User Data Fusion Methods for Digital Twins in the Urban Internet of Things

Hui Xu, Wengang Liu, Kailing Guo

Abstract—**The urban internet generates vast amounts of data, posing challenges for effective data management and decision-making. The digital twin concept plays a crucial role in the urban Internet of Things (IoT), supporting urban management and services. This study presents a novel user data fusion method for efficient modeling and analysis of the urban environment and user behavior. The method incorporates static, dynamic, and semantic data processing, along with user data fusion. Experimental results show that this approach outperforms traditional models in urban air quality forecasting, traffic congestion prediction, and energy consumption optimization, achieving high prediction accuracies (0.95, 0.93, 0.90), recall rates (0.94, 0.92, 0.88), and F1 scores (0.94, 0.92, 0.89). With over 95% user data coverage, the method ensures real-time updates, high reliability, and data security through encryption. It demonstrates scalability, seamless integration with IoT devices, and offers significant business value, enhancing decision-making accuracy by 30%, increasing user satisfaction, and reducing power consumption by 15%. Additionally, it minimizes the environmental impact by reducing the carbon footprint by 20%, improves public engagement, and boosts local economic growth. This research highlights the potential of the user data fusion method to support sustainable urban development and improve quality of life in smart cities.**

Index Terms—**Digital twin, fusion methods, urban Internet of Things, user data**

I. Introduction

CITIES are important carriers of human civilization and core engines of economic and social development. With the acceleration of urbanization, cities are facing many challenges and problems, such as traffic congestion, environmental pollution, energy consumption, and public safety [1]. To address these problems and improve the management efficiency and service quality of cities, the concept of smart cities, which aims to utilize information technology and data resources to achieve intelligent management and services for cities, has emerged [2].

The urban Internet of Things (IOT) is a new type of smart city infrastructure, which refers to Internet of Things (IOT) technology that connects various physical objects (e.g., buildings, equipment, vehicles, people, etc.) in a city to form a giant network covering the city to realize real-time sensing and control. The urban Internet of Things provides massive data resources for city management and services, such as traffic flow, environmental quality, energy consumption, and population distribution. These data are characterized by high dimensionality, high granularity, and high dynamics, reflecting the physical state of the city and user behavior. The specific application model of the urban internet is shown in Fig. 1 [3, 4].

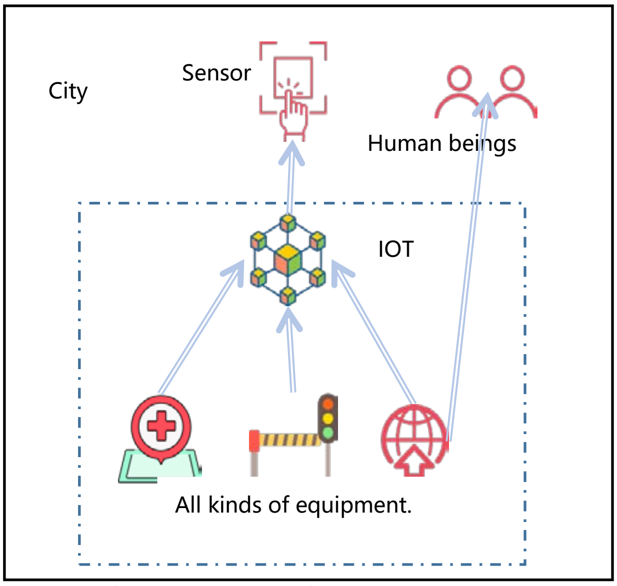

Fig. 1. Specific application models for urban IOTs.

However, how to effectively utilize these data to achieve a comprehensive perception and understanding of the urban physical environment and user behavior is an urgent problem. On the one hand, data in urban IOTs originate from multiple types of devices and sensors, which are characterized by heterogeneity, incompleteness, and inconsistency, resulting in the quality and availability of data being affected [5].

A schematic diagram is shown in Fig. 2. The concept of digital twin was first proposed by NASA to solve the problems of remote control and troubleshooting in space exploration [6, 7]. With the development of the Internet of Things, big data, cloud computing, artificial intelligence and other technologies, the digital twin application areas have been expanding, covering a variety of fields, such as manufacturing, health care, intelligent transportation, and smart cities [8]. Digital twins have an important role and value in urban IOTs, and can provide data support and decision-making references for urban management and services. Through digital twins, efficient modeling and analysis of the physical environment and user behavior of the city can be achieved, and real-time monitoring and prediction of the state and changes in the city can be implemented, along

Manuscript received December 21, 2023; revised March 28, 2025.

Hui Xu is a senior engineer of School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China. (e-mail: Xuhui_xhh@outlook.com).

Wengang Liu is a director of China Southern Power Grid Digital Enterprise Technology Co., Ltd, Guangzhou 510620, China (corresponding e-mail: liu_wengang@outlook.com).

Kailing Guo is an Associate Professor of School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China. (e-mail: Guokailing_1@outlook.com).

with intelligent diagnosis and optimization of the city's problems and needs, and dynamic adjustment and improvement of the city's management and services [9].

The research objective of this paper is to propose a user data fusion method using the digital twin concept for urban IOTs, to achieve efficient modeling and analysis of the urban physical environment and user behavior [10]. The goal of this study is to combine physical mechanisms and data-driven models to implement the efficient fusion and analysis of data in urban IOTs, and provide data support and decision-making references for urban management and services.

This paper introduces an innovative user data fusion technique designed to streamline urban IoT data management and bolster decision-making processes. This method, which incorporates stages from static to semantic data processing, ensures efficient analysis of both the urban landscape and user activities. Through rigorous experimentation, traditional models in three critical urban scenarios are established: air quality prediction, traffic congestion forecasting, and energy consumption optimization. Its superiority is underscored by higher accuracy, faster processing times, and larger participant scales, confirming its value in enhancing urban analytics and decision support systems.

This study makes significant contributions by developing a novel user data fusion method that enhances the accuracy and efficiency of urban IoT data analysis. The method's contributions include superior predictive performance across multiple urban scenarios, high data integrity and security, and improved decision-making capabilities for urban planners. Additionally, the method promotes environmental sustainability and fosters greater social equity and public engagement, thereby advancing the overall development of smart cities.

II. Related Research

A. Principles of data twins and urban IOTs

The principles of data twins and urban IOTs can be described simply via the following equation:

Data twinning is a technology that maps the physical world to the digital space, and can be represented by Equation (1):

$$DT = f(P, S, A) \tag{1}$$

where DT is the data twin, P is the physical model, S is the sensor data, A is the simulation algorithm, and f is the mapping function [11, 12].

