TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 14/15 pp759–771
DOI: 10.26599/TST.2021.9010026
Volume 26, Number 5, October 2021

Collaborative City Digital Twin for the COVID-19 Pandemic: A Federated Learning Solution

Junjie Pang, Yan Huang, Zhenzhen Xie, Jianbo Li*, and Zhipeng Cai

Abstract: The novel coronavirus, COVID-19, has caused a crisis that affects all segments of the population. As the knowledge and understanding of COVID-19 evolve, an appropriate response plan for this pandemic is considered one of the most effective methods for controlling the spread of the virus. Recent studies indicate that a city Digital Twin (DT) is beneficial for tackling this health crisis, because it can construct a virtual replica to simulate factors, such as climate conditions, response policies, and people's trajectories, to help plan efficient and inclusive decisions. However, a city DTsystem relies on long-term and high-quality data collection to make appropriate decisions, limiting its advantages when facing urgent crises, such as the COVID-19 pandemic. Federated Learning (FL), in which all clients can learn a shared model while retaining all training data locally, emerges as a promising solution for accumulating the insights from multiple data sources efficiently. Furthermore, the enhanced privacy protection settings removing the privacy barriers lie in this collaboration. In this work, we propose a framework that fused city DT with FL to achieve a novel collaborative paradigm that allows multiple city DTs to share the local strategy and status quickly. In particular, an FL central server manages the local updates of multiple collaborators (city DTs), providing a global model that is trained in multiple iterations at different city DT systems until the model gains the correlations between various response plans and infection trends. This approach means a collaborative city DT paradigm fused with FL techniques can obtain knowledge and patterns from multiple DTs and eventually establish a "global view" of city crisis management. Meanwhile, it also helps improve each city's DT by consolidating other DT's data without violating privacy rules. In this paper, we use the COVID-19 pandemic as the use case of the proposed framework. The experimental results on a real dataset with various response plans validate our proposed solution and demonstrate its superior performance.

Key words: COVID-19; Digital Twin (DT); Federated Learning (FL); deep learning

- Junjie Pang is with the College of Computer Science and Technology, Qingdao University, Qingdao 266000, China, and is also with the Business School, Qingdao University, Qingdao 266000, China. E-mail: pangjj18@163.com.
- Yan Huang is with the College of Computing and Software Engineering, Kennesaw State University, Atlanta, GA 30060, USA. E-mail:yhuang24@kennesaw.edu.
- Zhenzhen Xie is with the College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: xiezz14@mails.jlu.edu.cn.
- Jianbo Li is with the College of Computer Science and Technology, Qingdao University, Qingdao 266000, China, E-mail: lijianbo@188.com.
- Zhipeng Cai is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA. E-mail: zcai@gsu.edu.

∗ To whom correspondence should be addressed.

Manuscript received: 2021-02-26; accepted: 2021-03-18

1 Introduction

Coronavirus (COVID-19), an infectious disease caused by the recently discovered coronavirus, was identified on December 31th 2019[1] (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). The virus has spread worldwide in less than three months, infected more than 116 million people, and caused over 2 575 196 deaths (https://www.worldometers.info/coronavirus/). This widespread coronavirus outbreak received tremendous attention from the research and medical perspective. However, a specific antiviral treatment of COVID-19 remains unavailable. Therefore, an early and radical government response can be considered the most effective method when facing a novel infectious disease. However, determining the response plan properly can be challenging because of a lack of experience and efficient data sources.

A mathematical model is a possible solution for the intervention and surveillance of the infectious disease[2]. For example, the Susceptible-Infected-Susceptible (SIS) epidemic model is widely used in describing the spreading process for a virus in a static network with an assumption of a constant population. This model can also combine with a time-varying dynamic network to describe more complex propagation. We also observe that the significant proliferation of machine learning techniques has resulted in the rapid development of intelligent forecasting models[3]. Recent works demonstrate their comparable performance in capturing non-trivial atypical trends and typical patterns for epidemic control, such as the Wiener-series-based machine learning model for measuring the H1N1 virus spread after an intervention[4], and the representation learning model that generates interpretable epidemic forecasting results for seasonal influenza forecasting[5].

However, these models still have several challenges and limitations in predicting infection trends of a novel infectious disease, such as COVID-19:

Uncertain influence: In contrast to other pandemic predictions, the prediction model of unknown infectious diseases, such as COVID-19, must learn the influence of various response plan settings, such as mask-wearing, shelter in place, and statewide school closures.

Cold start problem: When a new virus starts to spread, the local health department always needs a long time to properly collect sufficient data to generate a response to the pandemic. Note that the same response plan could have varied effects in different locations: a radical response plan may only bring economic risks to a low-risk areas, while the same actions could result in losing control of the spreading virus and economic damage for severely affected areas.

Privacy protection: The data resources related to a health crisis, such as COVID-19 pandemic, unavoidably contain sensitive information. This situation means that we cannot collaboratively share these data unless we can provide a strong privacy guarantee[6]. However, medical institutions and local governments may expect a high-performance model for epidemic control, which means massive data collection is required for deep learning-based models. Because of privacy and confidentiality concerns, these applications can possibly be prevented, such that data silos emerge[7]. These silos are isolated islands of data, which can make health data management disorganized and inefficient. Moreover, they make it prohibitively costly for the local agencies to extract knowledge, share insights, and realize collaborations with other regions[8].

As shown in Fig. 1, we proposes a Digital Twin (DT) enabled collaborative training framework based on a federated learning paradigm to resolve the above problems. We use a city DT to build a virtual replica of the city/state that provides a digital view of

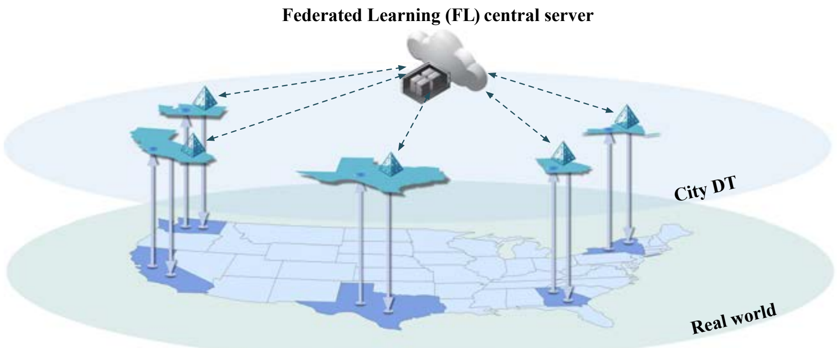


Fig. 1 Overview of the collaborative framework for a multiple city DT.

city/state facilities, human activities, and other types of information to enable information convergence in multiple aspects of infection trend, thus enabling the prediction of the uncertain influence caused by different events. City DT allows each region to accumulate historical data efficiently, while demonstrating a remarkable potential for offering continuous interaction with the physical world to refine prediction[9, 10]. Specifically, Time Convolutional Networks (TCN) is adopted to implement a city DT, ensuring superior performance for modeling the temporal information dynamics and the future infection trend prediction under a local response plan.

