

# Spatial-temporal data imputation for predictive modeling in intelligent transportation systems

Yohanes Pracoyo Widi Prasetyo<sup>1</sup>, Linawati<sup>2</sup>, Dewa Made Wiharta<sup>2</sup>, Nyoman Putra Sastra<sup>2</sup>

<sup>1</sup>Faculty of Engineering, Universitas 17 Agustus 1945, Banyuwangi, Indonesia

<sup>2</sup>Faculty of Engineering, Udayana University, Bali, Indonesia

---

## Article Info

### Article history:

Received May 21, 2024

Revised Oct 24, 2024

Accepted Nov 10, 2024

---

### Keywords:

Data imputation

Intelligent transportation system

Missing data

Predictive modeling

Spatial-temporal

---

## ABSTRACT

Data imputation is necessary to overcome data loss in intelligent transportation systems (ITS) due to the many sensors used to monitor traffic conditions. Sensor malfunction, hardware limitations, and technical glitches can lead to incomplete data, potentially leading to errors in traffic data analysis. This analysis investigated spatial-temporal data imputation approaches applied for predictive modeling in ITS. Each approach's strengths, weaknesses, and applicability in the context of ITS are evaluated. We analyzed various imputation approaches involving statistical, machine learning, and combined methods. Statistical methods are more straightforward but could effectively handle modern traffic's complexity. On the other hand, machine learning and combined approaches, such as hybrid convolutional neural network (CNN)- long short-term memory (LSTM), offer more robust capabilities in capturing non-linear patterns present in spatio-temporal data. This research aims to investigate the effectiveness of each approach in overcoming data incompleteness and the accuracy of predicting future traffic conditions with the widespread adoption of IoT, electric vehicles, and autonomous vehicles. The results of this investigation provide an understanding of the most suitable approaches to address the challenges of spatio-temporal data imputation and provide practical guidance for predictive modeling in ITS.

This is an open access article under the [CC BY-SA](#) license.



---

### Corresponding Author:

Yohanes Pracoyo Widi Prasetyo

Faculty of Engineering, Universitas 17 Agustus 1945

Banyuwangi, Indonesia

Email: widiprasetyo@untag-banyuwangi.ac.id

---

## 1. INTRODUCTION

Lost data is a common problem in ITS due to the increasing number of vehicles. Improvements to the ITS framework are needed as sensor failures and hardware limitations result in incomplete traffic data. The future development of autonomous vehicles, poor traffic management in developing countries resulting in congestion, and the drive towards smart, green, and sustainable cities make predictive modeling in ITS very important. Historically missing data was addressed by historical and statistical methods [1]. However, these methods are less effective due to the large data size and unnatural patterns caused by deleted data. Therefore, researchers began to investigate better imputation methods for predictive modeling in ITS. From previous research, the development of data imputation algorithms used interpolation, extrapolation, or model-based prediction techniques. By addressing these issues, the quality of traffic data and the system will work more efficiently [2].

The interpolation method can be used effectively [3]. However, they cannot capture complex spatial or temporal relationships in the data, hence the need to combine them with other methods. Proposing an

imputation model by combining a self-attention mechanism, automatic encoder, and generative adversarial network into a self-attention generative adversarial imputation net (SA-GAIN) [4]. However, the training of GAN is notoriously complicated on balance between generator and discriminator, this technique can capture spatio-temporal dependencies and correlations in the data, which is very useful in ITS. Meanwhile, [5] proposed the spatio-temporal generative adversarial network (STGAN) method with the concept of minimizing the reconstruction error from missing data entries and ensuring that the data entries fit the local spatio-temporal distribution, but STGAN requires large and high quality data because poor data leads to inaccurate output. Hence, it needs attention in training and interpretation. To reconstruct spatio-temporal states based on GANs, [6] proposed the traffic state reconstruction GAN (TSR-GAN) model. However, like other GAN models, it faces the problem of instability during training, so it needs careful customization. Specifically, the traffic lane states are converted into traffic state diagrams (TSDs), whose colors represent the values of traffic variables (e.g., speed or density). To reduce instability but with better accuracy, [7] proposed spatio-temporal learnable bidirectional attention generative adversarial networks (ST-LBAGAN) that use a deep learning framework. Although it has excellent potential, this method risks overfitting, so it needs a proper approach. Sometimes, data is lost on a large scale because some roads need sensors. To overcome this problem, Wang *et al.* [8] proposed an integrated deep learning for traffic state reconstruction (IDL-TSR) framework, using convolutional neural networks (CNNs) to capture spatial features and long short-term memory (LSTM) to capture temporal features to reconstruct traffic state using sensor data from limited links.

Real-time traffic prediction modeling [9] proposed the dynamic temporal adjacent graph convolutional network (D-TAGCN), a deep learning model designed to analyze spatio-temporal data with a dynamic graph structure. Although D-TAGCN is robust in capturing dynamic spatio-temporal patterns, its application requires special attention to model design, data, and parameter tuning. Proposed a spatiotemporal generative adversarial imputation net (ST-GAIN) model that relies heavily on the quality of available data, as data containing many outliers can lead to poor imputation results [10]. Spatio-temporal attention-gated recurrent neural network (ST-AGRNN) is a deep learning model designed to process and analyze spatio-temporal data, i.e., spatial and temporal dimensions. While these models have the power to handle complex spatio-temporal data, their application requires special attention to data, hyperparameters, and model design [11]. The GANs family can face the vanishing gradient problem, where the discriminator becomes too good, so the generator cannot learn effectively. The choice of method for predictive modeling in ITS depends on the characteristics of the data and the specific needs [12]. For spatial-temporal data, hybrid methods such as CNN-LSTM are very effective as they combine the strengths of CNN and LSTM to handle spatial-temporal data in predictive modeling. Ensemble and deep learning can provide a competitive advantage in predicting complex and dynamic conditions in intelligent transportation systems [13].

This article is divided into five sections. The first section introduces the study of spatial-temporal data imputation for predictive modeling in ITS. The next section presents methods that are often used in data imputation. Section 3 discusses each method's strengths, weaknesses, applicability, and effectiveness. Section 4 presents the results of the investigation and discussion of these methods. Finally, the paper concludes with conclusions about the best method for each and future research directions.

## 2. THE COMPREHENSIVE THEORETICAL BASIS

Studies using historical traffic data incorporating 3D convolutional generative networks and GANs to account for missing traffic data are [14]. In contrast to most studies that account for missing data at the granularity of road segments and combined time intervals, the imputation approach based on gated attentional generative adversarial networks (GaGANs) is highly responsive to dynamic traffic environments on signalized road networks. However, it requires high computational resources and longer training times [15]. Although many spatiotemporal approaches have been presented to overcome the problem of missing spatiotemporal data, Wang *et al.* [16] stated that there are limitations to capturing spatiotemporal dependencies in spatiotemporal graphs, as most imputation methods do not consider the dynamic data hidden in graph nodes, so an attention based message passing and dynamic graph convolution network is proposed by considering the traffic patterns of neighboring nodes and temporal changes in the data.

