method

The general federated learning process is depicted in Figure 3 [43]. The process step as follow: (1) Each client receives the learning model from the server; (2) Client data is used to train the received models; (3) Each client delivers to the server its trained model; (4) The server aggregates all of the collected models into a single updated model; and (5) The server updates each client's model, and processes 1 through 5 are repeated.

The proposed federated learning model aggregation process is depicted in Figure 4. The process step as follow: (1) The server asks each client for information about available resources; (2) The client sends information about the resources to the server; (3) The server chooses clients with enough resources and sends the learning model to them; (4) Client data is used to train the received models; (5) We send the loss value and fairness metrics value to the server instead of the entire new trained model; (6) The server analyzes the

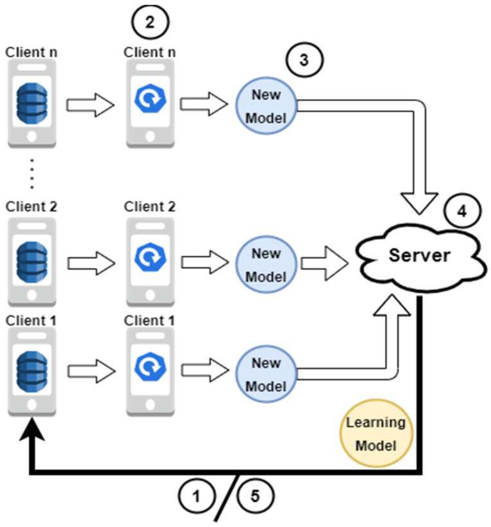

Fig. 3. Federated learning protocol

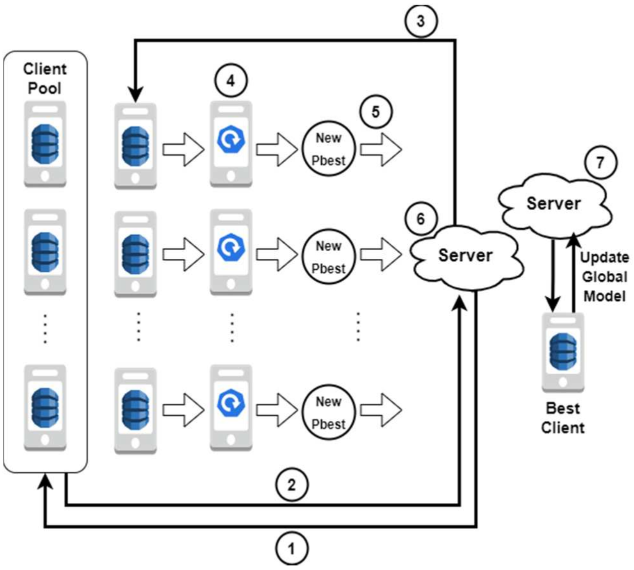

Fig. 4. The proposed federated learning protocol

data, chooses the best option, and saves the best client *gid*; and (7) The server requests to the best client using *gid* to transmit the model to the server, and the server updates the global model, repeating steps 1 through 7.

In most cases, federated learning will involve a large number of users/clients [44], [45]. Because the client's computational resources are heterogeneous. It becomes a problem if all clients participate in the aggregation process, because each client's processing time for training the local model may differ significantly. The other resources that will be checked is client's network condition. So, in order to avoid issues during the aggregation process, we must select clients that have similar and sufficient resources before we start the aggregation process and then setup the deadline for clients to do downloading, training, and uploading the result. The method given by Nishio and Yonetani [44] for the client selection process will be used in this research. However, clients must upload the entire updated local model to the server in the Nishio and Yonetani approach, which is called FedCS [44], causing a network bottleneck in the federated server. For instance, using the 549 MB VGG19 [46] baseline model, which will be utilized in federated learning, will be expensive in terms of bandwidth as client numbers increase. To address this problem, we combine FedCS with Park et al. FedPSO method [43]. FedPSO presented a practical solution to the problem of excessive bandwidth utilization during the upload of a new local model. Instead of uploading the whole model, clients send the loss value to the server as particle best (*pbest*), and the server selects the global best (*gbest*) from the *pbest* values of clients, updating *gid* with the selected client ID. The server will then ask the client with the same *gid* to upload their updated local model to the server, which will be used to update the global model.