To further resolve the cold start problem and privacy concerns, FL[11] is introduced as the collaborative training paradigm. It only involves the parameters shared among multiple parties in training collaborative machine learning models. Thus, FL can significantly lower the privacy risks in collaborative knowledge exchange[12]. These features, combined with the high-quality contribution from local city DT, are essential for establishing a prediction model and accumulating knowledge and insights for an unknown virus, such as COVID-19, in a short period.

Our contributions can be summarized as follows:

- To resolve the uncertain influence challenge for COVID-19 pandemic management, we are among the first to propose a novel collaborative learning framework with city DT embedding.
- The proposed TCN-based city DT helps determine the effects of various local response plans for each city/area, which is the first attempt to utilize a non-trivial deep learning model for epidemic forecasting considering fine-granularity time pattern features.
- Considering the cold start problem and privacy concerns, we use the FL as the solution, which offers collaborative learning via only parameter-sharing not to disturb each city DT's privacy rules.
- Extensive simulations with a real dataset reveal that our proposed framework significantly outperforms the non-trivial baseline and the non-FL city DT solution with a strong privacy guarantee.

The remainder of the paper is organized as follows. Section 2 introduces related works. The basic definitions and problem statements are presented in Section 3. Section 4 explains the detailed structure and methodology of the proposed framework. The experiments and results are analyzed in Section 5. Finally, conclusions and future work are presented in Section 6.

2 Related Work

In this section, we start with a brief review of traditional methods for epidemic prediction, and then discuss the related techniques and the need for the collaborative training framework.

Deep learning-based epidemic control: Historical insights from temporal infection data have been crucial for epidemic control and prevention, and could benefit other problems in smart city systems[13, 14] or enhanced social network analysis[15]. Deep learning-based techniques have demonstrated a remarkable performance to model such temporal correlations and recognize multiple patterns[16, 17], including the deep neural network-based short-term and high-resolution epidemic forecasting for influenza-like illness[18], the semi-supervised deep learning framework that integrates computational epidemiology and social media mining techniques for epidemic simulation, called SimNest[19] and EpiRP[20], which use representational learning methods to capture the dynamic characteristics of epidemic spreading on social networks for epidemics-oriented clustering and classification.

Moreover, recent breakthroughs in infectious disease modeling, forecasting, and real-time disease surveillance have further convinced us that these activities mitigate the effects of disease outbreaks. In addition, with the rapid growth of cloud computing and wireless data communication architectures[21, 22], deep learning-models demonstrate constantly improving efficiency. Given various application scenarios and objectives, deep learning-based models can be different. A typical solution for localized flu "nowcasting" and flu activity inferring is ARGONet[23], which is a network-based approach leveraging spatio-temporal correlations across different states to improve the prediction accuracy. ARGONet uses a spatial network to capture the spatio-temporal correlations across different states and produces more precise retrospective estimates based on the information from influenza-related Google search frequencies, electronic health records, and historical influenza trends. Instead of leveraging multiple data source, such as ARGONet, the studies in Ref. [24] proposed a multi-task learning-based model that is only uses user-generated content (Web search data). They investigate linear and nonlinear model capabilities and find that disease rate estimates can be significantly improved in the case study of an influenza-like illness.

However, these successful attempts are based on large-scale data sources or massive historical information of the disease with similar spreading patterns, which means that high-dimensionality, irregularity forms, noise, privacy concerns, or sparsity problems may affect these learning-based models' performance[25, 26], especially when we face unexpected infectious disease outbreaks, such as the COVID-19 pandemic.

For filling the data gap, the city DT is proposed as a promising solution. It is a virtual representation of a device or a specific application scenario that can interact with the target environment to collect data continuously for real-time decision-making. Several successful research attempts include a disaster city DT[27, 28], energy management[29], and city-scale Light Detection and Ranging (LiDAR) point clouds[30]. Furthermore, Singapore[31] and Germany[32] have launched the city-scale DT to monitor and improve utilities, which enhance the transparency, sustainability, and availability of a DT.

In this way, the city DT offers us a high-quality and real-time data resource to describe the spread of an epidemic, whereas data silos naturally emerge because of privacy barriers[33, 34]. To maintain the advantages of DT and tolerate the data sparsity challenge, FL, which allows multiple stack-holders to share data and train a global model, has become a preferred scheme[11]. In typical FL scheme settings, each data owner (FL client) engages in a collaborative training process without transferring the raw data to the others. Through FL, the central server manages each client's local training updates and aggregates their contributions to enhance the global model's performance. Several concrete scenarios, including Google's Gboard[35], health AI[36], and smart banking[37], show the advantages of FL in handling collaborative training issues and data difficulties among diverse data owners. Therefore, we are motivated to utilize FL techniques to resolve the data sparsity challenges and design a collaborative city DT for COVID-19 pandemic control.

3 Preliminary and System Model

In this section, we first explain the preliminaries of the proposed framework. The structural design, which combines DT and FL for COVID-19 pandemic control, will be explained with a mathematical definition of the problem objective. The detailed methodology and proposed solution will be illustrated in Section 4.

3.1 Preliminaries

TCN: Given these advantages and a delicate-designed convolutional architecture, TCN can handle variable length inputs, such as those of Recurrent Neural Network (RNN)-based methods[38], and convincingly outperform baseline recurrent architectures across various sequence modeling tasks. By leveraging a much simpler, 1-D fully-convolutional network, TCN can build a very long sufficient history size for a variable length of a input sequence, avoiding large memory requirements and intricate network architecture, such as those of gated RNNs. Its model pipeline has two distinguishing features: causal convolution and dilated convolution. The causal convolutions consider that the output at time t is convoluted only with elements that occurred before t, which suggests that current spatial-temporal information depends only on the past and not on any future inputs. Then, to further achieve longer history data without introducing an extremely deep network or very large filters, a TCN uses a dilated convolution to enlarge the sequence data's maximum length (receptive field). Notably, the receptive field can be changed by stacking more dilated convolution layers or increasing the filter sizes, which fully explain the robustness and flexibility.

FL: FL is a privacy-enhanced distributed learning framework with an emphasis on using mobile and edge devices for collecting data and scaling the computation resources[11]. Unlike previous research handling with training data in a centralized manner, FL's essential property uses a "parameter-only" collaborative training to avoid disturbing each FL clients' privacy rules. Thus, various participating clients can solve the learning task through a hub-and-spoke topology for model aggregation while maintaining the raw data on their devices. In particular, for a new FL training task, (1) the FL central server trains a global model for initialization, then distributes this model to the existing collaborators (clients); (2) after receiving the global model, each collaborator uses the local dataset to update the local parameters and generates the local updates; (3) based on specified synchronization settings, all these updates are sent to the FL central server for aggregation, and the global model is improved; (4) these distributed update iterations are repeated until the global model converges or achieves the expected performance.