The grid division method [17] is an approach to missing data imputation in traffic passenger flow by dividing a geographical area into grids or boxes and analyzing the temporal dynamics in each grid. However, too large grids do not capture local variations, while tiny grids are too sensitive to noise. A deep learning model designed to handle tasks involving spatiotemporal data, such as missing data imputation or prediction in datasets that have both time and space dimensions, is the spatiotemporal feature-enhanced generative adversarial network (ST-FVGAN) [18]. However, this model is sensitive to noise in spatial and temporal data, which causes less accurate predictions or generates unrealistic synthetic data. Due to complex spatiotemporal relationships [11], [19] proposed a traffic data completion model based on a graph

convolutional network model to account for missing values from a deep learning perspective. This model uses graph convolution to model local spatial dependencies, combining self-attention, graph convolution, and residual network mechanisms to cope with complex graph data. Using multi-view passenger flow (MVPF), the OD matrix-based prediction method is a two-component method that uses multi-view data to predict congestion and optimize transportation routes [20]. Tensor decomposition-based methods are the most popular for data imputation, followed by GANs and GNN, which rely on extensive training data sets. Using AI and deep learning models for data imputation offers flexibility and the ability to capture complex patterns by combining multiple data sources [21]. Research methods related to filling in missing data and comparing them on the California performance measurement system (PeMS) need to design research that includes important aspects such as representative methods, assumptions, imputation styles, implementation conditions, limitations, and the use of public datasets [22]. However, there is a limitation in that the imputation effectiveness is only based on the PeMS dataset and may not reflect the complexity of the actual data.

The iterative generative adversarial networks for imputation (IGANI) method proposed by [23] is a variant of GANs designed to perform missing data imputation iteratively. This method utilizes the strength of GANs in generating realistic data and combines it with an iterative approach to improve the quality and accuracy of missing data imputation. However, the challenge is that this approach becomes very complex and requires significant computational resources. Deep convolutional generative adversarial networks (DCGANs) are an exciting approach to imputing missing data, especially traffic time series data [24]. Traffic data from PeMS is converted into images, and each specific time window (e.g., 24 hours) is converted into an image representation, where traffic values at a specific time are represented as pixels in the image [25]. However, this training still requires significant computation. Adding a multimodal deep learning model for heterogeneous traffic data imputation using two parallel stacked autoencoders is an innovative step and can consider spatial and temporal dependencies simultaneously, it assesses whether the model can be generalized to other traffic datasets beyond PeMS [26]. Proposed organizing lane-scale traffic data into tensor patterns that can simultaneously consider the spatio-temporal dependence of traffic flow with an improved tucker decomposition-based imputation (ITDI) method to recover the missing values from traffic data by extending the Tucker decomposition model with an adaptive rank calculation algorithm and an improved objective function [26]. Proposed an attentive graph neural process (AGNP) method for short-term traffic speed prediction and imputation at the network level while considering reliability first, a gaussian process (GP) is used to model the observed traffic speed state [27].

As a tool to account for missing traffic data [28], designed a novel deep learning architecture called dynamic graph convolutional recurrent imputation network (DGCRIN). DGCRIN uses graph generators and dynamic graph convolutional gated recurrent units (DGCGRU) to perform detailed modeling of the dynamic spatiotemporal dependencies of road networks. An innovative model that combines graph attention networks (GATs) and recurrent neural networks (RNNs) to impute missing traffic data by considering spatial and temporal dependencies in a bidirectional manner by [29] called bidirectional graph attention recurrent neural network (GARNN) which needs development for other datasets. It can be an innovative solution to address the missing data problem in traffic data for the imputation of multistate time series data [30]. Proposing multistate time series imputation using a generative adversarial network operates by using a generator and discriminator. The generator aims to generate an imputation of missing values in the time series data, while the discriminator learns to discriminate between those generated by the generator. The interaction of these two components results in statistically sound imputations consistent with the underlying pattern of the time series. Using a latent factor model-based approach for imputing traffic data with road network information efficiently fills data gaps while considering the road network structure [31]. The model incorporates latent factors to capture complex patterns in traffic data and road network information to improve imputation accuracy.

In model fusion, the hybrid CNN-LSTM ensemble method is an approach that combines CNNs and LSTMs in an ensemble model. This method is designed to handle spatial-temporal data effectively, making it particularly suitable for predictive modeling applications in ITS [32]. Utilizing the strengths of CNN in capturing spatial information and LSTM in capturing temporal patterns, it is well suited for complex spatial-temporal data. Using an ensemble, the model is more resilient to noise and missing data and can produce more accurate predictions.

## 2.1. Data collection

We provide an overview of investigations into traffic data collection methods, definitions of missing data types, and data preprocessing methods that can be used to improve limited datasets. In Figure 1, spatial-temporal data collection through datasets includes identifying IoT sensor data sources, historical data, and geospatial data [33]. Determining random missing data, missing data within a specific time range, and missing data blocks. The second part is the data imputation method used, including the statistical method of spatial interpolation by estimating missing values based on a statistical model of the distance between data

points, machine learning based on GANs used to generate synthetic data that resembles the original data, and the fusion method refers to the fusion of various data sources, techniques, or models to achieve more accurate and reliable results. The third part refers to the process of selecting the model that best suits the purpose of the analysis and the characteristics of the data. Involving the selection of features, algorithms, evaluation methods, and ways of combining models across road network types and types, data loss refers to situations where data that should be available is lost, corrupted, or inaccessible. Fuzzy models uncertain spatial-temporal variables, such as travel time or road conditions. Finally, we review future research challenges related to spatial-temporal data limitations, historical data, and data quality in the dataset. Challenges in system workload as well as the ability of the system to recover from disruptions with cloud services for elastic scalability.

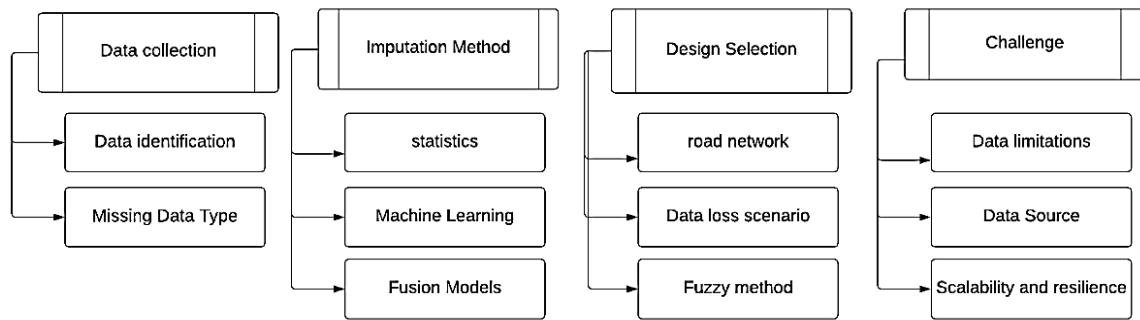


Figure 1. Flowchart of research review steps

**2.1.1. Data identification**

Traffic data identification refers to collecting and analyzing information related to the flow of vehicles or road users at a specific location. Standard methods for identifying traffic data based on online technologies such as GPS location, modeling, and video analysis image processing are discussed, as well as the complexity of the data source [34]. Using cameras to record images and identify vehicles is suggested by [35] for image processing and video analysis in detecting movements, vehicle types, and traffic patterns. Camera sensors are also frequently used in advanced traffic monitoring systems [36].