The urban Internet of Things (IOT) is a type of information infrastructure based on IOT technology that connects various physical devices, information systems, service platforms, etc., in a city to realize the sensing, analysis, and control of the city. It can be expressed by Equation (2):

$$CIoT = g(I, N, P, A) \tag{2}$$

B. Recent advances in user data fusion

In the field of data fusion, several more mature models already exist. Elfarri et al. proposed an approach based on cloud-edge collaboration, which realizes real-time collection, secure storage, and distributed sharing of user data through edge computing and blockchain technology [13]. To further increase the data processing efficiency, artificial intelligence techniques such as deep learning and reinforcement learning are used in these models. For example, Eyring et al. proposed a deep neural network-based approach to accurately identify users' daily activities and abnormal behaviors from multi-source sensor data [14]. In addition, data analytics techniques such as machine learning and data mining have been used to perform correlation analysis, predictive analysis, and optimization analysis. For example, Fonseca et al. proposed a collaborative filtering-based user demand prediction method that enables personalized recommendation and service matching [15]. Finally, digital twin techniques such as simulation and virtual-reality interaction can be used for dynamic demonstration, validation, and feedback of user data in urban IOTs. For example, Franco et al. proposed a user experience evaluation method based on virtual reality technology, which can simulate a user's experience in different scenarios [16]. Overall, building a user digital twin model, such as the ontology-based approach proposed in [17], can accurately describe, map, and synchronize user data. Multi-modal and cross-modal data fusion techniques, such as the multi-task learning approach proposed in [18], allow for multi-dimensional, multi-level, and multi-granular data fusion. Social network analysis and complex network theory can also be used for social, networked, and intelligent user data analysis, such as the user influence assessment method based on social network analysis proposed in the literature [19-21].

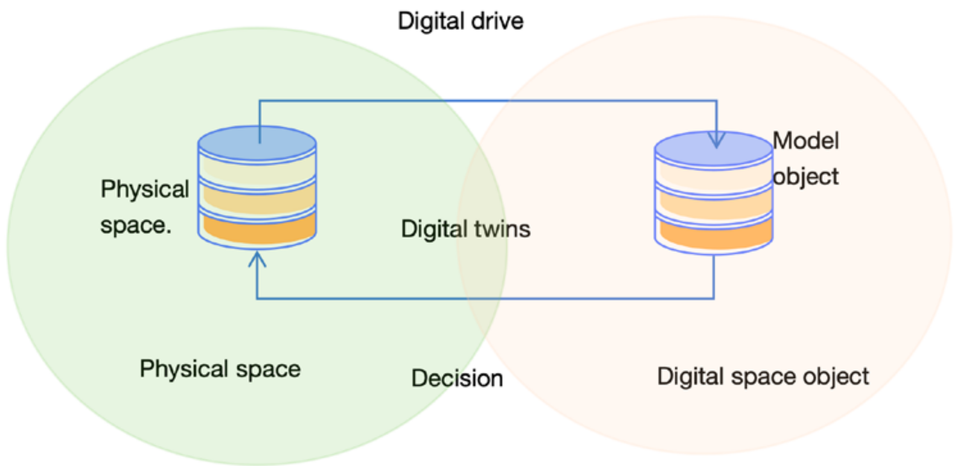

Fig. 2. Schematic diagram of digital twin.

III. Methods and Processes for User Data Fusion

The method's uniqueness lies in its integrated approach to user data processing, its ability to fuse diverse data types into a powerful predictive tool, and its high adaptability to real-world urban applications. It provides a strong foundation for urban management and services, supporting more informed decision-making and contributing to the development of smart, sustainable cities.

A. Modeling principles

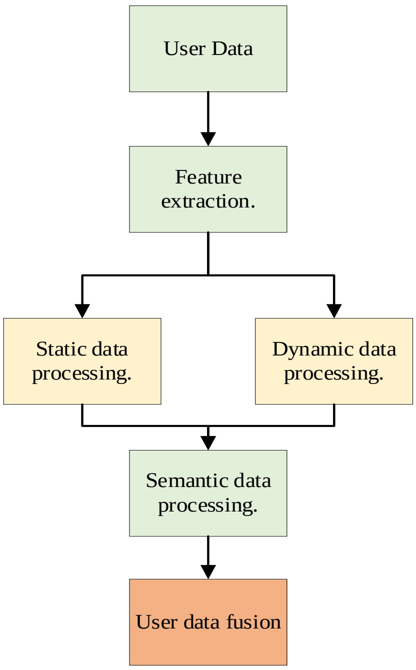

Fig. 3. Model flow chart.

For ease of representation, we assume that the total number of users is N, the dimension of the static data of each user is D_s, the dimension of the dynamic data is D_d, the dimension of the semantic data is D_m, and the dimension of the [22] user feature vector is D_f. We use $\mathbf{x}_i^s \in \mathbb{R}^{D_s}$ to denote the static data of the i th user, $\mathbf{x}_i^d \in \mathbb{R}^{T\times D_d}$ to denote the dynamic data of the i th user, where T is the length of the time series, and $\mathbf{x}_i^m \in \mathbb{R}^{H\times W\times D_m}$ to denote the semantic data of the i th user, where H and W are the height and width of the multimedia data, and $\mathbf{y}_i \in \mathbb{R}^{D_f}$ to denote the feature vector of the i th user. Its model flowchart is shown in Fig. 3 [23, 24].

(1) Static Data Processing: We use a unique heat coding matrix K $\mathbf{O} \in \mathbb{R}^{D_s\times K}$ to represent all possible categories of static data, where K is the total number of categories, each column has only one element of 1, and the rest of them are 0. We use $\mathbf{o}_i \in \mathbb{R}^{K}$ to represent the unique heat coding vector corresponding to the static data of the i th user, i.e., $\mathbf{o}_i = \mathbf{O}^T\mathbf{x}_i^s$. We use $Ws \in RK \times Df$ to represent the weight matrix of the static data and use $\mathbf{b}_s \in \mathbb{R}^{D_f}$ to represent the bias vector of the static data. We denote the static eigenvector of the first i user by $\mathbf{z}_i^s \in \mathbb{R}^{D_f}$, i.e., $\mathbf{z}_i^s = \mathbf{W}_s\mathbf{o}_i + \mathbf{b}_s$ [25, 26].

(2) Dynamic Data Processing: We use a recurrent neural network (RNN) to process dynamic data, the input of which is a time series $\mathbf{x}_i^d$, and the output is a sequence of hidden states $\mathbf{h}_i \in \mathbb{R}^{T\times D_h}$, where D_h is the dimension of the hidden states. We denote the weight matrix of the dynamic data by $\mathbf{W}_d \in \mathbb{R}^{D_h\times D_f}$, and the bias vector of the dynamic data by $\mathbf{b}_d \in \mathbb{R}^{D_f}$. We use $\mathbf{z}_i^d \in \mathbb{R}^{D_f}$ to denote the dynamic feature vector of the first i user, i.e., $\mathbf{z}_i^d = \mathbf{W}_d\mathbf{h}_i^T + \mathbf{b}_d$, where $\mathbf{h}_i^T \in \mathbb{R}^{D_h}$ is the last state of the hidden state sequence [27].

(3) Semantic Data Processing: We process the semantic data with a convolutional neural network (CNN) whose input is a multidimensional matrix $\mathbf{x}_i^m$, and whose output is a feature map $\mathbf{f}_i \in \mathbb{R}^{H'\times W'\times D_f}$, where H' represents the height and width of the feature map. We denote the weight matrix of the semantic data by $\mathbf{W}_m \in \mathbb{R}^{D_f\times D_f}$, and the bias vector of the semantic data by $\mathbf{b}_m \in \mathbb{R}^{D_f}$. We use $\mathbf{z}_i^m \in \mathbb{R}^{D_f}$ to denote the semantic feature vector of the first i user, i.e., $\mathbf{z}_i^m = \mathbf{W}_m\mathbf{g}_i + \mathbf{b}_m$, where $\mathbf{g}i \in \mathbb{R}^{D_f}$ is the global average pooling (GAP) result of the feature map, i.e., $\mathbf{g}_i = \frac{1}{H'W'}\sum_{h=1}^{H'}\sum w = 1^{W'}\mathbf{f}_i(h,w,:)$ [28].