As mentioned above the original FedPSO send only the loss value to the server to be aggregated [43]. Since the retraining process has been done on the client-side with their local data we need to make sure the updated model is free from bias and has no discrimination in its decision, so we can guarantee our system is fair [3]. In order to achieve that, we employ AIF360 [40] in the process of local model training to get the fairness metrics values [32], and then instead of sending only the loss value we also send the fairness metrics value to the server to be considered in the aggregation process. With this scenario, we can assure the updated model is fair.

B. *Achieving Trustworthy AI with the Proposed Architecture*

As Floridi [2] highlighted, at least seven important requirements should be considered in order to achieve trustworthy AI. As a result, in order to verify the proposed architecture, we will go over the important requirements one by one and show how the proposed architecture meets them.

1) *Human agency and oversight:* According to the idea of respect for human autonomy, AI systems should support human autonomy and decision-making. This necessitates AI systems acting as enablers of a democratic, prosperous, and egalitarian society by promoting user agency and fostering fundamental rights, as well as allowing for human monitoring and do necessary intervention [36]. In order to achieve these requirements, the result or output of AI should be explainable. The explainable output should also be understandable by a human. In the proposed architecture, we implement explainable AI algorithms from AIX360 [39], such as LIME [47], to make the output understandable why it produced a decision like that. We also consider that the model should be able to be monitored. That's why in the proposed architecture we implement methods such as Alibi-detect [48] to do drift detection, AIF360 [40] to make sure the trained model is fair, and Adversarial Robustness Toolkit (ART) [41] to make sure the trained model is secure and robust from adversarial attack.

2) *Robustness and safety:* Technical robustness, which is strongly tied to the idea of damage prevention, is a critical component of achieving trustworthy AI. It takes a proactive approach to risks and dependably operates as intended while minimizing unintended and unforeseen harm and preventing unacceptable harm. AI systems should be secured from vulnerabilities that could allow adversaries to exploit them, such as hacking. Data poisoning, model leaking, and attacks on the underlying infrastructure, both software and hardware, are all possibilities. When an AI system is attacked, such as in adversarial attacks, the data and behavior of the system might be altered, forcing the system to make incorrect decisions or shut down entirely [36]. To make sure the proposed architecture is robust and safe/secure, we apply defensive methods from ART [41] to detect and mitigate adversarial attacks. So, the accuracy of the model to make a decision remains accurate.

3) *Privacy and data governance:* Throughout the lifecycle of an AI system, privacy and data protection must be guaranteed. This covers both the information the user initially submitted and the information created about the user during their interaction with the system (e.g., outputs that the AI system generated for specific users or how users responded to particular recommendations). The quality of the datasets used is critical to AI system performance. When data is collected, it is possible that it will contain socially created biases, inaccuracies, errors, and mistakes. This must be resolved before any dataset is used for training. Furthermore, the data's integrity must be ensured. Data access protocols should be implemented in any organization that deals with personal information. These protocols should specify who

has access to data and when they can do so [36]. In order to achieve the aforementioned requirements, the proposed architecture adopted federated learning, which has a good reputation for preserving data privacy. Along with FL, to make sure collected data on the clients' side is unbiased, we employ AI Fairness checking methods from AIF360 [40] to calculate the fairness metrics as well as mitigate the bias.

4) Transparency: This requirement is tied to the notion of explicability and involves the transparency of elements essential to an AI system, such as data, systems, and business models. To enable traceability and increased transparency, we use datasets from open data sources in the training process so that the data is traceable and findable. Also, to make the output of the AI system transparent, the proposed architecture employs the AI explainability method from AIX360 [39].