Using Google Maps, Bing Maps, Waze, Navigation Pro, and Tom Tom Traffic platforms [37] and [38] provide flexibility in obtaining traffic data using real-time methods. However, the complexity of the traffic data requires improved simulation models that can utilize the technology so that real-time traffic data can be used as a reference for traffic speed services [39]. Like PeMS public data, the transportation department can provide an easily accessible, internet-available source of real-time historical traffic data containing various analysis capabilities to support various users [40].

**2.1.2. Missing data**

Research on data imputation has different classifications for the type of missing data [41] describing random, univariate, and multivariate missing data. In other papers, such as [42] and [43], univariate and multivariate will be named fiber missing data and block or panel missing data; other papers may also give different names to similar types of missing data, such as continuous missing data to represent fiber missing data [44]. We illustrate the categorization of missing data in Figure 2.

Figure 2(a) shows that random missing data can occur due to sudden power disconnection of sensors, failed data transmission due to network interference, and random errors in survey data collection and GPS devices. Other variants of random missing data are missing at random (MAR) and missing not at random (MNAR) but [45], [46] state that MNAR is generally not considered. Therefore, MCAR is the standard test case used in most studies, followed by missing layer and block. The missing data layer in Figure 2(b), in the context of spatio-temporal data or ITS data, refers to a specific time segment, geographic location, or a specific category of data, e.g., vehicle type or weather conditions. The missing data block in Figure 2(c) often occurs in the context of time series data or spatio-temporal data, where an entire range of time, geographic location, or other variables are successively missing.

**2.2. Data preprocessing**

Data preprocessing identifies and handles missing data, validating the missing values with appropriate estimates. Proposed a data denoising and compression method based on wavelet transform and data model construction [47]. Said that the process of removing unreasonable outliers should adapt to the

general pattern of the data [48]. That is, if the dataset is too large, then preprocessing using data samples to make it easier to handle, then checking the quality of the data and ensuring that the data meets the specified criteria before use for further analysis.

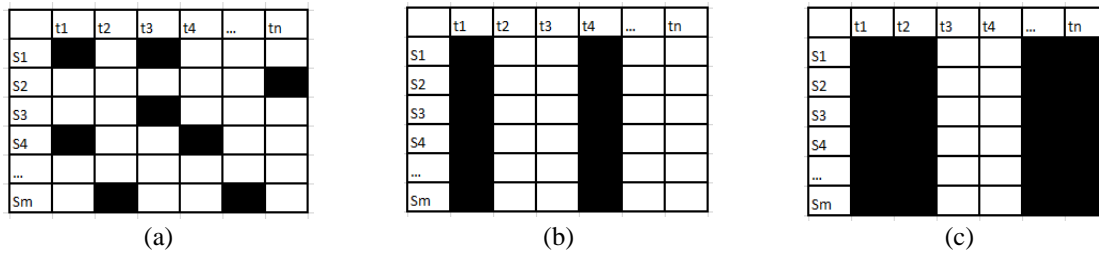


Figure 2. Illustration of missing data where black cells are missing data: (a) random missing data, (b) missing data layers, and (c) missing data blocks

### 3. INVESTIGATION METHOD

Literature reviews on the calculation of missing traffic data often focus on the result of the method used, such as [7]-[10], but may require further exploration in a more specific context, such as the road network or the type of missing data. There are three categories of imputation methods for missing traffic data: statistical, machine learning, and fusion model. Statistical imputation to estimate missing values uses statistical mean, median, and regression models. Machine learning methods involve using machine learning algorithms to predict and fill in missing values in the data set, and fusion methods are statistical, machine learning, and deep learning-based approaches. The hybrid CNN-LSTM ensemble method is a method that combines the advantages of CNN and LSTM with information from available data to perform imputation and predictive modeling accurately.

#### 3.1. Statistical methods

Statistical methods analyze existing data to develop representative models and are independent of the amount of data [49]. As the most popular and easy-to-process method, it replaces missing values with the mean, median, or mode of all data in the column. Probability principal component analysis (PPCA) is a statistical analysis technique used to reduce the dimensionality of complex data sets. This technique is similar to the principal component analysis (PCA) method but uses a probabilistic approach to determine the principal components of the data. PPCA generally works by finding a probabilistic representation of the data generated by a linear combination of multiple principal components. Has favorably reviewed a spatiotemporal PPCA-based data imputation method for traffic flow data in urban networks [50]. To overcome the shortcomings, a new method was proposed to improve the imputation performance of missing traffic data by fully utilizing the available spatial-temporal correlation data; fuzzy means (FCM) was selected as the basic algorithm [51].

##### 3.1.1. Spatial-temporal interpolation

Spatial-temporal interpolation is a technique used in geospatial data analysis to estimate values at unobserved locations and times based on observations of surrounding data [52]. As part of the data interpolation technique, the Kriging method fully utilizes the spatiotemporal correlation in traffic data and does not assume that the data follows a distribution [53]. The performance of the proposed method is compared with two other popular methods, namely historical averaging and KNN. The results show that the proposed method has the highest imputation accuracy and is more flexible than other methods. When the missing data rate is lower than 1%, the performance of the historical average method is better than the proposed imputation method [54]. Although Kriging is a powerful and popular interpolation method, it has some drawbacks related to the stationarity assumption, dependence on variogram models, sufficient data, and sensitivity to outliers.

##### 3.1.2. Tensor decomposition and factorization method

This method belongs to multivariate statistics and multidimensional data analysis, which involves using tensor structures to fill in missing values in multidimensional data. The technique involves breaking down the tensor structure into more structured components to model the complex patterns contained in the data [21], [55]. The advantage of data imputation with this method is the ability to handle multidimensional

data with complex patterns and interactions between dimensions through a Bayesian approach combined with imputation techniques to predict missing values in the data [56]. This approach uses a robust probabilistic approach to estimate missing values by considering the uncertainty in the prediction. Basic tensor factorization methods have shown significant improvement in the field of missing traffic data imputation, as stated by [3], [10], [13], [18], [20], [25], [56]-[58]. It can also be seen that most of these models have been tested for robustness in accounting for missing traffic data at levels ranging from 1% to 90% while still having a high level of accuracy [57], [59].