DT: A DT is a digital representation of a physical asset, environment, or system, that was initially developed to automatically aggregate, analyze, and visualize complex information through continuous interactions with the physical world.

3.2 City DT for COVID-19 pandemic control

From the above facts, we observe explicit advantages of using FL to establish the collaborative training framework of multiple city DTs. First, by separating local model training and global model updates, FL offers a strong capability to deal with the isolated data island problem between multiple DTs. Secondly, with enhanced privacy settings, each city DT can obtain the collaboration achievements without violating its privacy rules. These properties are essential for COVID-19 pandemic control, because different regions need a collaboration paradigm with lower privacy risks to quickly realize an effective response plan. Furthermore, for each city DT using a TCN as the time-series data modeling method, the shared global model can provide more temporal correlation perspectives, which is a complementary approach to make the city DT quickly converge to a robust performance.

In our proposed work, a city DT has three primary components: the physical environment of the city, a virtual replica describing the city's architecture, functions, and behaviors, and active communications between the two to obtain real-time spatiotemporal data from various infrastructure and human systems[39]. According to the three components, we compose a city DT for COVID-19 pandemic control using the following metrics:

COVID-19 case number: The COVID-19 case number is the number of identified confirmed cases. It is the direct evidence to describe the characteristics of human-to-human transmission. Daily updates of case numbers represent infection trend changes and show whether a response plan is operated efficiently. In our framework, each DT model is from a specific area, so that the case number is bounded with the area and time information.

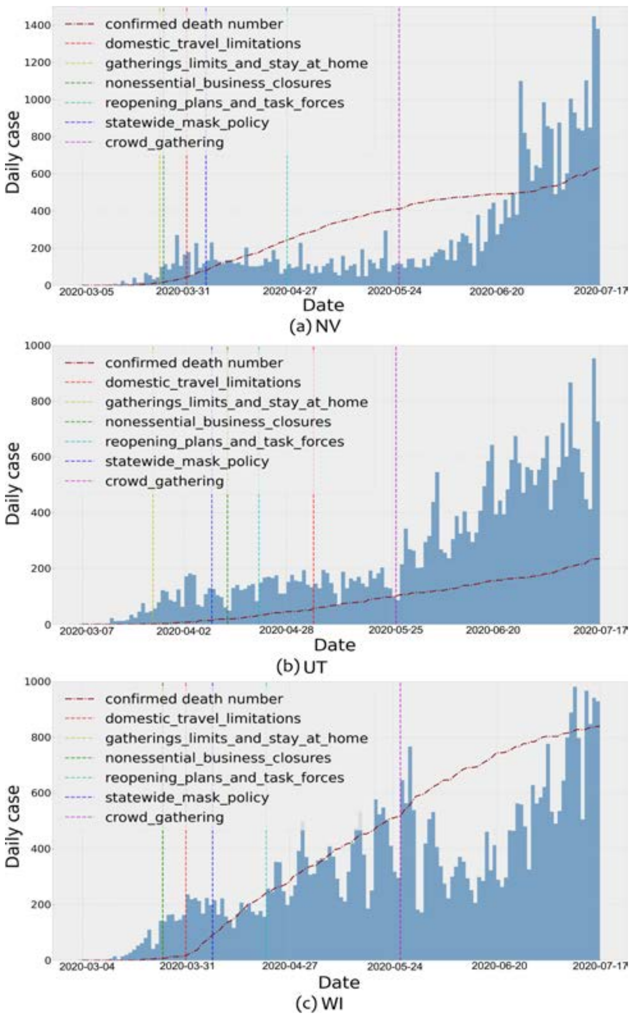
COVID-19 testing number: This metric measures how many individuals get tested of COVID-19 in the affected regions. The actual total number of people infected with COVID-19 cannot be obtained. In this situation, the number of confirmed cases depends on the testing number, because it can be used to further interpret and revise the COVID-19 case number. Meanwhile, the positive rate, computed as the testing number in a particular time window, is an essential metric for describing if the target area controls the spread properly. Therefore, we must use both numbers to estimate the current infection status and mitigate the risks of under-reporting cases and deaths.

COVID-19 confirmed death number: The confirmed death number describes the ability of COVID-19 to cause death, which is another direct piece of evidence of how a region is affected. Furthermore, it is an important metric for identifying at-risk populations and guiding the response plan to adjust the medical resource allocations. The confirmed death number and case number can have very different trends because the same response plan may affect these metrics differently. For example, several infected regions can bring the number of deaths down for the same response plan, but other areas may only lower the case number. Thus, the death rate helps us understand the severity of this virus and evaluate each response plan's fine-grained function.

Response plan: For COVID-19 pandemic control, various organizations and governments develop several local-level response plans or even a country-level response plan to prepare for and respond to COVID-19. In our DT model, we use $R_i = (l_i, t_{st}, t_{end})$ to represent a response plan, where l_i is the location, with t_{st} and t_{end} denoting the starting time and end time of R_i. We include the following response plans in the proposed model: 14-day quarantine, domestic travel limitations, gathering limits and stay-at-home orders, nonessential business closures, reopening plans, mask policy, etc. The effectiveness of different response plans can vary because they may be affected by several external factors, such as a sudden emergency, adverse weather conditions, or vaccinations.

Temporal effects: In our work, two types of temporal effects are considered as the primary factors in each city DT model: temporal effects of historical infection status (e.g., historical case numbers and historical deaths) and external factors (e.g., selected response plans, events, and gatherings). Note that our proposed city DT model's primary goal is to determine whether the specific response plan can flatten the infection curve and evaluate the period of validity of the plan. We thus need a robust epidemic forecasting model that can consider multiple temporal factors and hidden periodicity.

Historical infection status: For a fast-evolving pandemic, such as the COVID-19 pandemic, the historical case numbers are direct evidence of the correlation between past conditions and the current infection status. In Fig. 2, we take the historical daily case information of three states (NV: Nevada, UT: Utah,



(a) NV

(b) UT

(c) WI

Fig. 2 Correlation between the current infection trend and the historical infection numbers.

and WI: Wisconsin) as examples of these temporal effects. From the early March data of all three states, we observe the same immediate effect of historical infection numbers, because they lead to a continuously increasing number of infections until April 2nd, which indicates that the temporal correlations can play an essential role in explaining and predicting future infection trends.

External factors: To determine whether external factors can have an immediate or delayed effect on future infection trends, we observe the correlation between each specific factor and the infection status in the next few days. In our work, the response plans are considered the primary external factor, because the choice of a specified response plan can also significantly affect the number of infections. This effect can be various, depending on the strictness of that policy, people's acceptance of it, and many other factors, such as various climate conditions or the population density. For example, in Fig. 2, we observe that after taking a specified response plan, such as domestic travel limitations or gathering limits, the infection trend of all three states can be significantly decreased. However, for different reasons, the validity period of the response plan can vary, so all three states exhibit an increasing infection trend over time. Thus, the temporal effect of a specific response plan can be complicated because external factors, such as the 14-day time window, the indeterminate period that a response plan starts to take effect, and a paroxysmal public crisis, may also lead to infection trend changes, which suggests that it is a challenge to estimate the temporal effects of a specific response plan from such a complicated physical environment.