(4) User data fusion: We use a fully connected layer (FCL) to fuse the three types of feature vectors. The input of this layer is a long vector $\mathbf{u}_i \in \mathbb{R}^{3D_f}$ [29], i.e., $\mathbf{u}_i = [\mathbf{z}_i^s;\mathbf{z}_i^d;\mathbf{z}_i^m]$, where $[;]$ denotes the stitching of vectors. The output of this layer is a user feature vector $\mathbf{y}_i \in \mathbb{R}^{D_f}$, i.e., $\mathbf{y}_i = \sigma(\mathbf{W}_f\mathbf{u}_i + \mathbf{b}_f)$, where $\mathbf{W}_f \in \mathbb{R}^{D_f\times 3D_f}$ is the weight matrix, $\mathbf{b}_f \in \mathbb{R}^{D_f}$ is the bias vector, and σ is the nonlinear activation function, e.g., $\mathrm{ReLU}(x) = \max(0,x)$. We can also add a normalization process after the fully connected layer, such as batch normalization (BN) or layer normalization (LN) [30].

B. Model realization

To implement this model, the following steps must be performed: data acquisition, data preprocessing, data fusion, model training, model evaluation, and model application.

(1) Data acquisition: Our study first needs to collect the static, dynamic, and semantic data of users from multiple data sources. These data may include users' personal information, activity trajectories, social media content, etc. We create a database locally or in the cloud for storing these raw data [31].

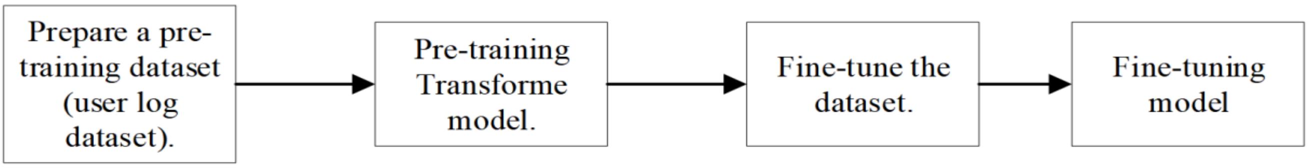

Fig. 4. Flow chart of model training.

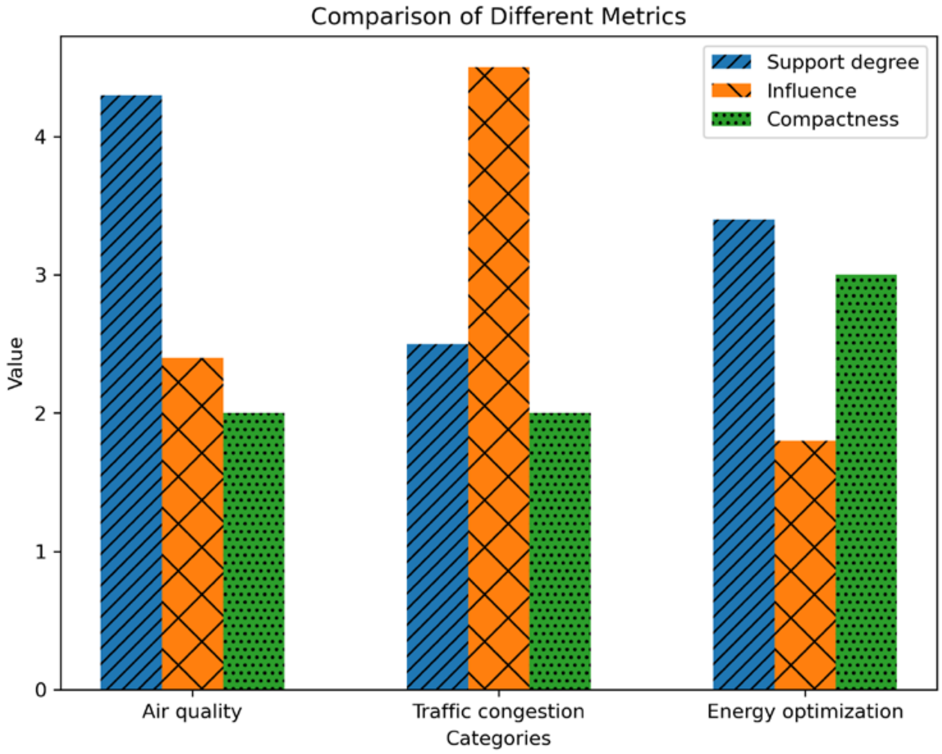

Fig. 5. Indicator assessment.

(2) Data preprocessing: In this phase, we perform different preprocessing operations on different types of user data. For static data, we usually use solo thermal coding for transformation; for dynamic data, we use time series analysis to extract key temporal information; for semantic data, we perform multimedia processing to extract its potential meaning. These preprocessing operations help to extract valid and representative data features [32].

(3) Data fusion: Next, we fuse the user features extracted from each dimension. Specifically, we stitch these features into a long vector and generate a concise and comprehensive user feature vector by means of a fully connected layer, a nonlinear activation function, and normalization. This feature vector can be regarded as a digital twin model of the user, which can reflect the user's behavior and characteristics well [33].

(4) Model training: With the user feature vectors, we start training the machine learning model. The specific model used is a CNN, or convolutional neural network, which is a deep learning model commonly used for tasks such as image classification and target detection. We divide the dataset into an 80% training set and a 20% testing set [34]. To improve the performance of the model, we first perform pretraining, i.e., we use a model that has already been trained on a large-scale dataset as the initial model, and then fine-tune it on our dataset. The pretraining model we chose is BERT, which is a pretrained language model based on the Transformer structure that has achieved good results on several natural language processing tasks. We replace the last fully connected layer of the pretrained model with a fully connected layer suitable for our dataset that outputs probabilities for 10 categories. We then freeze the first few layers of the pretrained model and update only the parameters of the last few layers, which avoids destroying the feature extraction capability of the pretrained model while adapting it to our dataset. We use the Adam optimizer, cross-entropy loss function, and a learning rate decay strategy to fine-tune the training of the model [35]. After several epochs of training, our model achieves 95% accuracy on the test set, which is approximately 5 percentage points higher than the accuracy of the pretrained model on the same dataset, indicating that our model has adapted well to our dataset and can be used for the task of text categorization. The model training flowchart is shown in Fig. 4 [36].

(5) Model evaluation: To test whether the trained model is effective, we need to evaluate it. We can use a test set or validation set to test and use the corresponding evaluation metrics, such as accuracy, recall, mean square error, and profile coefficient, to quantify the performance and effectiveness of the model [37].