5) Diversity, non-discrimination and fairness: This requirement primarily concerns the AI system's fairness. Discrimination should not be embedded in the AI model. Poor, biased, or imbalanced datasets throughout the training procedure could cause embedding discrimination. As a result, high-quality, diversified, and balanced datasets during AI model training should be guaranteed to meet these criteria. We apply trustworthy data governance and refer to Janssen et al.'s research on how to organize data for trustworthy AI [49]. We utilize ProtoDash [50] to properly describe datasets to analyze dataset representation because some datasets may have a complex distribution. We use appropriate algorithms from AI Fairness (AIF360) [40] to analyze fairness metrics and prevent bias in the technical implementation of the AI system lifecycle.

6) Societal and environmental well-being: Sustainable and environmentally friendly AI: the system's development, deployment, and use processes, as well as its entire supply chain, should be evaluated in this regard, for example, through a critical examination of resource usage and energy consumption during training, with less harmful options chosen. We integrate FedCS [44], FedPSO [43], and AIF360 [40] to achieve this requirement. Only clients with adequate resources and steady network connections are selected to participate in the aggregation process. We also reduce network cost by simply transmitting the loss value and fairness metric values to the server instead of the whole retrained model from clients, which could cause a network bottleneck. We are concerned that the architecture planned will have a beneficial societal impact. As a result, the first applications built on top of this architecture in the future will be pollution analysis and road traffic analysis. We collaborate with law and social science academics to examine the proposed architecture to ensure that social and democratic aspects are addressed.

7) Accountability: Accountability is a requirement that complements the above requirements and is strongly tied to the notion of fairness. It requires the implementation of processes to assure ownership and accountability for AI systems and their outcomes, both before and after their creation, deployment, and use. With the implementation of all the aforementioned methods, we make sure the proposed architecture is trustworthy. With the inclusion of fairness metrics, measurable results make stakeholders simple to comprehend and boost their trust in the system, proving that we care about accountability.

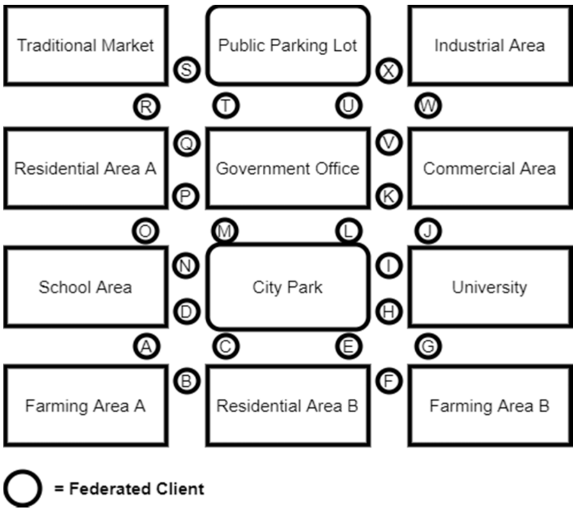

Fig. 5. Scenario of federated learning for road traffic analysis in smart city

V. Road Traffic Analysis Study Case Scenario

We illustrated the scenario of a road traffic analysis application that will be deployed on top of the proposed architecture as seen in Figure 5. Totally, 24 computational devices with one camera each will act as clients of federated learning. It is proposed that we place each client in a traffic juncture with the possibility of the type of vehicle and traffic density as listed in Table I. We assumed all clients had enough computational resources and a stable network connection, so we selected all clients to be included in the aggregation process. Then we send the global model to all clients, and each client starts the retraining process. As seen in Table I, the traffic density on each client point varies with the category, i.e., low (Client A, B, F, and G), medium (Client C, D, E, and H), high (Client I, J, N, Q, R, S, T, U, V, W, and X), and very high (K, L, M, O, and P). As for the detected vehicles, in some clients' points, they mostly detected particular kinds of vehicles (Clients A, G, R, S, and X). And in some clients' points, due to restrictions, medium trucks (medium trucks can be passed in Client R, S, T, U, V, W, and X) and big trucks (big trucks can be passed in Client W and X) are not allowed to be passed.