3.3 Problem statement

To place the COVID-19 pandemic under control, different local agencies in each city/region may choose their own strategy to meet the local requirements. This divergence occurs mainly because different regions should consider the local intrinsic properties. For instance, Area A, which is a thinly populated district with very low infection rates, would prefer to choose a less radical response plan; while the another Area B, where has severe infection conditions, is very likely to choose a less radical response plan, like restricting activities and closing most of the facilities. This situation means that each region can only obtain knowledge by trial-and-error operation schemes for seeking an effective response plan, and the increasing time cost could lead to a delayed response plan with poor performance. Moreover, to train a city DT model to predict future infection trends after a response plan, enough features must be used to construct the temporal correlations, which suggests that a collaborative city DT-training framework must be considered instead.

In coping with these challenges and limitations, the FL protocol is used in our collaborative city DT framework. In this paper, we study the problem of forecasting future infection trends for specific response plans. Formally, this problem is stated as follows: Given multiple city DTs, $D = D_1, D_2, \ldots, D_i$, each expects collaborations and is bounded with a local data sensing method to generate individualize data source $s_{i,1}, s_{i,2}, \ldots, s_{i,m_i}$, our federated training problem is to optimize the following function:

$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) \triangleq \sum_{i=1}^{N} p_i F_i(\mathbf{w}) \right\},$$

where N is the number of city DTs, $\mathbf{w}$ represents the parameter of FL global model, $p_i = m_i/m$, where m is the number of data points in all city DT's data source and m_i is the number of data points of i-th city DT. For city DT D_i, $f(\cdot)$ is the loss function of a data point, so that the local objective $F_i(\cdot)$ of D_i can be defined as

$$F_i(\mathbf{w}) \triangleq \frac{1}{m_i}\sum_{j=1}^{m_i} f\left(\mathbf{w}; s_{i,j}\right).$$

4 Exploring the Collaborative Framework for Multiple City DT

A city DT can actively collect real-time data from various human systems in a targeted region and provide automatic decision-making or possible future behavior predictions, which is beneficial for tracking an infectious disease's progress in real-time and accumulating local information knowledge. However, a city DT likely lacks the experience to quickly determine a response plan when facing an unknown virus, such as COVID-19, that has data sparsity challenges. Thus, we propose a novel collaborative training process, enabling multiple city DTs to train a global model to help each city DT improve the response policy efficiently. Specifically, TCN architecture is used in the DT model for temporal sequence modeling, and an FL-based collaborative training process is implemented, so that every city DT can be automatically improved with privacy protection by design.

4.1 System description

In our proposed collaborative training framework for multiple city DTs shown in Fig. 3, each city DT model can update and evolve itself in two ways:

Local updates: This process is similar to the updates of the typical DT model, which actively collects real-time COVID-19 conditions (historical case numbers, testing number, confirmed death number, and external factors, such as crowd gatherings, population age, and vaccinations) through on-device sensing or directly uses data sources from institutions, such as public health agencies or hospitals.

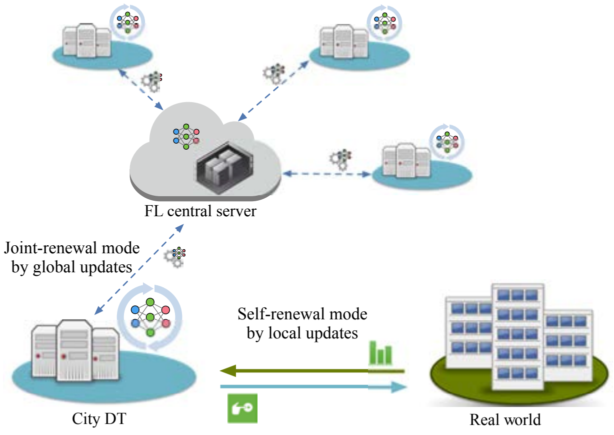


Fig. 3 Illustration of the collaborative training process for multiple city DTs.

Global updates: Our proposed FL framework offers a platform for all city DTs with a willingness to share their parameter-only knowledge. Thus, each city DT can learn from others to update its local model during "global updates". Specifically, this mode is implemented by the FL aggregation, at which all the city DTs would upload the local parameters to generate an aggregated update to ensure that each city DT can always benefit from the last updated global model.

For example, as shown in Fig. 4, it supposes a city DT A decides to take a new response operation R_i (similar to a statewide mask-wearing policy) to handle a crowd-gathering caused infection outbreak. At that time, city DT B, which has previously used the same response policy R_i, can broadcast its experience through the FL platform by uploading the local updates on time. Then, after obtaining the new global model, city DT A can simulate the feedback of the mask policy R_i by its next iteration of local updates to decide if this policy is necessary. Meanwhile, the existing city DTs can also enhance their model's performance from the historical experience of city A through global updates. Thus, under the FL-based collaborative scheme, multiple city DTs can share the historical experience of different practical conditions by uploading the local parameter updates. Meanwhile, the inherent character of city DT can ensure the high quality of local updates, because each DT is bounded with a real-time data collection basis.

4.2 Local city DT

For the local city DT system, we utilize TCN to model the temporal correlation hidden in the historical data and predict the arrival of future infection trends

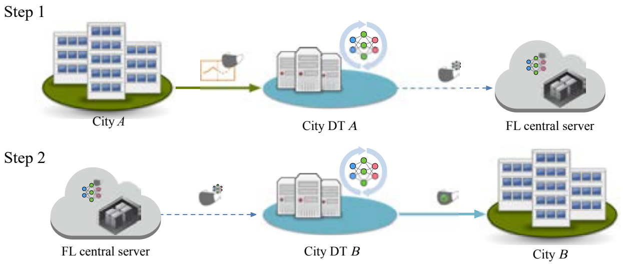


Fig. 4 Example of knowledge exchange between city DTs.

concerning various response plans as **output**. The model is illustrated by Fig. 5, from which we can see a 2-D dilated convolution is adopted in the local training model.

For predicting the future infection trends under various response plans properly, we utilize the historical number of current confirmed cases n_i, the testing number n_t, the confirmed death number n_d, and the response plan p that performed in the past as the **input** of each local city DT. Additionally, normalization is adopted for having all the input features on a similar scale.