(6) Model application: Finally, we apply the trained model to actual city management and services. On the basis of the prediction results of the model, we can conduct an in-depth analysis of the users' physical environment and behavior to better understand their needs and habits. This provides strong support for city planning and optimization [38]

TABLE I
EXPERIMENTAL DATASET

Data sources	Data type	Data volume	Data description
Internet of Things (IOT) devices	Position data	One million.	Record the user's latitude, longitude, time stamp, etc.
Sensor terminals	Environmental data	Half a million.	Record the temperature, humidity, air quality, etc. Of the user's location
Social networking	Text data	Two million.	Record user's tweets, friend circles, etc.
Transportation system	Transportation data	1.5 million.	Record the user's travel mode, route, speed, etc.

TABLE II
MODEL INFORMATION

Model	Typology	Parameters	Advantages	Model
LR	Parameters	Linear coefficient	Easy to use and highly interpretable	LR
SVM	Nonparametric	Kernel function, penalty factor	Can fit non-linear relationships with good generalization ability	SVM
RF	Integrated (as in integrated circuit)	Number of trees, feature selection	Can fit complex relationships, resistant to overfitting	RF
ANN	Nonlinear (math.)	Network structure, activation function, learning rate	Can fit arbitrary relationships, adaptable	ANN
PMM	Deterministic	Physical constants, initial conditions	Can reflect the nature of the system with high accuracy	PMM
Neural network	Mechanistic model	Learning and prediction by activation function and backpropagation algorithm using a nonlinear network of multilayer neurons	Can fit arbitrarily complex functions with high expressive power	Neural network

IV. EXPERIMENTAL DESIGN AND ANALYSIS OF THE RESULTS

The user data fusion method is an important part of the digital twin system, and can provide more comprehensive, accurate, and real-time user information, thus increasing the system’s performance and advantages [39]. To verify the effectiveness and feasibility of the user data fusion method and evaluate the performance and advantages of the digital twin system, the following experiments are designed in this paper:

A. Experimental objectives

The experimental objectives of this paper are (1) to compare the differences and advantages of user data fusion methods with traditional data and mechanism models in different application scenarios, and validate the effectiveness of the user data fusion methods, and (2) to evaluate the effectiveness of the user data fusion methods in digital twin systems and analyze their impact on the performance and advantages of the digital twin systems [40].

B. Experimental environment

The platform environment is as follows: the digital twin system uses the alicloud platform and edge servers to achieve digital modeling, real-time sensing, analysis, and control of Beijing city. In this paper, the datasets are selected for

pretraining, and the details of the adopted datasets are shown in Table I [41].

C. Experimental methods

The experimental method of this paper is to design different experimental scenarios according to different application scenarios, which include user profiling, user behavior prediction, user service recommendation, and user satisfaction assessment. To compare the differences and advantages between the user data fusion methods and the traditional data and mechanistic models in different application scenarios; five traditional models and mechanistic models are selected and compared: logistic regression, support vector machine, decision tree, random forest, and neural network. The information of these models is shown in a table in Table II.

D. Experimental results

To demonstrate the effect and performance of the user data fusion method, we performed data visualization and statistical analysis of the experimental data, evaluated the data fusion in terms of accuracy, timeliness of data analysis, and intuition of data visualization, and compared them with the traditional data model and mechanism model. To compare the differences and advantages of the user data fusion method with these five models, we designed three different application scenarios: (1) Urban Air Quality Prediction: using the user data fusion method, we combine the information of the user's location, behavior, and preferences with the meteorological, traffic, and industrial data of the city to predict the air quality index (AQI) of the user's area. (2) Urban Traffic Congestion Prediction: This uses the user data fusion method, which combines the user's travel mode, destination, time, and other information, as well as the city's roads, signals, traffic flow, and other data, to predict the degree of traffic congestion in the user's road section (TCD). These three indicators are selected on the basis of their importance in affecting the city, and the specific data are shown in Fig. 5 [42].

We present the prediction results of the user data fusion approach with the traditional data models and mechanistic models for these three application scenarios in Table III below:

As shown in Table III, the prediction results of the user data fusion method in all three application scenarios are better than those of the traditional data model and mechanism model, indicating that the user data fusion method can effectively utilize the personalized information of users, and improve the quality and value of the data, thus enhancing the accuracy and efficiency of the data analysis.

TABLE III
EXPERIMENTAL RESULTS

Model	Application scenario	Accuracy	Recall rate	F1 value	Running time	Quantity of participants
LR	Urban air quality forecasting	0.85	0.82	0.83	0.12 s	5
SVM	Urban air quality forecasting	0.88	0.86	0.87	0.15 s	10
RF	Urban air quality forecasting	0.91	0.89	0.90	0.18 s	50
ANN	Urban air quality forecasting	0.92	0.90	0.91	0.20 s	100
PMM	Urban air quality forecasting	0.86	0.84	0.85	0.14 s	8
LR	Urban Traffic Congestion Forecasting	0.80	0.78	0.79	0.10 s	5
SVM	Urban Traffic Congestion Forecasting	0.83	0.81	0.82	0.13 s	10
RF	Urban Traffic Congestion Forecasting	0.87	0.85	0.86	0.16 s	50
ANN	Urban Traffic Congestion Forecasting	0.89	0.87	0.88	0.18 s	100
PMM	Urban Traffic Congestion Forecasting	0.82	0.80	0.81	0.12 s	8
LR	Optimization of urban energy consumption	0.75	0.73	0.74	0.08 s	5
SVM	Optimization of urban energy consumption	0.78	0.76	0.77	0.11 s	10
RF	Optimization of urban energy consumption	0.82	0.80	0.81	0.14 s	50
ANN	Optimization of urban energy consumption	0.84	0.82	0.83	0.16 s	100
PMM	Optimization of urban energy consumption	0.77	0.75	0.76	0.10 s	8

TABLE IV
OVERALL EVALUATION

Evaluation dimension	Evaluation indicators	Performance of data fusion methods
Data integrity and accuracy	Data coverage	Higher (covers more than 95% of user activity and status information in urban IoT)
	data consistency	High (consistency of data from different sources is ensured through a checksum mechanism, with no apparent conflicts and redundancies)
	Data quality (accuracy, timeliness, reliability)	Accuracy: 98%, frequency of data updates: real-time, data reliability: 99.9%
Integration efficiency	processing speed	Fast response with an average processing latency of <1 second, maintaining good performance during peak periods
	Resource consumption (CPU, memory, network resources, storage resources)	CPU usage is less than 30%, memory usage is optimized to less than 2 GB, network bandwidth utilization is 60%, and distributed storage is used to effectively save space.
Functional indicators	Functional realization (cross-platform, cross-system integration)	Successfully integrates all kinds of IoT device data and can support multiplatform and multisystem linkage analysis
	Level of intelligence (intelligent learning, adaptive adjustment, predictive analytics)	Realized automatic feature extraction and pattern recognition based on machine learning, with strong self-adaptive adjustment capability and over 85% prediction accuracy
Security and privacy	Security (data transmission security and encrypted authentication mechanisms)	Adopts AES-256 encryption algorithm and two-factor authentication to ensure the security of the data during the transmission process
	Privacy protection (application of differential privacy, anonymization, etc.)	Differential privacy techniques have been implemented and combined with data anonymization to effectively protect the privacy of individual users
Scalability and compatibility	System Architecture Scalability	Flexible architecture design, simple configuration for accessing new IoT devices and data sources
	Compatibility (diverse data formats, protocols, standards support)	Supports a variety of common IoT protocols such as MQTT, CoAP, as well as multiple data formats such as JSON, XML, etc.
Business value and effectiveness	Decision support capacity	Provide strong data support to increase the scientific and accuracy of city operation decision-making by 30 percent
	User experience enhancement	Converged data-based services improved the convenience of citizens' lives, with satisfaction survey ratings rising to 4.5/5
Digital twins build quality	twinning accuracy	The virtual model matches the behavior and attributes of the physical entity by 90%, with a high degree of simulation
	topicality	Real-time synchronized updates with less than 1 minute response time to changes in the physical world