Based on the above mentioned conditions, clients that only detect and collect data for a particular vehicle, such as clients A, G, R, S, and X, for example, a client A that mostly detects small trucks when the retaining process is complete, will have high accuracy in detecting small trucks while being worse at detecting other kinds of vehicles. Also, if we check the datasets' fairness metrics, the disparity value would be highly negative, which means that the dataset is imbalanced.

The level of traffic density may also influence the best model selection. More data can be used for local training as the density level rises. The best model would theoretically be chosen from the client with the highest traffic density, but because the proposed architecture also considers fairness metrics, the best model is not always chosen from the client with the highest traffic density but also from the client with the highest fairness metrics value. Clients K, L, M, O, and P have the highest density over other clients in Table I, but if we look closely, we can see that those clients have not detected any big trucks at all, compared to client W, which has only high density but detects and saves data from all kinds of vehicles, so our proposed architecture will ideally choose the best model from client W. Of course, the best client is not always chosen from client W. It will always be updated based on the real conditions on the road. For example, during certain events, the road conditions may change dynamically at certain points. For example, road repair work, big holidays that cause density and disparity of vehicles moving to the other clients' points, are some factors that would affect the model updating conditions.

TABLE I. POSSIBLE ROAD TRAFFIC DISTRIBUTION BASED ON FIGURE 5

Fed Client	Possible Detected Vehicle	Traffic Density	Balance Data Possibility
A	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Low	Mostly small truck
B	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Low	All kind of vehicle
C	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Medium	All kind of vehicle
D	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Medium	All kind of vehicle
E	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Medium	All kind of vehicle
F	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Low	All kind of vehicle
G	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Low	Mostly small truck and motor cycle
H	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Medium	All kind of vehicle
I	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
J	Bicycle, Motor Cycle, Small Truck, Medium Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
K	Bicycle, Motor Cycle, Small Truck, Medium Truck, Sedan, SUV, MPV, Small Car, Public Bus	Very High	All kind of vehicle
L	Bicycle, Motor Cycle, Small Truck, Medium Truck, Sedan, SUV, MPV, Small Car, Public Bus	Very High	All kind of vehicle
M	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Very High	All kind of vehicle
N	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
O	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Very High	All kind of vehicle
P	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	Very High	All kind of vehicle
Q	Bicycle, Motor Cycle, Small Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
R	Bicycle, Motor Cycle, Small Truck, Medium Truck, Big Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	Mostly motor cycle and small truck
S	Bicycle, Motor Cycle, Small Truck, Medium Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	Mostly motor cycle and small truck
T	Bicycle, Motor Cycle, Small Truck, Medium Truck, Big Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
U	Bicycle, Motor Cycle, Small Truck, Medium Truck, Big Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
V	Bicycle, Motor Cycle, Small Truck, Medium Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
W	Bicycle, Motor Cycle, Small Truck, Medium Truck, Big Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	All kind of vehicle
X	Bicycle, Motor Cycle, Small Truck, Medium Truck, Big Truck, Sedan, SUV, MPV, Small Car, Public Bus	High	Mostly big truck and medium truck

VI. CONCLUSIONS

This paper introduced the Federated Trustworthy Artificial Intelligence (FTAI) Architecture with the coupling of the TAI architecture between KServe and Federated Learning. We suggested a new model aggregation strategy that combines FedCS and FedPSO, as well as using AIF360 in the client-side training process to ensure that the model is embedded discrimination-free so that fairness is achieved. The contributions of this paper are as follows: (1) a trustworthy and secure user data privacy platform architecture; (2) a heterogeneous and low-bandwidth tolerant client; and (3) an updated model with desirable fairness. We can prove that the proposed architecture can satisfy the technical requirements of a trustworthy AI system. The scenario also clearly describes how the best model will be selected based on the current road conditions. This architecture is still in the planning stages. Proof of work is still required. As a result, we will look at the experimental results in the future and compare them to other architecture and aggregation methods. Then, based on the results, we can improve the FTAI architecture's performance.