The historical data of confirmed cases n_i is convoluted by TCN, including several temporal convolution blocks to generate an output tensor n'_i. Moreover, the features n_t, n_d, and p compose the 3-D tensor v, which is convoluted by a 2-D convolutional network to generate another output tensor v'. We concatenate n_i and v' as n_v, which is further convoluted by TCN to obtain n'_v. Finally, the future infection trend can be considered as the output of batch normalization and a fully connected neural network. For clarity, we also illustrated the **model pipeline** in Fig. 5, which contains 2-D dilated causal convolution layer, normalization layer, activation function, dropout, and residual connection. Instead of directly using the sum of the input and the output in the residual function, our input is convoluted by a 2-D convolution to transform it into the same shape of the output for the addition operation.

4.3 Federated city DTs

The federated training process aims to provide a collaborative training protocol without violating the privacy rules of each city DT, and the entire training diagram is depicted in Fig. 6. This training process includes the following steps:

Step 1: In the beginning, the FL central server trains a global model using a pubic data source or voluntary data set from a DT as pre-training, then opens the platform for collaborators to join the training task. In addition, when a new city DT enters the FL paradigm, the FL central server starts the initialization step again and uses the global model of the last iteration as the new global model.

Step 2: The FL central server distributes the global model to all the existing city DTs, and each DT trains the model by the latest local data set to generate the local updates.

Step 3: Each city DT uploads the regional updates to the FL central server for the aggregation step.

Step 4: The FL central server aggregates all the local updates using the aggregation algorithm to generate an updated global model and distributes this model to each city DT.

This iteration repeats several times until the global model achieves the expected performance. Finally, each city DT can always obtain the latest model with another city DT's local updates. That is, our proposed FL platform can provide an enhanced model for prediciting of COVID-19 infection trend under different response plans, by which each city DT can determine the response based on crowd-sourcing intelligence without sacrificing privacy.

5 Simulation

This section validates our proposed FL-based collaborative framework for multiple city DTs through extensive experiments. First, we give a detailed description of the applied dataset and experimental settings of the framework. Then, multiple aspects of experimental results, comparisons, and analysis are provided, especially the comparisons between non-federated methods and our proposed method in terms of prediction accuracy.

5.1 Datasets and experimental settings

Dataset description: We used the COVID-19 tracking project dataset (https://www.dolthub.com/repositories/Liquidata/corona-virus-state-action) and the COVID-19

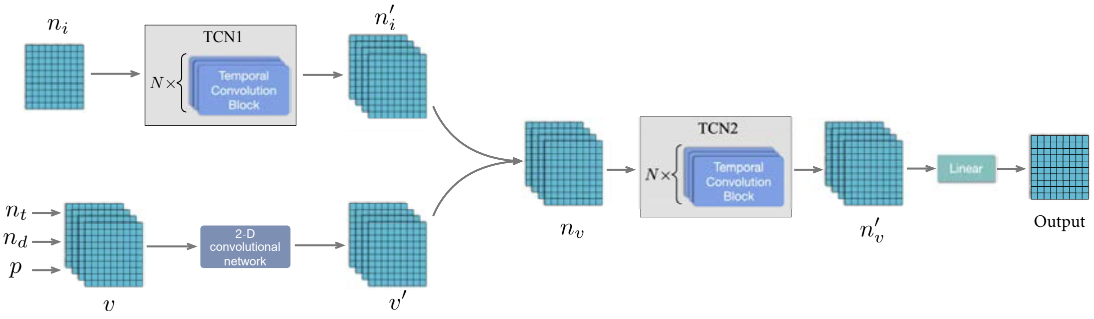


Fig. 5 Local city DT model: Structure of the TCN model.

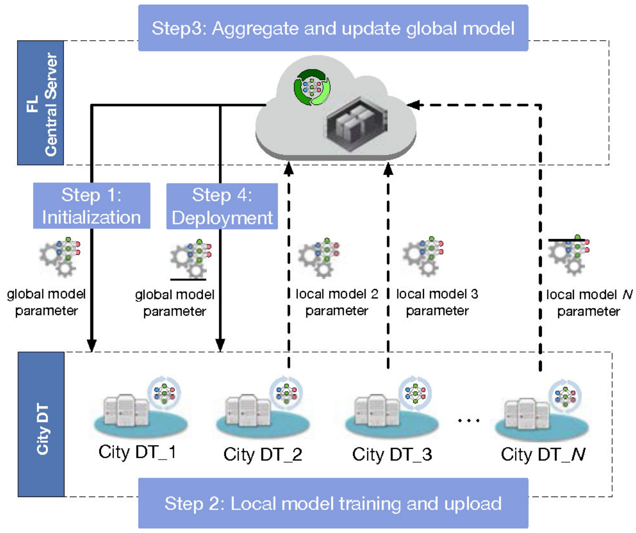


Fig. 6 Illustration of the proposed federated training process.

State Actions Dataset (https://covidtracking.com/data) to conduct experiments. The COVID-19 tracking project dataset contains each state's daily epidemic data from the first case in the United States from January 2020 to July 2020, including information on the number of nucleic acid tests, the number of confirmed cases, and the number of severe cases. The COVID-19 State Actions Dataset mainly contains specific information about the policies adopted by various states in the United States during the epidemic, including the response policy's name, the start time, and the end time. Specifically, we also consider several external factors, such as the change in the number of infections in each state in a time zone and the implementation of various response policies. Therefore, we combine the two datasets for our model training. In our combined dataset,

- *State name* indicates the state where the current data are located.
- *Date* represents the current data time.
- *Data*—($t - 13, \dots, t + 7$) contains a total of 21 days of epidemic data starting at time t (13 days forward and 7 days backward), and each day's data are determined by the number of people diagnosed and the current response policy.

A one-hot vector represents every response plan, which includes several response policies (e.g., gathering limits, statewide school closures, and statewide mask policies). It is represented as 1 if selected; otherwise 0.

Baseline: We compare our performance with the baseline: Seq2Seq. It is adopted for COVID-19 forecasting and is a commonly used method of encoder-decoder for predicting time-series data. In our simulation, the encoder is with three-layer Gated Recurrent Unit (GRU) and the decoder is with three fully connected layers and a three-layer GRU.

Simulation settings and parameters: We conduct all the experiments on the PyTorch platform, and use PySyft[40] framework to implement our FL protocol. We set the learning rates of the neural networks as 0.005, mini-batch size as 60, and evaluate the accuracy of prediction results with Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE), which are defined as follows:

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{y_i - x_i}{y_i} \right|}{n},$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n},$$

where y_i is the prediction results, x_i is the true value, and n represents the total number of data points.

5.2 Experimental results

We compare the performance of all the methods, and the results for each method are summarized in Table 1. The Seq2Seq model is the baseline model. The centralized solution to simulate a central server that has all historical nationwide data with no privacy protection guarantee. The local method indicates a normal DT model, which has only local historical data instead. Both of the centralized solution and local method make the prediction using the TCN model, which is identical to each city DT model in our proposed FL framework. As shown in the Table 1, the performance of all four methods decreases when the date of prediction is extended, while our proposed FL method outperforms

Table 1 Performance comparison of all methods.