This method's advantages lie in its simplicity and applicability, ease of interpretation, computational efficiency, and suitability for structured data. Its weaknesses are that it is limited to linear data, less flexible, and less adaptive. Since it works well on linear data patterns, it is more effective when working with structured data.

### 3.2. Machine learning

Machine learning algorithms are designed to learn from existing data, identify patterns, and adapt to environmental changes or new data [60]. Models are trained using datasets containing predefined input and output pairs in supervised learning. In contrast, the model is given data with no labels or categories in unsupervised learning. The goal is to discover natural patterns in the data, such as clusters or hidden structures. Neural networks are the model most synonymous with machine learning, although not completely [61], it is a powerful tool for data imputation, including spatial-temporal data imputation. In this context, neural networks can predict missing values based on patterns in the available data. Some approaches to using neural networks in data imputation are GANs.

GANs are a type of neural network architecture that consists of two neural network models: generators and discriminators. The generator is responsible for creating new data, for example, images, sounds, or texts, similar to the training data. At first, the generator generates random data. During training, the generator learns to generate data that is increasingly similar to the training data through feedback from the discriminator [62]. The discriminator is responsible for distinguishing between the data generated by the generator and the original training data. The discriminator is trained to distinguish between "real" data (from the training dataset) and "fake" data (generated by the generator). The discriminator updates the parameters based on the error in classifying the fake or original data [22], [24]. The GAN training process involves iterations where the generator and discriminator play against each other.

A neural network graph is a visual representation of the architecture and structure of a neural network used in machine learning. It shows how neurons are organized in layers and connected through weighted connections. Recently, [63] has conducted a comprehensive survey on GNNs and classified various GNN models into four categories-recurrent GNNs, convolutional GNNs, graph auto-encoders, and spatio-temporal GNNs. Among them, it has been found that convolutional GNN has recently become a more popular choice in traffic research, as shown by [5], [15], [36]. Graph convolutional networks (GCNs) are neural network architectures designed to perform learning on data structured as graphs or networks. They extend the convolution concept of conventional neural networks to the graph domain, enabling the use of topological information in data representation. Convolutional GNNs, or GCNs, utilize convolutional neural networks to embed graph information into tensors, resulting in a uniform framework from irregular data sets [64]. GNNs can learn complex representations of spatio-temporal structures in graphs and extract patterns, relationships, and dependencies between graph entities [65].

Machine learning methods can handle complex data such as spatial-temporal, adaptive to extensive data, and more accurate prediction results. Machine learning models such as neural networks or random forests can capture complex data patterns and work with various data types, including those with missing values. Although they have drawbacks, such as large data requirements and the risk of overfitting, approaches such as regularization and careful feature selection can help overcome them. In ITS, this method works well when the data is dynamic and thus can predict traffic, detect anomalies, signal optimization, and travel time estimation with precision.

### 3.3. Fusion model

Model fusion for data imputation refers to using different techniques and models to fill in missing values in a data set. This approach exploits the strengths of each model to improve the quality of imputation and reduce the weaknesses of a single imputation method. The ensemble approach combines predictions from several different imputation models [66]. For example, bagging, boosting, or stacking techniques can integrate the results from multiple imputation models and produce more accurate predictions. Combination models combine imputation results from linear and non-linear models (e.g., random forest or neural network) to obtain better results. The fusion method with hybrid CNN-LSTM ensemble in the context of imputation of spatio-temporal data involves combining two types of models to utilize the strengths of each [67]. Fusion with multiple imputation methods is an approach that involves using several different imputation methods,

such as KNN, regression, and interpolation, and then combining the imputation results from the various methods to produce more stable predictions [68].

It utilizes the advantages of both statistical and machine learning methods to improve predictive accuracy and has the flexibility and adaptiveness to overcome overfitting. Although it has the disadvantage of high complexity and requires extensive computational resources, it works well on complex and heterogeneous data processing, making it a flexible solution for dynamic data patterns. Effectiveness for heterogeneous and balanced data on various data variation models, this fusion method is the most effective.

### 3.4. Overview of research methods

To provide more specific information, the following is a detailed description of the primary methods, example reference papers, road network types, data acquisition methods, and types of missing data tested in the context of spatial-temporal data imputation research in intelligent transportation in Table 1.

From Table 1, most of the literature reviewed was conducted on urban networks. This is because urban networks are the most fluctuating and busiest and thus require the support of intelligent transportation systems. However, if we look deeper into the data sets used, it can be seen that most of them are the same. Most studies considering traffic flow shown in Table 1 use taxi GPS data, which may not have accurate traffic speeds. Traffic volume shows fewer missing data imputation studies, likely due to data availability and the inaccurate nature of traffic volume. However, traffic volume also provides a good picture of traffic conditions, travel time, and congestion levels; these are parameters to consider. Imputation methods should consider both spatial and temporal aspects to produce accurate estimates, as transportation data are often derived from sensors and direct observations to overcome uncertainties and disturbances that may occur in observational data. Table 2 lists the characteristic advantages and gaps of the popular methods mentioned in Table 1.

Table 1. Summary of literature studies on variable data imputation

Methods	Article	Road network	Data loss	Fuzzy method	Data limitations	Data source
Statistics	[49], [50]-[54]	2	3	0	0	0
Spatial-Temporal						
Interpolation	[52]-[54]	2	1	0	0	0
Tensor	[3], [10] [13], [18], [20], [21], [20],	9	5	0	0	0
Decomposition	[25], [55], [56]-[59]					
Machine Learning	[5], [15], [23], [24], [36], [61]-[65], [60]	1	9	0	1	0
Fusion Model	[28], [32], [66], [67], [68]	2	3	0	0	0

Table 2. Characteristics of popular methods

Methods	State of the art	Research gaps
Statistics	– Ability to handle time and space variability supported by accuracy in estimating missing values and flexibility in models and approaches.	– Handling extensive data and accommodating uncertainty in estimation are limitations, as spatial-temporal data tends to be full of uncertainty.
Tensor decomposition	– It handles complex and multidimensional data well and has hidden pattern recognition. – It accommodates uncertainty in data and estimation and is more computationally efficient than other computationally intensive approaches.	– Limitations in handling noise and uncertainty in the data. – It requires large resources for implementation in real-time or big data environments.
Machine learning	– Able to handle large volumes of data well. – Able to handle data with high dimensionality and complex structures, such as spatial-temporal data, with sound capabilities. – Enables data-driven model training that can improve imputation performance.	– Requires a deep understanding of the concepts and algorithms involved – The performance of machine learning models is highly dependent on the data quality used for training. – Machine learning models could be more effective when dealing with imbalanced data. – Overfitting can lead to inaccurate or unreliable imputation results.
Fusion models	– We are utilizing each model's advantages and managing each model's shortcomings. – It has improved performance and flexibility in resource utilization. – Resilient to changes in data and environment and increased robustness for reduced overfitting.	– Requires more excellent computing resources. – This complexity can complicate the interpretation of results and increase the computational and management costs of the model. – Requires more complex customization and maintenance than single models.