REFERENCES

[1] M. Brundage *et al.*, “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” *ArXiv200407213 Cs*, Apr. 2020, Accessed: May 10, 2022. [Online]. Available: http://arxiv.org/abs/2004.07213

[2] L. Floridi, “Establishing the rules for building trustworthy AI,” *Nat. Mach. Intell.*, vol. 1, no. 6, pp. 261–262, Jun. 2019, doi: 10.1038/s42256-019-0055-y.

[3] H. Liu *et al.*, “Trustworthy AI: A Computational Perspective,” *ArXiv210706641 Cs*, Aug. 2021, Accessed: May 10, 2022. [Online]. Available: http://arxiv.org/abs/2107.06641

[4] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M.-A. Sicilia, “Traceability for Trustworthy AI: A Review of Models and Tools,” *Big Data Cogn. Comput.*, vol. 5, no. 2, p. 20, May 2021, doi: 10.3390/bdcc5020020.

[5] S. Thiebes, S. Lins, and A. Sunyaev, “Trustworthy artificial intelligence,” *Electron. Mark.*, vol. 31, no. 2, pp. 447–464, Jun. 2021, doi: 10.1007/s12525-020-00441-4.

[6] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, “Federated Learning,” *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 13, no. 3, pp. 1–207, Dec. 2019, doi: 10.2200/S00960ED2V01Y201910AIM043.

[7] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, “A survey on federated learning,” *Knowl.-Based Syst.*, vol. 216, p. 106775, Mar. 2021, doi: 10.1016/j.knosys.2021.106775.

[8] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, “A review of applications in federated learning,” *Comput. Ind. Eng.*, vol. 149, p. 106854, Nov. 2020, doi: 10.1016/j.cie.2020.106854.

[9] K. Bonawitz *et al.*, “Towards Federated Learning at Scale: System Design,” in *Proceedings of Machine Learning and Systems*, 2019, vol. 1, pp. 374–388. [Online]. Available: https://proceedings.mlsys.org/paper/2019/file/bd686fd640be98efaae0091fa301e613-Paper.pdf

[10] 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, May 2020, doi: 10.1109/MSP.2020.2975749.

[11] Y. Jararweh, S. Otoum, and I. A. Ridhawi, “Trustworthy and sustainable smart city services at the edge,” *Sustain. Cities Soc.*, vol. 62, p. 102394, Nov. 2020, doi: 10.1016/j.scs.2020.102394.

[12] S. P. Ramu *et al.*, "Federated learning enabled digital twins for smart cities: Concepts, recent advances, and future directions," *Sustain. Cities Soc.*, vol. 79, p. 103663, Apr. 2022, doi: 10.1016/j.scs.2021.103663.

[13] N. Zohrabi *et al.*, "OpenCity: An Open Architecture Testbed for Smart Cities," in *2021 IEEE International Smart Cities Conference (ISC2)*, Sep. 2021, pp. 1–7. doi: 10.1109/ISC253183.2021.9562813.

[14] A. Albaseer, B. S. Ciftler, M. Abdallah, and A. Al-Fuqaha, "Exploiting Unlabeled Data in Smart Cities using Federated Learning," *ArXiv200104030 Cs*, Mar. 2020, Accessed: May 10, 2022. [Online]. Available: http://arxiv.org/abs/2001.04030

[15] A. Imteaj and M. H. Amini, "Distributed Sensing Using Smart End-User Devices: Pathway to Federated Learning for Autonomous IoT," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, Dec. 2019, pp. 1156–1161. doi: 10.1109/CSCI49370.2019.00218.