Method	Day 1		Day 2		Day 3		Day 4		Day 5		Day 6		Day 7	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Seq2Seq	4374.72	0.2137	3996.51	0.1822	4213.53	0.1477	4779.57	0.0764	5681.11	0.2924	6731.24	0.6018	7850.29	0.5856
Local	6598.40	0.1531	6789.24	0.1562	7064.19	0.1616	7354.41	0.1624	7704.93	0.1675	8024.07	0.1733	8103.88	0.1768
Centralized	565.14	0.0211	659.06	0.0236	948.52	0.0298	1335.30	0.0359	1769.01	0.0434	2191.02	0.0505	2628.33	0.0573
FL	2424.68	0.0631	2989.68	0.0829	3459.44	0.0943	4025.78	0.0971	4493.98	0.0993	4949.64	0.1016	5394.88	0.1042

the Seq2Seq and the local DT methods.

A similar observation pertains to Fig. 7, where we select several states to further perform the performance comparison of the centralized method (the red line), the ground truth (the yellow line), and our FL method (the red dotted line). This is because of the fact that FL enable each local DT to share historical experiences/knowledge, which results in more accurate forecasting. Note that, the centralized method can always obtain the best performance since it can access to all the raw data from available states, while the local DT model has the worst performance due to the very limited local historical data. Thus, the centralized method, represented by the red line in Fig. 7, shows the same trend as the ground truth. Meanwhile, our method's performance is close to that of the centralized method, which confirms that our FL method can ensure a similar prediction performance without sacrificing privacy.

Moreover, we conduct regional forecasting for COVID-19 to further analyze the above prediction results. For the regional forecasting task, we run the three solutions (centralized, local DT, and FL) under different response plan settings (no response plan, implement current response plan, and implement all response plans) for each state in the USA. For our federated learning model, FedAvg is adopted as the aggregation algorithm to save communication resources. The selected response policy includes domestic travel limitations, gathering limits and stay-at-home orders, nonessential business closures, reopening plans and task forces, and statewide mask policies.

For clarity, we select eight states to illustrate the performance comparison results, which are shown in Fig. 7, and we observe a performance gap among the three methods similar to that in Table 1. The centralized method (the red line) is very close to the ground truth (the yellow line), while our FL method (the red dotted line), the centralized method, and the ground truth method have very similar trends. The performance gap between our FL method and the centralized method is significantly smaller than that between the local method (the red dashed line) and centralized method.

In Fig. 7, we also observe a significant difference in the case number prediction depending on whether no response plan, all response plans, or the current response plan is implemented. The predicted infected number is much lower when all response plans or the current response plan are implemented compared to the no response plan. This result further highlights the fact that a positive response plan can benefit COVID-

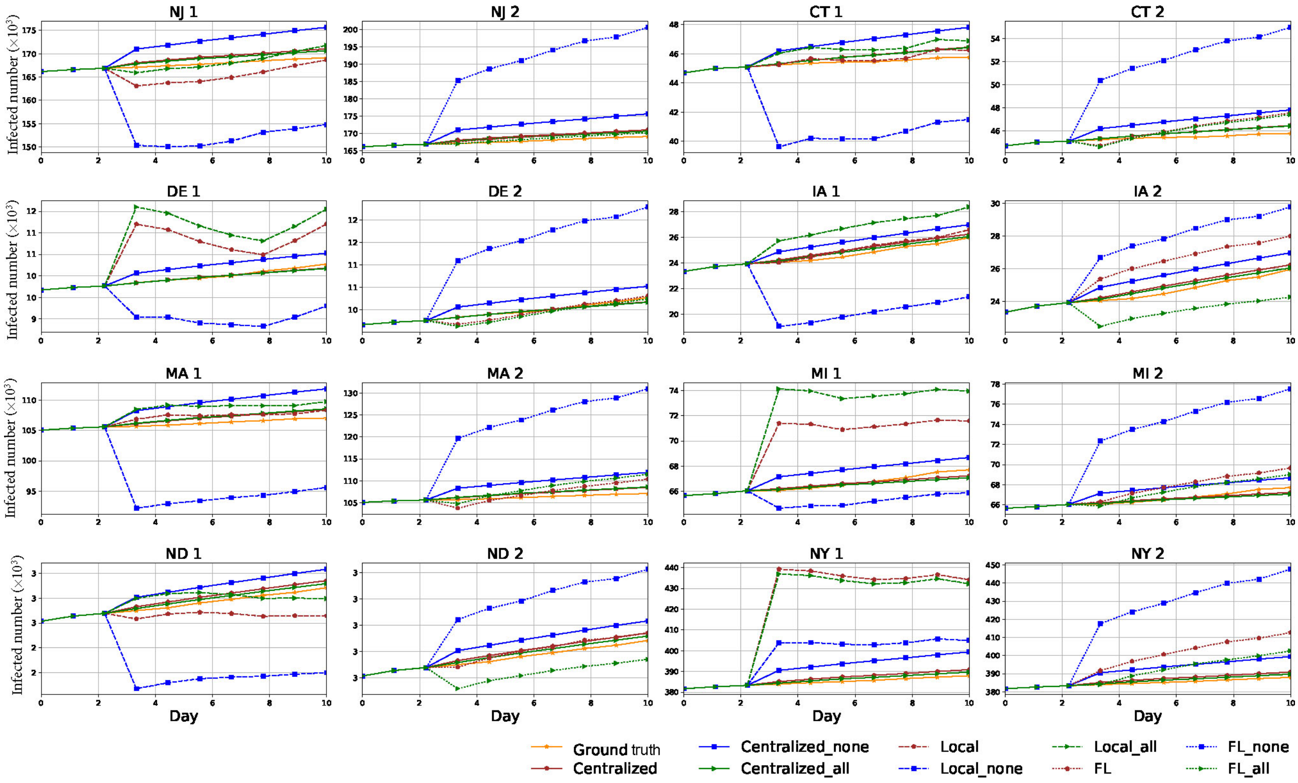


Fig. 7 Prediction results comparison analysis: future infection trends of different states under different response plans using federated and non-federated solutions (for each state, e.g., NJ 1 uses only local data and NJ 2 uses FL).

19 pandemic control and decrease the case number effectively, and it indicates that the predicted case number of all 8 states could be dramatically increased without a response plan for epidemic control. We observe a small difference in the predicted case number when the current response plan and all response plans are implemented, and this is due to the overlap of the two response plan sets. In particular, in all 8 states, the centralized method has the closet trends with the ground truth curve, whereas the local method has the least similar trends. Our proposed FL method and the centralized method have similar trends across all scenarios. This result indicates that through FL-training, each city DT can exchange its epidemic information to improve its DT model, so that the prediction performance for the COVID-19 trend under selected response policies exploits a more accurate forecasting than that of a local city DT.