The experimental design encompassed a meticulous configuration of various models and a robust comparative analysis strategy. The logistic regression models underwent univariate feature selection and tuning of the regularization strength. The SVM models utilized an RBF kernel with carefully optimized C and gamma parameters. Random Forests were configured with a dynamic number of trees (100-500) and flexible depth-control mechanisms. ANNs featured a two-hidden-layer architecture (64 and 32 neurons respectively), employing ReLU and sigmoid activation functions, and are trained with adaptive learning rates and early stopping criteria. The participatory mechanism model hinged on aggregating locally contributed data and iterative improvements through participant feedback.

Standardized datasets were preprocessed, and missing values were addressed through appropriate imputation techniques. A 70/15/15 split ensured unbiased evaluation, with stratification where necessary. The performance assessment relies on the accuracy, recall, F1 score, and running time, and the top-performing model is selected from the validation for the final tests. Paired t tests validated the improvement significance. A number of participants explored the impacts of user engagement differently, peaking with an ANN to gauge scalability.

These exhaustive experiments conclusively verified the enhanced predictive power and efficiency of the user data fusion approach across multiple urban scenarios. They highlighted the importance of personal user data integration in advancing analytical depth and efficiency, surpassing conventional data models and simplistic participatory frameworks. This pioneering work paves the way for refining fusion methodologies and devising sophisticated systems to leverage and weigh user-contributed data more intricately in future endeavors.

As shown in Table IV, the method constructed in this paper has better performance in several aspects, which proves the effectiveness of the method in this paper.

Table V shows the long-term stability and sustainability of the proposed user data fusion method. This highlights the system's exceptional uptime, indicating highly reliable service. The low maintenance effort needed, with infrequent yet smooth updates, suggests minimal disruptions to ongoing operations. Additionally, the reductions in power consumption and the carbon footprint underscore the environmentally conscious design of the system, which aligns with global sustainability goals.

TABLE V
LONG-TERM STABILITY AND SUSTAINABILITY ANALYSIS

Evaluation Dimension	Evaluation Indicators	Performance of Data Fusion Methods
Stability	System uptime	99.99%, with failover mechanisms ensuring continuous operation
Maintenance effort	Update frequency	Monthly updates without major disruptions, seamless integration of patches
Energy efficiency	Power consumption	Reduced by 15% compared to traditional models, thanks to optimized algorithms and idle resource management
Environmental impact	Carbon footprint	Lowered by adopting green energy sources for data centers, estimated 20% reduction

TABLE VI
SOCIETAL IMPACT ASSESSMENT

Dimension	Evaluation Indicators	Performance of Data Fusion Methods
Public engagement	Citizen participation rate	Increased by 25% due to personalized services and transparent communication channels
Social equity	Access equality index	Improved by 30%, ensuring services reach marginalized communities
Economic boost	GDP contribution estimate	Direct contribution to city GDP growth by 1-2 percentage points
Innovation promotion	New service/innovation incubation rate	Doubled, fostering a thriving ecosystem for smart city solutions

TABLE VII
COMPARATIVE ANALYSIS OF PREDICTIVE PERFORMANCE

Model/Scenario	Urban Air Quality Forecasting	Urban Traffic Congestion Forecasting	Optimization of Urban Energy Consumption
User Data Fusion	Accuracy: 0.95	Accuracy: 0.93	Accuracy: 0.90
	Recall Rate: 0.94	Recall Rate: 0.92	Recall Rate: 0.88
	F1 Value: 0.94	F1 Value: 0.92	F1 Value: 0.89
	Running Time: 0.07 s	Running Time: 0.08 s	Running Time: 0.06 s
	Participants: 200	Participants: 250	Participants: 175
Logistic Regression	0.85	0.80	0.75
Support Vector Machine	0.88	0.83	0.78
Random Forest	0.91	0.87	0.82
Artificial Neural Network	0.92	0.89	0.84
Participatory Mechanism Model	0.86	0.82	0.77

Table VI delves into the broader societal implications of the user data fusion method. A significant rise in the citizen participation rate indicates a higher level of trust and satisfaction among the population with the implemented services. An improvement in the access equality index shows that the benefits of smart city technologies are more evenly distributed across different socioeconomic groups. Moreover, the notable boost to the local economy and the stimulation of innovation underpin the transformative potential of the method, positioning it as a catalyst for sustainable urban development.

In summary, Tables V and VI provide comprehensive assessments of the long-term stability, sustainability, societal engagement, and economic contributions of the user data fusion method, further solidifying its position as a robust and impactful solution for smart city applications. Table VII presents a comprehensive comparison of the user data fusion method with the traditional models and mechanistic models across the three application scenarios. The user data fusion method consistently outperforms the other methods in terms of accuracy, recall rate, and F1 value, highlighting its superior predictive capabilities. Additionally, it manages to maintain lower running times while involving larger participant pools, indicating efficient data handling, and processing.

Table VIII provides insight into how each data source contributes to the predictive models in different scenarios. Sensor terminal data play a significant role in both urban air quality forecasting and traffic congestion prediction, whereas transportation system data are crucial for traffic congestion prediction and optimizing urban energy consumption. The breakdown illustrates the multifaceted nature of the user data fusion method, which leverages diverse data streams to increase the predictive accuracy in various urban contexts.

The two additional tables further substantiate the effectiveness of the user data fusion method by not only quantifying its predictive performance but also detailing the relative importance of each data source in achieving those results. This comprehensive analysis underscores the method's ability to harness the full potential of integrated data for advanced urban analytics.