#### 4. RESULTS AND DISCUSSION

This paper investigates the impact of spatial-temporal data loss on ITS. Although previous papers have explored model imputation and predictive modeling, they did not examine the common mechanisms used among the various models reviewed and focused more on the overall quality of each model. Spatio-temporal attention networks (STAN) incorporate attention mechanisms to capture the complex relationships between spatial and temporal data [69]. In addition, a spatial-temporal fusion layer combines spatial and temporal feature representations and an encoder-decoder architecture that produces the desired output. STAN can take into account the interactions between sensor locations and time, resulting in feature representations for more accurate predictions [70]. The discussion offers a hybrid CNN-LSTM ensemble method that combines the strengths of CNN and LSTM to model the complexity of spatial-temporal data. Initialization parameters on the number of CNN layers, LSTM units, ensemble size, and method are determined appropriately, and then data preprocessing is performed to handle missing values during training. The CNN captures the spatial patterns, while the LSTM handles the temporal dependencies. Then, the CNN-LSTM model is combined, the final model is used to impute the missing values, and the performance is evaluated using standard metrics. We find that the prediction accuracy is correlated with the complexity of the model and the number of spatio-temporal features used. The method proposed in this paper has a much higher proportion of accuracy under dynamic traffic conditions than traditional regression-based or interpolation-based imputation methods.

##### 4.1. Handling lost data

Fusion models utilize the advantages of each technique to overcome the limitations of individual methods [71]. CNNs capture spatial patterns in data, such as images or maps, with a two-dimensional structure. To generate a feature map, CNNs use convolution and pooling operations to hierarchically extract spatial features by applying filters or small kernels to the input data (e.g., images or maps). Here is the equation of the convolution operation.

$$X_{i,j} = \sum_{m=1}^M \sum_{n=1}^N I_{i+m-1,j+n-1} \cdot W_{m,n} + b \quad (1)$$

$X_{i,j}$ : output feature map;  $I$ : input data;  $W$ : filter;  $b$ : bias;  $M$  and  $N$  are the filter sizes. This operation allows CNNs to recognize basic features such as edges, corners, or textures across the input data, which are then combined in the next layer to form a more complex representation.

Multiple CNN-LSTM models are trained independently with different parameter variations or data subsets, and all models' predictions are combined. For example, predictions are combined by taking the average of all predictions (bagging) or weighting the predictions based on model performance (boosting). For each missing value, the predictions from all CNN-LSTM models in the ensemble are combined to produce the final estimate, and the missing value is replaced with the predicted result from the ensemble.

##### 4.2. Feature extraction

Spatial feature extraction CNNs perform spatial data processing, such as images or maps, where CNNs automatically learn to detect essential patterns in the data. Spatial data such as images, maps, or road network grids are represented as two-dimensional or three-dimensional matrices, such as channels for RGB images. Each element in this matrix represents certain information, such as pixel intensity in the image or a specific value in the grid. LSTM temporal feature extraction is used to capture patterns in sequential data, such as time sequences, which are essential in many applications, such as time series prediction, weather analysis, and intelligent transportation systems. The input of LSTM is usually a sequence of data with specific dimensions as in (2):

$$X = \{x_1, x_2, x_3, \dots, x_T\} \quad (2)$$

Where  $T$  is the length of the time sequence.

The LSTM mechanism at the forget gate regulates how much information from the previous time step  $h_{t-1}$  will be forgotten by the memory cell  $C_t$ .

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

The input gate contains how much new information  $C_t$  will be added to the memory cell.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ C_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \end{aligned} \quad (4)$$



To update the cell state, the memory cell is updated by combining the old information filtered by the forget gate and the new information selected by the input gate.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad (5)$$

The output gate determines the current output based on the updated memory cell information.

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \cdot \tanh(C_t) \end{aligned} \quad (6)$$

Where  $x_t$  is the input at time  $t$ ;  $h_{t-1}$  is the hidden state from the previous time;  $C_t$  is the cell state at time  $t$ ;  $W_f, W_i, W_c, W_o$  are the learned weights;  $b_f, b_i, b_c, b_o$  is the bias;  $\sigma$  is the sigmoid function, and  $\tanh$  is the activity function  $\tanh$ .

### 4.3. Fuzzy method

Fuzzy theory allows modeling uncertainty with fuzzy rules that can flexibly represent knowledge, such as general traffic patterns, road user habits, or vehicle movement patterns [72]. For incomplete data due to damaged sensors, outages, or system failures, fuzzy methods can provide a more accurate approach than deterministic ones [73] uses geographic information for spatial data through movement patterns, traffic distribution, and temporal data to capture trends and time patterns affecting traffic conditions. The fuzzy-spatial-temporal model is used to develop a hybrid model that combines fuzzy theory with spatial-temporal analysis methods for data imputation. The model can dynamically adjust to traffic conditions and other external factors [74]. Despite its minimal usefulness, the author argues that this method is worth mentioning because traffic data tends to be imprecise due to many external variables, and fuzzy theory can improve the performance of other models in a hybrid setting, as shown by the above research.

### 4.4. Challenge

In Figure 1, road and environmental conditions related to data availability and infrastructure are the main objects of data availability in modern transportation. The existing transportation infrastructure may be inadequate to handle the growing volume of vehicles, leading to congestion and accidents, which are still significant problems in modern transportation systems.

#### 4.4.1. Data limitations

Datasets collected from ITS systems are only sometimes complete due to sensor limitations, signal interference, or technical errors that affect data quality [75]. Data that is not real-time has implications for delays in data processing and collection, resulting in slow response to rapidly changing traffic conditions. The Hybrid CNN-LSTM Ensemble method is an approach that combines CNN and LSTM to handle spatial-temporal data effectively. When there is limited data available, this method can maximize information utilization by leveraging the strengths of each model. CNNs in this method can utilize transfer learning by using pre-trained models trained on similar large datasets to strengthen spatial features. At the same time, LSTMs can utilize transfer learning if pre-trained models are available for the relevant temporal data type. Ensemble learning combines multiple independently trained CNN-LSTM models with different initializations or different subsets of data. This approach improves the model's reliability and accuracy by reducing bias and variance, especially under data limitations.