We also tested the effects of parameter settings for the city DT model and FL aggregation process (FedAvg). For the city DT, we use different settings on the number of blocks and learning rate (lr), and the results are shown in Fig 8. We observe the impressive performance gains when the block is set to 5 and learning rate is set to 0.001. The effect of the number of rounds (the number of communications between the FL server and each city DT) and epochs (the number of local trainings at each city DT) on FedAvg, shown in Fig. 9, demonstrates that the different parameter settings affect the results, and we investigate the best performance when the round = 200 and epoch = 50. This result indicates that increasing the two parameters shows minimal performance gains if enough communication rounds and epochs are used, which further verifies that our proposed method is robust to the training parameters.

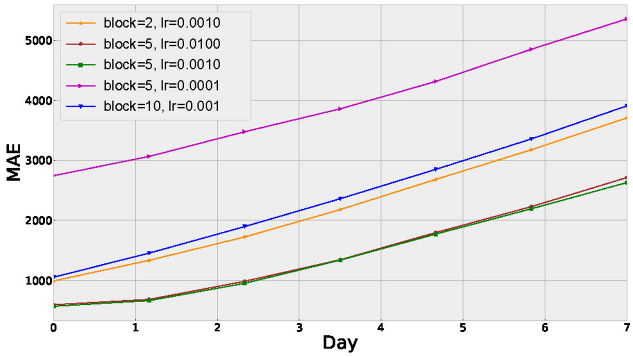


Fig. 8 MAE analysis of the TCN under different training parameters.

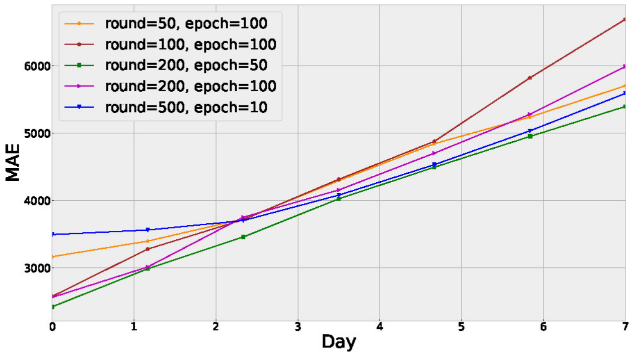


Fig. 9 MAE analysis of FL under different training parameters.

6 Conclusion

This paper proposes a novel collaborative city DT framework based on FL techniques for COVID-19 response plan management, with a TCN structure to better capture temporal contexts in historical infection data. Our work shows significant potential for establishing an intelligent model for novel infectious diseases. It shows that the combination of FL and city DTs helps alleviate the data sparsity challenge, achieve collaboration, and provide privacy protection by design. Our intensive experiments verify that our approach offers improved performance on a real dataset. Further, the proposed framework can be generalized to other collaborative training problems, such as disaster surveillance and prediction. In the future, we aim to involve more data sources (e.g., movement of people, seasonal changes, temperature, and humidity) to improve the proposed framework's performance.

Acknowledgment

This work was supported in part by the National Key Research and Development program of China (Nos. 2017YFA0604500 and 2018YFB2100303), in part by the National Natural Science Foundation of China (No. 61701190), in part by the Key Technology Innovation Cooperation Project of the Government and University for the Entire Industry Demonstration (No. SXGJSF2017-4), and the Program for Innovative Postdoctoral Talents in Shandong Province (No. 40618030001).

References

[1] WHO, Coronavirus disease (COVID-19) pandemic, https://www.who.int/emergencies/diseases/novel-coronavirus-2019, 2020.

[2] S. Kumar and M. Singh, Big data analytics for healthcare

industry: Impact, applications, and tools, *Big Data Mining and Analytics*, vol. 2, no. 1, pp. 48–57, 2019.

[3] W. Zhong, N. Yu, and C. Y. Ai, Applying big data based deep learning system to intrusion detection, *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 181–195, 2020.

[4] H. Nieto-Chaupis, Face to face with next flu pandemic with a wiener-series-based machine learning: Fast decisions to tackle rapid spread, in *Proc. 2019 IEEE 9^{th} Annu. Computing and Communication Workshop and Conf.*, Las Vegas, NV, USA, 2019, pp. 654–658.

[5] B. Adhikari, X. F. Xu, N. Ramakrishnan, and B. A. Prakash, EpiDeep: Exploiting embeddings for epidemic forecasting, in *Proc. 25^{th} ACM SIGKDD Int. Conf. Knowledge Discovery* & *Data Mining*, Anchorage, AK, USA, 2019, pp. 577–586.

[6] Z. B. He, Y. S. Li, J. Li, K. Y. Li, Q. Cai, and Y. Liang, Achieving differential privacy of genomic data releasing via belief propagation, *Tsinghua Science and Technology*, vol. 23, no. 4, pp. 389–395, 2018.

[7] X. Zheng, Z. P. Cai, and Y. S. Li, Data linkage in smart internet of things systems: A consideration from a privacy perspective, *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.

[8] Z. P. Cai, Z. B. He, X. Guan, and Y. S. Li, Collective data-sanitization for preventing sensitive information inference attacks in social networks, *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.

[9] A. Croatti, M. Gabellini, S. Montagna, and A. Ricci, On the integration of agents and digital twins in healthcare, *J. Med. Syst.*, vol. 44, no. 9, p. 161, 2020.

[10] N. Bagaria, F. Laamarti, H. F. Badawi, A. Albraikan, R. A. M. Velazquez, and A. El-Saddik, Health 4.0: Digital twins for health and well-being, in *Connected Health in Smart Cities*, A. El-Saddik, M. S. Hossain, and B. Kantarci, eds. Switzerland: Springer, 2020, pp. 143–152.

[11] 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. 20^{th} Int. Conf. Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[12] J. J. Pang, Y. Huang, Z. Z. Xie, Q. L. Han, and Z. P. Cai, Realizing the heterogeneity: A self-organized federated learning framework for IoT, *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3088–3098, 2021.

[13] J. J. Li, H. H. Jiao, J. Wang, Z. G. Liu, and J. Wu, Online real-time trajectory analysis based on adaptive time interval clustering algorithm, *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 131–142, 2020.

[14] K. Yang, J. H. Zhu, and X. Guo, POI neural-rec model via graph embedding representation, *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 208–218, 2021.

[15] Z. Wang, C. K. Wang, X. J. Ye, J. S. Pei, and B. Li, Propagation history ranking in social networks: A causality-based approach, *Tsinghua Science and Technology*, vol. 25, no. 2, pp. 161–179, 2020.

[16] H. X. Chen, S. Feng, X. Pei, Z. Zhang, and D. Y. Yao, Dangerous driving behavior recognition and prevention using an autoregressive time-series model, *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 682–690, 2017.

[17] Z. L. Ye, H. X. Zhao, K. Zhang, Z. Y. Wang, and Y. Zhu, Network representation based on the joint learning of three feature views, *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 248–260, 2019.