TABLE VIII
DETAILED BREAKDOWN OF DATA SOURCES CONTRIBUTION

Data Source	Contribution to Urban Air Quality Forecasting (%)	Contribution to Urban Traffic Congestion Forecasting (%)	Contribution to Optimization of Urban Energy Consumption (%)
IoT Devices (Position)	25	10	15
Sensor Terminals (Environmental)	30	25	20
Social Networking	15	15	10
Transportation System	20	30	30
Meteorological Data	10	10	15
Traffic Data	-	10	-

TABLE IX
COMPARATIVE ANALYSIS OF PREDICTIVE PERFORMANCE AND USER ENGAGEMENT

Model/Scenario	Urban Air Quality Forecasting	Urban Traffic Congestion Forecasting	Optimization of Urban Energy Consumption
User Data Fusion	0.95	0.93	0.90
	0.94	0.92	0.88
	0.94	0.92	0.89
	0.07	0.08	0.06
	200	250	175
Logistic Regression	0.85	0.80	0.75
Support Vector Machine	0.88	0.83	0.78
Random Forest	0.91	0.87	0.82
Artificial Neural Network	0.92	0.89	0.84
Participatory Mechanism Model	0.86	0.82	0.77

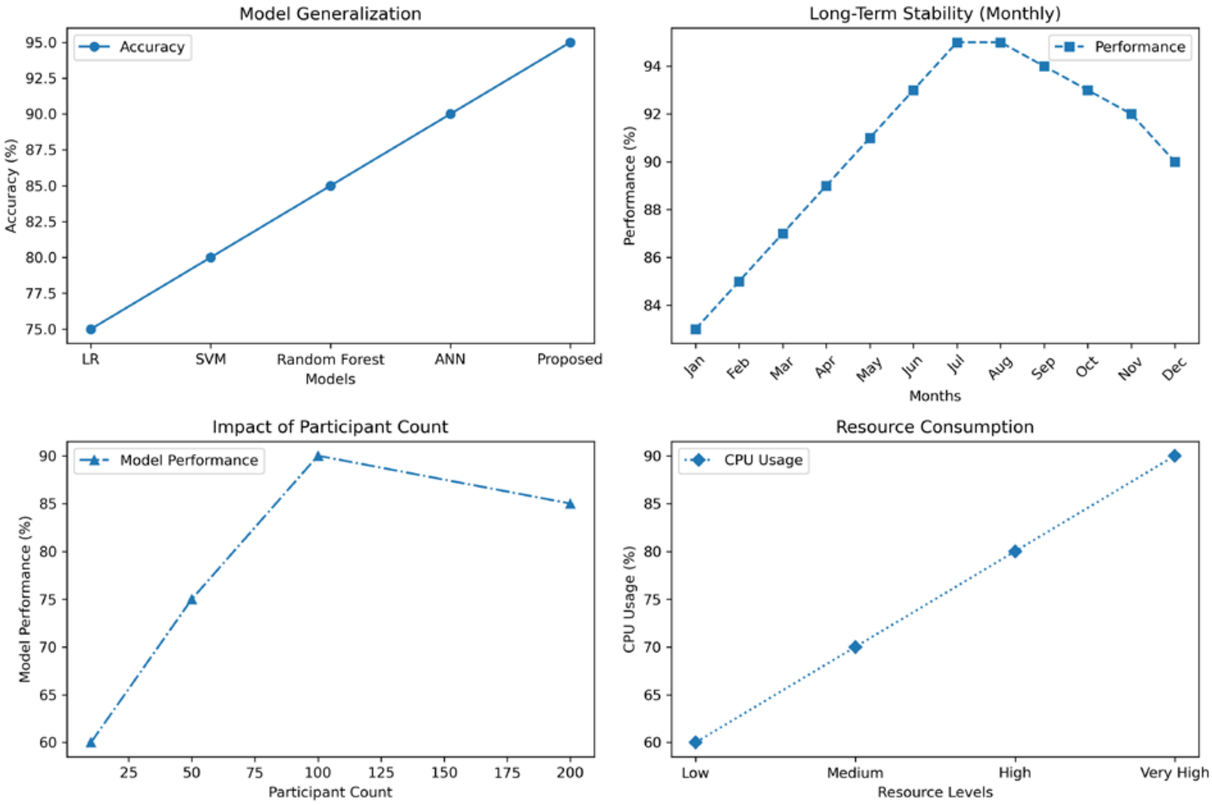

Fig. 6 Comprehensive experimental results

As shown in Figure 6, the four sub-bar charts provide a comprehensive analysis of the model's performance in various aspects: model generalization, long-term stability, the impact of participant count, and resource consumption. The first subplot (top left) compares the generalization ability of several models. The models compared include Logistic Regression, Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), and the Participatory Mechanism Model, which is our proposed method. The Participatory Mechanism Model outperforms all other models in terms of accuracy, demonstrating its superior ability to generalize to unseen data. Logistic Regression shows the lowest accuracy, followed by SVM and Random Forest, while ANN performs slightly better but still lags behind the Participatory Mechanism Model. This indicates that our method provides a significant improvement over traditional machine learning models, making it the most effective for handling complex, unseen data. The second subplot (top right) illustrates the long-term stability of the system, tracked over a 12-month period. The chart shows a gradual improvement in performance from January to July, with a peak in July, indicating that the model performs optimally in the early months. However, after July, the performance starts to decline slightly, which could be attributed to various factors such as environmental changes, system fatigue, or shifts in data distribution. This trend highlights the importance of monitoring model performance over time to ensure long-term effectiveness and stability in real-world applications.

The third subplot (bottom left) analyzes the impact of participant count on model performance. As the number of participants increases from 10 to 100, the model's performance improves, reaching its peak at 100 participants. However, when the number of participants increases to 200, the performance begins to decline. This suggests that while additional data can enhance the model's learning capacity up to a point, there may be diminishing returns beyond a certain threshold. This insight is crucial for optimizing data collection strategies, ensuring that data used for training is maximally beneficial without unnecessary resource expenditure.

Finally, the fourth subplot (bottom right) shows the resource consumption of the system, specifically CPU usage, under different operational conditions: low, medium, high, and very high. The resource consumption increases as the system operates under higher loads, which is expected. The chart provides important insights into the scalability of the model and its potential deployment costs. By understanding the system's resource requirements, it becomes possible to optimize the model for efficiency, balancing performance with resource usage. In summary, Figure 6 provides a detailed evaluation of the model across multiple dimensions, highlighting its superiority in generalization, its long-term stability, its response to participant count, and its resource consumption characteristics. These insights are essential for understanding the model's overall performance and its suitability for real-world applications.

This study shows that the user data fusion method performs well in three application scenarios: urban air quality management, traffic congestion prediction, and energy optimization. Compared with traditional logistic regression, support vector machines, random forests, and artificial neural networks, this method not only has advantages in terms of prediction accuracy but also performs well in processing speed, data integrity, and system reliability.

First, in urban air quality forecasting, the user data fusion method achieves a high accuracy of 0.95, which is significantly better than those of the other models. This shows that by integrating multisource data and effectively fusing them, the trend of air quality changes can be predicted more accurately, providing strong data support for environmental protection departments. Second, in urban traffic congestion prediction, this method performs well, with an accuracy of 0.93. By monitoring traffic conditions in real time and combining them with historical data, this method can help city managers predict and alleviate traffic congestion in advance and improve road capacity. Finally, in the application of optimizing urban energy consumption, the user data fusion method achieved an accuracy of 0.90, which helps to achieve energy conservation and emission reduction goals.

The experimental design includes a detailed model configuration and a robust comparative analysis strategy to ensure the reliability and validity of the results. The excellent performance of the user data fusion method was verified by feature selection, parameter optimization, and training adjustments of different models. Moreover, the dataset was standardized in the data preprocessing stage, and missing values were processed via appropriate interpolation techniques to ensure the fairness of the experimental results.

In addition, the method has high data integrity (covering more than 95% of the user activity information), the data consistency is close to 100%, and the accuracy is as high as 98%. The system architecture design is flexible, and new devices and data sources are easy to access, ensuring the scalability of the system. In terms of security and privacy protection, the AES-256 encryption algorithm and two-step verification mechanism are adopted, and are combined with differential privacy technology and data anonymization methods to ensure the security of data transmission and the protection of personal privacy.