#### 4.4.2. Data source

ITS relies on various data sources to optimize transportation systems, and the sensors used can vary in type and scope [76]. Some sensors only measure traffic data on roads, while others include data from public transportation or other modes of transportation. Data quality limitations, such as irregularities and noise, can complicate the imputation process and make the imputation results less accurate [77], [78]. Urban areas may have denser sensor networks compared to rural areas. This may lead to an imbalance in the availability of spatial-temporal data. Data obtained from sensors may have privacy and ownership restrictions that affect its accessibility and use for imputation purposes [79]. The limitations of the format and structure of data obtained from various sensors can be different formats and structures [80]. Large volumes of spatio-temporal data require large storage and processing capacities, and these limitations can hinder the ability to store and process data efficiently [81]. According to [82], [83], data security limitations and transportation data are sensitive to privacy security, so they must be protected from unauthorized access. The

hybrid CNN-LSTM ensemble method can be adapted to handle different data in form, type, and source. This approach involves the combination of CNN and LSTM in an ensemble framework to maximize model performance when working with different types of data, such as spatial, temporal, or a combination of both. In this hybrid architecture, CNN extracts spatial features from different data. After spatial feature extraction, LSTM is used to extract temporal features from the data that CNN has processed. LSTM captures the relationship between time and emerging spatial patterns. Ensemble learning combines predictions from different CNN-LSTM models that may be trained independently on different data types. Techniques such as voting, averaging, or stacking can be applied to produce more accurate final predictions.

**4.4.3. Scalability and resilience**

Scalability refers to the system's ability to handle increasing volumes of data without experiencing performance degradation, given the growth in the number of users, data traffic, and service requests [84]. To achieve scalability, the ITS system architecture should be designed considering distributed architecture, cloud computing technology, and sufficient network capacity [85]. Early detection technologies and management systems are important to help respond quickly to disruptions or incidents in the transportation system [86]. Integrating new technologies, such as IoT, big data analytics, and AI, can help improve the scalability and resilience of ITS systems by enabling real-time monitoring, predictive analytics, and more efficient centralized management [87], [88]. Hybrid CNN-LSTM can be implemented using parallel processing. CNN and LSTM can run independently on different data before combining their outputs in the ensemble stage. This method can efficiently process large volumes of data by utilizing parallel computing, thus improving scalability. In this scenario, extensive data can be divided into several parts and processed and distributed across multiple nodes, significantly improving the model's ability to handle large-scale data.

Robustness is improved through ensemble learning, where multiple CNN-LSTM models are trained independently, and the results are combined. By combining predictions from multiple models, ensemble learning reduces the risk of model failure caused by data outliers or noise. Techniques such as iterative imputation to fill in missing data or spatial/temporal filters to remove noise can be applied, ensuring that the model remains accurate despite problems with the data. Using incremental learning or fine-tuning approaches to adjust the model based on the latest data ensures that predictions remain accurate despite changing environmental conditions. Table 3 shows the results of the discussion of the various methods used.

Table 3. Results of discussion on method characteristics

Methods	Strength	Weakness	Application	Effectiveness
Statistics	– Simple and easy to implement, efficient on computing resources, interpretability, and structured data.	– Limited to linear relationships, flexible, less adaptive.	– Works well when the data pattern is simple and linear while predicting in the short-term.	– Effective only on simple variables and structured linear data.
Machine learning	– Able to handle complex and dynamic data, more adaptive to changes in data patterns, effective with large-scale data, and more accurate prediction results.	– It requires extensive and high-quality data, is challenging to interpret, requires high computational resources, and is prone to overfitting.	– Works well on dynamic, large, and complex data and for long-term predictions. – Prediction accuracy in ITS is stronger	– Effectively handles complex and dynamic data. – Suitable for problems that require accurate predictions in dynamic and non-linear data
Fusion models	– Utilize statistical and machine learning methods, improve accuracy, be flexible and adaptive, and avoid overfitting.	– High complexity, requires large computing resources, and difficulty in tuning.	– Works well on complex and heterogeneous data. – Prediction accuracy and model robustness across different types of dynamic data.	– Most effective for heterogeneous data. – It has a balance on various data models and is highly accurate in prediction.

Our study shows that higher model complexity is not associated with poor performance. When applied to large traffic datasets, the proposed method can benefit from increased features and spatiotemporal variables without adversely affecting computational efficiency or prediction accuracy. This study explores a comprehensive spatio-temporal data imputation approach using various methods, including machine learning and combined techniques. However, further in-depth studies may be needed to confirm this method's robustness and generalizability, especially regarding the influence of traffic environment variations and lower sensor data quality on prediction performance.

Our research shows that CNN-LSTM-based imputation methods are more robust than traditional interpolation methods in the face of dynamic spatiotemporal data incompleteness. Future research can

explore feasible ways to incorporate other machine learning techniques, such as GANs or attention-based models, to produce more accurate and efficient traffic predictions in complex future traffic scenarios.

## 5. CONCLUSIONS

Recent observations suggest that changes influence the phenomenon in ITS in spatial-temporal dynamics and fluctuating traffic conditions. Our findings provide conclusive evidence that the phenomenon is related to changes in vehicle movement patterns, irregular traffic density, and travel time distribution rather than due to an increase in the number of vehicles alone. Of the various methods investigated, if the priority is simple data and easy to implement, statistical methods are suitable; when having extensive data and complex non-linear relationships, machine learning methods will be more effective. If the main goal is prediction accuracy and the ability to handle heterogeneous data flexibly, then model fusion is the best choice. Fusion methods with hybrid approaches, such as CNN-LSTM ensemble, provide alternative solutions in ITS spatial-temporal data imputation and prediction accuracy. CNN handles spatial features well, while LSTM handles temporal dependencies, allowing the model to utilize both types of information synergistically. Fusion models often yield the best ITS performance because they integrate multiple information sources and data types. Future research should focus on developing transportation solutions that are adaptive to changes in technology and data, as well as the challenges of real-time data, privacy, autonomous vehicles, and climate change.