[18] L. J. Wang, J. Z. Chen, and M. Marathe, DEFSI: Deep learning based epidemic forecasting with synthetic information, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 9607–9612, 2019.

[19] L. Zhao, J. Z. Chen, F. Chen, W. Wang, C. T. Lu, and N. Ramakrishnan, SimNest: Social media nested epidemic simulation via online semi-supervised deep learning, in *Proc. 2015 IEEE Int. Conf. Data Mining*, Atlantic City, NJ, USA, 2015, pp. 639–648.

[20] B. Y. Shi, J. N. Zhong, Q. Bao, H. J. Qiu, and J. M. Liu, EpiRep: Learning node representations through epidemic dynamics on networks, in *Proc. 2019 IEEE/WIC/ACM Int. Conf. Web Intelligence*, Thessaloniki, Greece, 2019, pp. 486–492.

[21] Y. Z. Zhou, D. Zhang, and N. X. Xiong, Post-cloud computing paradigms: a survey and comparison, *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 714–732, 2017.

[22] H. Yang, F. Li, D. X. Yu, Y. F. Zou, and J. G. Yu, Reliable data storage in heterogeneous wireless sensor networks by jointly optimizing routing and storage node deployment, *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 230–238, 2021.

[23] F. S. Lu, M. W. Hattab, C. L. Clemente, M. Biggerstaff, and M. Santillana, Improved state-level influenza nowcasting in the United States leveraging internet-based data and network approaches, *Nature Communications*, vol. 10, p. 147, 2019.

[24] B. Zou, V. Lampos, and I. Cox, Multi-task learning improves disease models from web search, in *Proc. 2018 World Wide Web Conf.*, Lyon, France, 2018, pp. 87–96.

[25] Z. P. Cai and X. Zheng, A private and efficient mechanism for data uploading in smart cyber-physical systems, *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 766–775, 2020.

[26] Z. P. Cai, X. Zheng, and J. G. Yu, A differential-private framework for urban traffic flows estimation via taxi companies, *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.

[27] C. Fan, Y. C. Jiang, and A. Mostafavi, Social sensing in disaster city digital twin: Integrated textual–visual–geo framework for situational awareness during built environment disruptions, *Journal of Management in Engineering*, vol. 36, no. 3, p. 04020002, 2020.

[28] C. Fan, C. Zhang, A. Yahja, and A. Mostafavi, Disaster city digital twin: A vision for integrating artificial and human intelligence for disaster management, *International Journal of Information Management*, vol. 56, p. 102049, 2021.

[29] A. Francisco, N. Mohammadi, and J. E. Taylor, Smart city digital twin–enabled energy management: Toward real-time urban building energy benchmarking, *Journal of Management in Engineering*, vol. 36, no. 2, p. 04019045, 2020.

[30] F. Xue, W. S. Lu, Z. Chen, and C. J. Webster, From LiDAR point cloud towards digital twin city: Clustering city

objects based on gestalt principles, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 418–431, 2020.

[31] W. J. Holstein, Virtual Singapore-creating an intelligent 3D model to improve experiences of residents, business and government, http://www.3dexperiencecity.com/, 2016.

[32] F. Dembski, U. Wössner, M. Letzgus, M. Ruddat, and C. Yamu, Urban digital twins for smart cities and citizens: The case study of Herrenberg, Germany, *Sustainability*, vol. 12, no. 6, p. 2307, 2020.

[33] Z. P. Cai and Z. B. He, Trading private range counting over big IoT data, in *Proc. 39th IEEE Int. Conf. Distributed Computing Systems*, Dallas, TX, USA, 2019, pp. 144–153.

[34] X. Zheng and Z. P. Cai, Privacy-preserved data sharing towards multiple parties in industrial IoTs, *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 968–979, 2020.

[35] A. Hard, K. Rao, R. Mathews, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, Federated learning for mobile keyboard prediction, arXiv preprint arXiv: 1811.03604, 2018.

[36] Y. Q. Chen, X. Qin, J. D. Wang, C. H. Yu, and W. Gao, FedHealth: A federated transfer learning framework for wearable healthcare, *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, 2020.

[37] G. D. Long, Y. Tan, J. Jiang, and C. Q. Zhang, Federated learning for open banking, in *Federated Learning. Lecture Notes in Computer Science, vol 12500*, Q. Yang, L. X. Fan, and H. Yu, eds. Switzerland: Springer, 2020, pp. 240–254.

[38] S. J. Bai, J. Z. Kolter, and V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv: 1803.01271, 2018.

[39] N. Mohammadi and J. E. Taylor, Smart city digital twins, in *Proc. 2017 IEEE Symp. Series on Computational Intelligence*, Honolulu, HI, USA, 2017, pp. 1–5.

[40] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, A generic framework for privacy preserving deep learning, arXiv preprint arXiv: 1811.04017, 2018.

Junjie Pang received the MS and PhD degrees from the Department of Computer Science, Jilin University in 2013 and 2017, respectively. He currently holds a post-doctoral position at Qingdao University. He is the recipient of the Initiative Postdocs Program of Shandong Province. His research interests include federated learning, reinforcement learning, and next-generation networking.

Yan Huang received the PhD degree from the Department of Computing Science, Georgia State University, Atlanta, USA in 2019. He is currently an assistant professor at the Department of Software Engineering & Game Development, Kennesaw State University (KSU). He is broadly interested in privacy and security, with particular emphasis on deep learning-aided privacy protection solutions and cybersecurity challenges in the IoT environment. Now, his research focuses on improving the FL's performance and efficiency and alleviating the contradictions between multiple training tasks.

Jianbo Li received the PhD degree in computer science and technology from University of Science and Technology of China in 2009. From 2013 to 2014, he was a visiting scholar at Fordham University. He is currently a professor at the College of Computer Science and Technology, Qingdao University. He is the chairman of ACM Qingdao Branch, the deputy secretary general of Qingdao Computer Society, a senior member of the China Computer Federation, and a member of the Internet of Things professional committee of the China Computer Federation. His research interests include urban computing, mobile social networks, and data offloading.

Zhenzhen Xie received the MEng degree in computer science from Jilin University, China in 2014, where she is currently a PhD candidate at the College of Computer Science and Technology. Her research areas are reinforcement learning, IoTs, and representation learning.

Zhipeng Cai received the BS degree from Beijing Institute of Technology, China in 2001, the MS and PhD degrees from University of Alberta, USA in 2004 and 2008, respectively. He is currently an assistant professor at the Department of Computer Science, Georgia State University. His research interests include networking, privacy, and big data. He has published more than 50 journal papers, including more than 20 IEEE/ACM transaction papers, such as *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE/ACM Transactions on Networking*, and *IEEE Transactions on Mobile Computing*. He is the recipient of an NSF CAREER award. He is an editor/guest editor for *Algorithmica*, *Theoretical Computer Science*, *Journal of Combinatorial Optimization*, and *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. He is a senior member of the IEEE.