[16] M. V. S. da Silva, L. F. Bittencourt, and A. R. Rivera, "Towards Federated Learning in Edge Computing for Real-Time Traffic Estimation in Smart Cities," in *Anais do Workshop de Computação Urbana (CoUrb 2020)*, Brasil, Dec. 2020, pp. 166–177. doi: 10.5753/courb.2020.12361.

[17] S. Lonare and R. Bhramaramba, "Model Aggregation Federated Learning Approach for Vehicular Traffic Forecasting," *J. Eng. Sci. Technol. Rev.*, vol. 14, no. 3, pp. 111–115, 2021, doi: 10.25103/jestr.143.13.

[18] Y. Qi, M. S. Hossain, J. Nie, and X. Li, "Privacy-preserving blockchain-based federated learning for traffic flow prediction," *Future Gener. Comput. Syst.*, vol. 117, pp. 328–337, Apr. 2021, doi: 10.1016/j.future.2020.12.003.

[19] Y. Liu, J. J. Q. Yu, J. Kang, D. Niyato, and S. Zhang, "Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7751–7763, Aug. 2020, doi: 10.1109/JIOT.2020.2991401.

[20] P. Chhikara, R. Tekchandani, N. Kumar, S. Tanwar, and J. J. P. C. Rodrigues, "Federated Learning for Air Quality Index Prediction using UAV Swarm Networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6. doi: 10.1109/GLOBECOM46510.2021.9685991.

[21] D.-V. Nguyen and K. Zettsu, "Spatially-distributed Federated Learning of Convolutional Recurrent Neural Networks for Air Pollution Prediction," in *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, Dec. 2021, pp. 3601–3608. doi: 10.1109/BigData52589.2021.9671336.

[22] S. Park, S. Jung, H. Lee, J. Kim, and J.-H. Kim, "Large-Scale Water Quality Prediction Using Federated Sensing and Learning: A Case Study with Real-World Sensing Big-Data," *Sensors*, vol. 21, no. 4, p. 1462, Feb. 2021, doi: 10.3390/s21041462.

[23] I. Kholod *et al.*, "Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis," *Sensors*, vol. 21, no. 1, p. 167, Dec. 2020, doi: 10.3390/s21010167.

[24] "TensorFlow Federated: Machine Learning on Decentralized Data," *TensorFlow*. https://www.tensorflow.org/federated (accessed May 10, 2022).

[25] FedAI, "An Industrial Grade Federated Learning Framework," *Fate*. https://fate.fedai.org/ (accessed May 10, 2022).

[26] B. Research, "Baidu PaddlePaddle Releases 21 New Capabilities to Accelerate Industry-Grade Model Development." http://research.baidu.com/Blog/index-view?id=126 (accessed May 10, 2022).

[27] "Syft + Grid: Code for computing on data you do not own and cannot see." OpenMined, May 10, 2022. Accessed: May 10, 2022. [Online]. Available: https://github.com/OpenMined/PySyft

[28] "Sherpa.ai | Privacy-Preserving Artificial Intelligence." https://www.sherpa.ai/ (accessed May 10, 2022).

[29] K. Cheng *et al.*, "SecureBoost: A Lossless Federated Learning Framework," *IEEE Intell. Syst.*, vol. 36, no. 6, pp. 87–98, Nov. 2021, doi: 10.1109/MIS.2021.3082561.

[30] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An Efficient Framework for Clustered Federated Learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 19586–19597. Accessed: May 10, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/e32cc80bf07915058ce90722ee17bb71-Abstract.html

[31] Y. Liu *et al.*, "A Communication Efficient Collaborative Learning Framework for Distributed Features," *ArXiv191211187 Cs Stat*, Jul. 2020, Accessed: May 10, 2022. [Online]. Available: http://arxiv.org/abs/1912.11187