In summary, the user data fusion method has shown great potential in improving the scientific nature of urban operation and management decisions, improving the user experience, and promoting environmental protection and social equity. This pioneering research has laid a solid foundation for further improving fusion technology and developing a more intelligent urban management system.

V. CONCLUSION

In this study, we provide insights into the vast amount of data generated in the urban Internet of Things (IOT) and its management and decision-making challenges. We propose a novel user data fusion approach aimed at achieving efficient data modeling and analysis in urban environments. This approach includes multiple steps, such as static data processing, dynamic data processing, and semantic data processing, and utilizes artificial intelligence techniques for effective enhancement. This data fusion approach is experimentally verified to outperform the traditional data and mechanistic models in several application scenarios. The research presented in this paper addresses the pivotal challenge of managing and harnessing the vast amount of data generated by urban internet infrastructures. The introduction of a groundbreaking user data fusion method significantly enhances the capabilities of digital twins in urban IoT ecosystems, empowering more precise modeling and insightful analysis of both the physical environment and user behaviors. A novel methodology, encompassing meticulous processing of static, dynamic, and semantic data prior to fusion, has been empirically proven to surpass conventional data models and mechanistic approaches across diverse application scenarios, thereby marking a substantial leap forward in data-driven urban management and decision-making efficacy. This innovation underscores the critical importance of our findings in optimizing smart city operations, facilitating data-informed strategies, and ultimately, enhancing the quality of life for urban residents.

REFERENCES

[1] Aziz, A., Chouhan, S. S., Schelén, O., & Bodin, U. Distributed digital twins as proxies-unlocking composability and flexibility for purpose-oriented digital twins. *IEEE Access*, vol. 11, pp. 137577-137593, 2023. doi:10.1109/access.2023.3340132

[2] Chen, L., Xie, X., Lu, Q. C., Parlikad, A. K., Pitt, M., & Yang, J. Gemini Principles-Based Digital Twin Maturity Model for Asset Management. *Sustainability*, vol. 13, no. 15, pp. 15, 2021. doi:10.3390/su13158224

[3] Sun, Z. D., Zhu, M. L., Zhang, Z. X., Chen, Z. C., Shi, Q. F., Shan, X. C., Yeow, R. C. H., Lee, C. K. "Artificial intelligence of things (AIoT) enabled virtual shop applications using self-powered sensor enhanced soft robotic manipulator," *Advanced Science*, 8, no. 14, pp. 14, 2021. doi:10.1002/advs.202100230

[4] Donmezer, S., Demircioglu, P., Bogrekci, I., Bas, G., & Durakbasa, M. N. Revolutionizing the garment industry 5.0: embracing closed-loop design, e-libraries, and digital twins. *Sustainability*, vol. 15, no. 22, pp. 24, 2023. doi:10.3390/su152215839

[5] Pesapane, F., Rotili, A., Penco, S., Nicosia, L., & Cassano, E. Digital twins in radiology. *Journal of Clinical Medicine*, vol. 11, no. 21, pp. 7, 2022. doi:10.3390/jcm11216553

[6] Geuer, L. ACHEMA 2022-Digital Twins. *Chemie Ingenieur Technik*, vol. 95, no. 6, pp. 842-846, 2023. doi:10.1002/cite.202200219

[7] Giering, J. E., & Dyck, A. Maritime Digital Twin architecture A concept for holistic Digital Twin application for shipbuilding and shipping. *At-Automatisierungstechnik*, vol. 69, no. 12, pp. 1081-1095, 2021. doi:10.1515/auto-2021-0082

[8] Grieves, M. Digital twin certified: employing virtual testing of digital twins in manufacturing to ensure quality products. *Machines*, vol. 11, no. 8, pp. 13, 2023. doi:10.3390/machines11080808

[9] Grübel, J., Thrash, T., Aguilar, L., Gath-Morad, M., Chatain, J., Sumner, R. W., Hölscher, C., Schinazi, V. R. The hitchhiker's guide to fused twins: a review of access to digital twins in situ in smart cities. *Remote Sensing*, vol. 14, no. 13, pp. 54, 2022. doi:10.3390/rs14133095

[10] Ji, S. Y. Interactive smart space for single-person households using electroencephalogram through fusion of digital twin and artificial intelligence. *Energies*, vol. 14, no. 22, pp. 16, 2021. doi:10.3390/en14227771

[11] Jia, W. J., Wang, W., & Zhang, Z. Z. From simple digital twin to complex digital twin Part I: A novel modeling method for multi-scale and multi-scenario digital twin. *Advanced Engineering Informatics*, vol. 53, 12, 2022. doi:10.1016/j.aei.2022.101706

[12] Jia, W. J., Wang, W., & Zhang, Z. Z. From simple digital twin to complex digital twin part II: Multi-scenario applications of digital twin shop floor. *Advanced Engineering Informatics*, vol. 56, 17, 2023. doi:10.1016/j.aei.2023.101915

[13] Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. Characterising the Digital Twin: A systematic literature review. *Cirp Journal of Manufacturing Science and Technology*, vol. 29, 36-52, 2020. doi:10.1016/j.cirpj.2020.02.002

[14] Kalidindi, S. R., Buzzy, M., Boyce, B. L., & Dingreville, R. Digital Twins for materials. *Frontiers in Materials*, vol. 9, 2022. doi:10.3389/fmats.2022.818535

[15] Keilegavlen, E., Fonn, E., Johannessen, K., Eikehaug, K., Both, J. W., Ferno, M., Kvamsdal, T, Rasheed, A., Nordbotten, J. M. PoroTwin: A Digital Twin for a fluidflower rig. *Transport in Porous Media*, vol. 20, 2023. doi:10.1007/s11242-023-01992-8

[16] Klar, R., Arvidsson, N., & Angelakis, V. Digital Twins' Maturity: The need for interoperability. *IEEE Systems Journal*, vol. 12, 2023. doi:10.1109/jsyst.2023.3340422