## REFERENCES

- [1] P. L. Wu, M. Ding, and Y. B. Zheng, "Spatiotemporal traffic data imputation by synergizing low tensor ring rank and nonlocal subspace regularization," *IET Intelligent Transport Systems*, vol. 17, no. 9, pp. 1908–1923, 2023, doi: 10.1049/itr2.12383.
- [2] X. Wang, Y. Ma, S. Huang, and Y. Xu, "Data imputation for detected traffic volume of freeway using regression of multilayer perceptron," *Journal of Advanced Transportation*, vol. 2022, 2022, doi: 10.1155/2022/4840021.
- [3] X. Su, M. Fan, Z. Cai, Q. Liu, and X. Zhang, "A light weight traffic volume prediction approach based on finite traffic volume data," *Journal of Systems Science and Systems Engineering*, vol. 32, no. 5, pp. 603–622, 2023, doi: 10.1007/s11518-023-5572-x.
- [4] W. Zhang, P. Zhang, Y. Yu, X. Li, S. A. Biancardo, and J. Zhang, "Missing data repairs for traffic flow with self-attention generative adversarial imputation net," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7919–7930, 2022, doi: 10.1109/TITS.2021.3074564.
- [5] Y. Yuan, Y. Zhang, B. Wang, Y. Peng, Y. Hu, and B. Yin, "STGAN: Spatio-temporal generative adversarial network for traffic data imputation," *IEEE Transactions on Big Data*, vol. 9, no. 1, pp. 200–211, 2023, doi: 10.1109/TBDATA.2022.3154097.
- [6] K. Zhang, X. Feng, N. Jia, L. Zhao, and Z. He, "TSR-GAN: generative adversarial networks for traffic state reconstruction with time space diagrams," *Physica A: Statistical Mechanics and its Applications*, vol. 591, 2022, doi: 10.1016/j.physa.2021.126788.
- [7] B. Yang, Y. Kang, Y. Y. Yuan, X. Huang, and H. Li, "ST-LBAGAN: Spatio-temporal learnable bidirectional attention generative adversarial networks for missing traffic data imputation," *Knowledge-Based Systems*, vol. 215, 2021, doi: 10.1016/j.knosys.2020.106705.
- [8] N. Wang, K. Zhang, L. Zheng, J. Lee, and S. Li, "Network-wide traffic state reconstruction: An integrated generative adversarial network framework with structural deep network embedding," *Chaos, Solitons and Fractals*, vol. 174, 2023, doi: 10.1016/j.chaos.2023.113830.
- [9] Q. Zheng and Y. Zhang, "DSTAGCN: Dynamic spatial-temporal adjacent graph convolutional network for traffic forecasting," *IEEE Transactions on Big Data*, vol. 9, no. 1, pp. 241–253, 2023, doi: 10.1109/TBDATA.2022.3156366.
- [10] L. Cai, C. Sha, J. He, and S. Yao, "Spatial-temporal data imputation model of traffic passenger flow based on grid division," *ISPRS International Journal of Geo-Information*, vol. 12, no. 1, 2023, doi: 10.3390/ijgi12010013.
- [11] J. Yang, J. Li, L. Wei, L. Gao, and F. Mao, "ST-AGRNN: A spatio-temporal attention-gated recurrent neural network for traffic state forecasting," *Journal of Advanced Transportation*, vol. 2022, 2022, doi: 10.1155/2022/2806183.
- [12] V. T. Ha and C. T. Thuy, "Model predictive control combined reinforcement learning for automatic vehicles applied in intelligent transportation system," *Telecommunication Computing Electronics and Control (TELKOMNIKA)*, vol. 22, no. 2, pp. 302–310, 2024, doi: 10.12928/TELKOMNIKA.v22i2.25274.
- [13] H. Song and H. Choi, "Forecasting stock market indices using the recurrent neural network based hybrid models: CNN-LSTM, GRU-CNN, and ensemble models," *Applied Sciences (Switzerland)*, vol. 13, no. 7, 2023, doi: 10.3390/app13074644.
- [14] Z. Li, H. Zheng, and X. Feng, "3D convolutional generative adversarial networks for missing traffic data completion," 2018, doi: 10.1109/WCSP.2018.8555917.
- [15] T. Zhang, J. Wang, and J. Liu, "A gated generative adversarial imputation approach for signalized road networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12144–12160, 2022, doi: 10.1109/TITS.2021.3110268.
- [16] Y. Wang *et al.*, "Attention-based message passing and dynamic graph convolution for spatiotemporal data imputation," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-34077-z.
- [17] L. Cai, C. Sha, J. He, and S. Yao, "Spatial-temporal data imputation model of traffic passenger flow based on grid division," *ISPRS International Journal of Geo-Information*, vol. 12, no. 1, 2023, doi: 10.3390/ijgi12010013.
- [18] B. Yang, Y. Kang, Y. Yuan, H. Li, and F. Wang, "ST-FVGAN: filling series traffic missing values with generative adversarial network," *Transportation Letters*, vol. 14, no. 4, pp. 407–415, 2022, doi: 10.1080/19427867.2021.1879624.
- [19] Y. Zhang, X. Wei, X. Zhang, Y. Hu, and B. Yin, "Self-attention graph convolution residual network for traffic data completion," *IEEE Transactions on Big Data*, vol. 9, no. 2, pp. 528–541, 2023, doi: 10.1109/TBDATA.2022.3181068.
- [20] F. Zheng, J. Zhao, J. Ye, X. Gao, K. Ye, and C. Xu, "Metro OD matrix prediction based on multi-view passenger flow evolution trend modeling," *IEEE Transactions on Big Data*, vol. 9, no. 3, pp. 991–1003, 2023, doi: 10.1109/TBDATA.2022.3229836.
- [21] R. K. C. Chan, J. M. Y. Lim, and R. Parthiban, "Missing traffic data imputation for artificial intelligence in intelligent transportation systems: review of methods, limitations, and challenges," *IEEE Access*, vol. 11, pp. 34080–34093, 2023, doi: 10.1109/ACCESS.2023.3264216.