[32] S. Divi, Y.-S. Lin, H. Farrukh, and Z. B. Celik, "New Metrics to Evaluate the Performance and Fairness of Personalized Federated Learning," *ArXiv210713173 Cs*, Jul. 2021, Accessed: May 10, 2022. [Online]. Available: http://arxiv.org/abs/2107.13173

[33] M. Kumar, B. A. Moser, L. Fischer, and B. Freudenthaler, "Information Theoretic Evaluation of Privacy-Leakage, Interpretability, and Transferability for Trustworthy AI." arXiv, Apr. 12, 2022. Accessed: May 16, 2022. [Online]. Available: http://arxiv.org/abs/2106.06046

[34] S. Schmager and S. Sousa, "A Toolkit to Enable the Design of Trustworthy AI," in *HCI International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence*, vol. 13095, C. Stephanidis, M. Kurosu, J. Y. C. Chen, G. Fragomeni, N. Streitz, S. Konomi, H. Degen, and S. Ntoa, Eds. Cham: Springer International Publishing, 2021, pp. 536–555. doi: 10.1007/978-3-030-90963-5_41.

[35] J. Baker-Brunnbauer, "TAII Framework for Trustworthy AI Systems," *ROBONOMICS J. Autom. Econ.*, vol. 2, pp. 17–17, Dec. 2021.

[36] S. Weiser, "Requirements of Trustworthy AI," *FUTURIUM - European Commission*, Apr. 08, 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1 (accessed May 16, 2022).

[37] "Trusted AI – LFAI & Data." https://lfaidata.foundation/projects/trusted-ai/ (accessed May 11, 2022).

[38] "KServe." KSERVE, May 10, 2022. Accessed: May 11, 2022. [Online]. Available: https://github.com/kserve/kserve

[39] "AI Explainability 360 (v0.2.1)." Trusted-AI, May 11, 2022. Accessed: May 11, 2022. [Online]. Available: https://github.com/Trusted-AI/AIX360

[40] "AI Fairness 360 (AIF360)." Trusted-AI, May 11, 2022. Accessed: May 11, 2022. [Online]. Available: https://github.com/Trusted-AI/AIF360

[41] "Adversarial Robustness Toolbox (ART) v1.10." Trusted-AI, May 11, 2022. Accessed: May 11, 2022. [Online]. Available: https://github.com/Trusted-AI/adversarial-robustness-toolbox

[42] Z. Zheng, Y. Zhou, Y. Sun, Z. Wang, B. Liu, and K. Li, "Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges," *Connect. Sci.*, vol. 34, no. 1, pp. 1–28, Dec. 2022, doi: 10.1080/09540091.2021.1936455.

[43] S. Park, Y. Suh, and J. Lee, "FedPSO: Federated Learning Using Particle Swarm Optimization to Reduce Communication Costs," *Sensors*, vol. 21, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/s21020600.

[44] T. Nishio and R. Yonetani, "Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–7. doi: 10.1109/ICC.2019.8761315.

[45] J. So, B. Güler, and A. S. Avestimehr, "Turbo-Aggregate: Breaking the Quadratic Aggregation Barrier in Secure Federated Learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 479–489, Mar. 2021, doi: 10.1109/JSAIT.2021.3054610.

[46] K. Team, "Keras documentation: Keras Applications." https://keras.io/api/applications/ (accessed Aug. 04, 2022).

[47] M. T. C. Ribeiro, "lime." May 16, 2022. Accessed: May 16, 2022. [Online]. Available: https://github.com/marcotcr/lime

[48] "Alibi Detect — alibi-detect 0.9.1 documentation." https://docs.seldon.io/projects/alibi-detect/en/stable/index.html (accessed May 16, 2022).

[49] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy Artificial Intelligence," *Gov. Inf. Q.*, vol. 37, no. 3, p. 101493, Jul. 2020, doi: 10.1016/j.giq.2020.101493.

[50] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient Data Representation by Selecting Prototypes with Importance Weights," arXiv, arXiv:1707.01212, Aug. 2019. doi: 10.48550/arXiv.1707.01212.