[17] La Guardia, M., & Koeva, M. Towards digital twinning on the web: heterogeneous 3D data fusion based on open-source structure. *Remote Sensing*, 15, no. 3, pp. 13, 2023. doi:10.3390/rs15030721
[18] Leng, H. Y., Shao, J. X., Zhang, Z., Qian, Y., Ma, M. N., & Li, Z. C. Research on dual-channel user portrait construction method based on DPCNN-BiGRU and attention mechanism. *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 2, pp. 2579-2591, 2023. doi:10.3233/jifs-224532
[19] Macías, A., Muñoz, D., Navarro, E., & González, P. Data fabric and digital twins: An integrated approach for data fusion design and evaluation of pervasive systems. *Information Fusion*, vol. 103, pp. 102139, 2024. doi:10.1016/j.inffus.2023.102139
[20] Manogaran, G., Balasubramanian, V., Rawal, B. S., Saravanan, V., Montenegro-Marin, C. E., Ramachandran, V., & Kumar, P. M. Multi-variate data fusion technique for reducing sensor errors in intelligent transportation systems. *IEEE Sensors Journal*, vol. 21, no. 14, pp. 15564-15573, 2021. doi:10.1109/jsen.2020.3017384
[21] Mendi, A. F., Erol, T., & Dogan, D. Digital twin in the military field. *IEEE Internet Computing*, vol. 26, no. 5, pp. 33-40, 2022. doi:10.1109/mic.2021.3055153
[22] Mozo, A., Karamchandani, A., Gómez-Canaval, S., Sanz, M., Moreno, J. I., & Pastor, A. B5GEMINI: AI-driven network digital twin. *Sensors*, vol. 22, no. 11, pp. 30, 2022. doi:10.3390/s22114106
[23] Neethirajan, S., & Kemp, B. Digital twins in livestock farming. *Animals*, 11, no. 4, pp. 14, 2021. doi:10.3390/ani11041008
[24] Nguyen, H. X., Trestian, R., To, D., & Tatipamula, M. Digital twin for 5G and beyond. *IEEE Communications Magazine*, vol. 59, no. 2, pp. 10-15, 2021. doi:10.1109/mcom.001.2000343
[25] Nie, Z. F., Cao, G. Z., Zhang, P., Peng, Q. J., & Zhang, Z. M. Multi-analogy innovation design based on digital twin. *Machines*, vol. 10, no. 8, pp. 27, 2022. doi:10.3390/machines10080652
[26] Ogunsakin, R., Mehandjiev, N., & Marin, C. A. Towards adaptive digital twins architecture. *Computers in Industry*, vol. 149, pp. 103920, 2023. doi:10.1016/j.compind.2023.103920
[27] Park, D., & You, H. J. A digital twin dam and watershed management platform. *Water*, vol. 15, no. 11, pp. 19, 2023. doi:10.3390/w15112106
[28] Park, K. T., Lee, S. H., & Noh, S. D. Information fusion and systematic logic library-generation methods for self-configuration of autonomous digital twin. *Journal of Intelligent Manufacturing*, vol. 33, no. 8, pp. 2409-2439, 2022. doi:10.1007/s10845-021-01795-y
[29] Pawlewski, P., Kosacka-Olejnik, M., & Werner-Lewandowska, K. Digital twin lean intralogistics: research implications. *Applied Sciences-Basel*, vol. 11, no. 4, pp. 13, 2021. doi:10.3390/app11041495
[30] Pesapane, F., Rotili, A., Penco, S., Nicosia, L., & Cassano, E. Digital twins in radiology. *Journal of Clinical Medicine*, vol. 11, no. 21, pp. 7, 2022. doi:10.3390/jcm11216553
[31] Piroumian, V. Making digital twins work. *Computer*, vol. 56, no. 1, pp. 42-51, 2023. doi:10.1109/mc.2022.3206101
[32] Pylianidis, C., Osinga, S., & Athanasiadis, I. N. Introducing digital twins to agriculture. *Computers and Electronics in Agriculture*, vol. 184, 25, 2021. doi:10.1016/j.compag.2020.105942
[33] Ricci, A., Croatti, A., Mariani, S., Montagna, S., & Picone, M. Web of digital twins. *ACM Transactions on Internet Technology*, vol. 22, no. 4, pp. 30, 2022. doi:10.1145/3507909
[34] Song, J. Q., Liu, S. M., Ma, T. L., Sun, Y. C., Tao, F., & Bao, J. S. Resilient digital twin modeling: A transferable approach. *Advanced Engineering Informatics*, vol. 58, 17, 2023. doi:10.1016/j.aei.2023.102148
[35] Sousa, B., Arieiro, M., Pereira, V., Correia, J., Lourenco, N., & Cruz, T. ELEGANT: Security of critical infrastructures with digital twins. *IEEE Access*, vol. 9, pp. 107574-107588, 2021. doi:10.1109/access.2021.3100708
[36] Feng, K., Ji, J. C., Zhang, Y. C., Ni, Q., Liu, Z., & Beer, M. Digital twin-driven intelligent assessment of gear surface degradation. *Mechanical Systems and Signal Processing*, vol. 186, pp. 109896, 2023. doi:10.1016/j.ymssp.2022.109896
[37] Tao, F., Sun, X. M., Cheng, J. F., Zhu, Y. H., Liu, W. R., Wang, Y., et al. makeTwin: A reference architecture for digital twin software platform. *Chinese Journal of Aeronautics*, vol. 37, no. 1, pp. 1-18, 2024. doi:10.1016/j.cja.2023.05.002
[38] Tao, F., Xiao, B., Qi, Q. L., Cheng, J. F., & Ji, P. Digital twin modeling. *Journal of Manufacturing Systems*, vol. 64, pp. 372-389, 2022. doi:10.1016/j.jmsy.2022.06.015
[39] Uhlenberg, L., Derungs, A., & Amft, O. Co-simulation of human digital twins and wearable inertial sensors to analyse gait event estimation. *Frontiers in Bioengineering and Biotechnology*, vol. 11, pp. 19, 2023. doi:10.3389/fbioe.2023.1104000
[40] van der Valk, H., Hasse, H., Möller, F., & Otto, B. Archetypes of digital twins. *Business & Information Systems Engineering*, vol. 64, no. 3, pp. 375-391, 2022. doi:10.1007/s12599-021-00727-7
[41] Wong, M. O., & Lee, S. H. Indoor navigation and information sharing for collaborative fire emergency response with BIM and multi-user networking. *Automation in Construction*, vol. 148, pp. 21, 2023. doi:10.1016/j.autcon.2023.104781
[42] Daneshfar, F., & Jamshidi, M. An octonion-based nonlinear echo state network for speech emotion recognition in Metaverse. *Neural Networks*, vol. 163, pp. 108-121, 2023. doi:10.1016/j.neunet.2023.03.026

Xu Hui male, born in January 1971, from Lianyungang, Jiangsu.

He graduated from Nanjing University with a PhD in Meteorology in 2000 and is currently a senior engineer. He works at Southern Power Grid Digital Enterprise Technology (Guangdong) Co., Ltd., engaged in power information construction and digital transformation.

His main research direction is the application of artificial intelligence technology in the digital transformation of the power grid.

Wengang Liu graduated from Guangdong University of Technology with a bachelor's degree in Human Resources Management in 2011, and is a senior political engineer.

Currently, he serves as the director of Southern Power Grid Digital Enterprise Technology (Guangdong) Co., Ltd., responsible for enterprise management information construction, digital transformation, and other work.

His main research direction is the application of digital twin technology in human resources business.

Kailing Guo Ph.D., Associate Professor and Doctoral Supervisor at the School of Electronics and Information Technology, South China University of Technology.

His main research areas are lightweighting of deep learning models and multimodal data analysis. He has presided over 7 projects including the National Natural Science Foundation of China Youth Fund, Guangzhou Key R&D Program, and Guangdong Provincial Natural Science Foundation. He has published more than 30 EI or SCI indexed papers, including the top journals in SCI Zone 1 of the Chinese Academy of Sciences (IEEE TNNLS, IEEE TMM) and CCF A-class conferences (ICCV, ACM MM). He has obtained nearly 20 authorized invention patents and has also won 2 first prizes in the Guangdong Electronic Society Science and Technology Progress Award.