- [22] T. Sun, S. Zhu, R. Hao, B. Sun, and J. Xie, "Traffic missing data imputation: a selective overview of temporal theories and algorithms," *Mathematics*, vol. 10, no. 14, 2022, doi: 10.3390/math10142544.
- [23] A. Kazemi and H. Meidani, "IGANI: Iterative generative adversarial networks for imputation with application to traffic data," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3103456.
- [24] T. Huang, P. Chakraborty, and A. Sharma, "Deep convolutional generative adversarial networks for traffic data imputation encoding time series as images," *International Journal of Transportation Science and Technology*, vol. 12, no. 1, pp. 1–18, 2023, doi: 10.1016/j.ijtst.2021.10.007.
- [25] L. Li, B. Du, Y. Wang, L. Qin, and H. Tan, "Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model," *Knowledge-Based Systems*, vol. 194, 2020, doi: 10.1016/j.knosys.2020.105592.
- [26] W. Lu, T. Zhou, L. Li, Y. Gu, Y. Rui, and B. Ran, "An improved tucker decomposition-based imputation method for recovering lane-level missing values in traffic data," *IET Intelligent Transport Systems*, vol. 16, no. 3, pp. 363–379, 2022, doi: 10.1049/itr2.12148.
- [27] M. Xu, Y. Di, H. Ding, Z. Zhu, X. Chen, and H. Yang, "AGNP: Network-wide short-term probabilistic traffic speed prediction and imputation," *Communications in Transportation Research*, vol. 3, 2023, doi: 10.1016/j.commtr.2023.100099.
- [28] X. Kong, W. Zhou, G. Shen, W. Zhang, N. Liu, and Y. Yang, "Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data," *Knowledge-Based Systems*, vol. 261, 2023, doi: 10.1016/j.knosys.2022.110188.
- [29] G. Shen, W. Zhou, W. Zhang, N. Liu, Z. Liu, and X. Kong, "Bidirectional spatial-temporal traffic data imputation via graph attention recurrent neural network," *Neurocomputing*, vol. 531, pp. 151–162, 2023, doi: 10.1016/j.neucom.2023.02.017.
- [30] H. Li, Q. Cao, Q. Bai, Z. Li, and H. Hu, "Multistate time series imputation using generative adversarial network with applications to traffic data," *Neural Computing and Applications*, vol. 35, no. 9, pp. 6545–6567, 2023, doi: 10.1007/s00521-022-07961-4.
- [31] X. Su, W. Sun, C. Song, Z. Cai, and L. Guo, "A latent-factor-model-based approach for traffic data imputation with road network information," *ISPRS International Journal of Geo-Information*, vol. 12, no. 9, 2023, doi: 10.3390/ijgi12090378.
- [32] H. M. Rai and K. Chatterjee, "Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data," *Applied Intelligence*, vol. 52, no. 5, pp. 5366–5384, 2022, doi: 10.1007/s10489-021-02696-6.
- [33] H. Varun Chand and J. Karthikeyan, "Survey on the role of IoT in intelligent transportation system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 3, pp. 936–941, 2018, doi: 10.11591/ijeecs.v11.i3.pp936-941.
- [34] A. Chabchoub, A. Hamouda, S. Al-Ahmadi, and A. Cherif, "Intelligent traffic light controller using fuzzy logic and image processing," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, pp. 396–399, 2021, doi: 10.14569/IJACSA.2021.0120450.
- [35] T. Dudek and A. Kujawski, "The concept of big data management with various transportation systems sources as a key role in smart cities development," *Energies*, vol. 15, no. 24, Dec. 2022, doi: 10.3390/en15249506.
- [36] H. T. Sadeeq, T. H. Hameed, A. S. Abdi, and A. N. Abdulfatah, "Image compression using neural networks: a review," *International journal of online and biomedical engineering*, vol. 17, no. 14, pp. 135–153, 2021, doi: 10.3991/IJOE.V17I14.26059.
- [37] Google Developers, "Google maps platform documentation," *Google Developers*, 2021. Access date: 25 January 2024, [Online] Available: <https://developers.google.com/maps/documentation>.
- [38] "Location technology for developers," *tomtom*, 2024, Access date: 25 January 2024, [Online] Available: <https://developer.tomtom.com/>.
- [39] V. Verendel and S. Yeh, "Measuring traffic in cities through a large-scale online platform," *Journal of Big Data Analytics in Transportation*, vol. 1, no. 2–3, pp. 161–173, 2019, doi: 10.1007/s42421-019-00007-7.
- [40] California Department of Transportation, "Performance measurement system (PeMS) data source," *Caltrans*, 2023.
- [41] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11651–11667, 2019, doi: 10.1109/ACCESS.2019.2891360.
- [42] X. Wu, M. Xu, J. Fang, and X. Wu, "A multi-attention tensor completion network for spatiotemporal traffic data imputation," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 20203–20213, 2022, doi: 10.1109/IJOT.2022.3171780.
- [43] Y. Liang, Z. Zhao, and L. Sun, "Memory-augmented dynamic graph convolution networks for traffic data imputation with diverse missing patterns," *Transportation Research Part C: Emerging Technologies*, vol. 143, 2022, doi: 10.1016/j.trc.2022.103826.
- [44] J. Tang, X. Zhang, W. Yin, Y. Zou, and Y. Wang, "Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 25, no. 5, pp. 439–454, 2021, doi: 10.1080/15472450.2020.1713772.
- [45] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values," *Transportation Research Part C: Emerging Technologies*, vol. 118, 2020, doi: 10.1016/j.trc.2020.102674.
- [46] J. Du, M. Hu, and W. Zhang, "Missing data problem in the monitoring system: a review," *IEEE Sensors Journal*, vol. 20, no. 23, pp. 13984–13998, 2020, doi: 10.1109/JSEN.2020.3009265.
- [47] H. Dou and G. Wang, "Data denoising and compression of intelligent transportation system based on two-dimensional discrete wavelet transform," *International Journal of Communication Systems*, vol. 34, no. 10, 2021, doi: 10.1002/dac.4809.
- [48] H. Xu, L. Zhang, P. Li, and F. Zhu, "Outlier detection algorithm based on k-nearest neighbors-local outlier factor," *Journal of Algorithms and Computational Technology*, vol. 16, 2022, doi: 10.1177/17483026221078111.
- [49] N. Karmitsa, S. Taheri, A. Bagirov, and P. Makinen, "Missing value imputation via clusterwise linear regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1889–1901, 2022, doi: 10.1109/TKDE.2020.3001694.
- [50] E. Joelianto, M. F. Fathurrahman, H. Y. Sutarto, I. Semajnski, A. Putri, and S. Gautama, "Analysis of spatiotemporal data imputation methods for traffic flow data in urban networks," *ISPRS International Journal of Geo-Information*, vol. 11, no. 5, 2022, doi: 10.3390/ijgi11050310.
- [51] Q. Shang, Z. Yang, S. Gao, and D. Tan, "An imputation method for missing traffic data based on FCM optimized by PSO-SVR," *Journal of Advanced Transportation*, vol. 2018, 2018, doi: 10.1155/2018/2935248.
- [52] M. L. Xu, T. Xing, and M. Han, "Spatial-temporal data interpolation based on spatial-temporal kriging method," *Zidonghua Xuebao/Acta Automatica Sinica*, vol. 46, no. 8, pp. 1681–1688, 2020, doi: 10.16383/j.aas.2018.c170525.
- [53] Z. Yao, T. Zhang, L. Wu, X. Wang, and J. Huang, "Physics-informed deep learning for reconstruction of spatial missing climate information in the antarctic," *Atmosphere*, vol. 14, no. 4, 2023, doi: 10.3390/atmos14040658.
- [54] H. Yang *et al.*, "A Kriging based spatiotemporal approach for traffic volume data imputation," *PLoS ONE*, vol. 13, no. 4, 2018, doi: 10.1371/journal.pone.0195957.
- [55] A. Baggag *et al.*, "Learning spatiotemporal latent factors of traffic via regularized tensor factorization: imputing missing values and forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2573–2587, 2021, doi: 10.1109/TKDE.2019.2954868.

- [56] J. Xing, R. Liu, K. Anish, and Z. Liu, "A customized data fusion tensor approach for interval-wise missing network volume imputation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12107–12122, 2023, doi: 10.1109/TITS.2023.3289193.
- [57] K. Lin, H. Zheng, X. Feng, and Z. Chen, "A novel spatial-temporal regularized tensor completion algorithm for traffic data imputation," 2018, doi: 10.1109/WCSP.2018.8555921.
- [58] H. Xie, Y. Gong, and X. Dong, "Spatial-temporal regularized tensor decomposition method for traffic speed data imputation," *International Journal of Data Science and Analytics*, vol. 17, no. 2, pp. 203–223, 2024, doi: 10.1007/s41060-023-00412-w.
- [59] Z. Zhao, L. Tang, M. Fang, X. Yang, C. Li, and Q. Li, "Toward urban traffic scenarios and more: a spatio-temporal analysis empowered low-rank tensor completion method for data imputation," *International Journal of Geographical Information Science*, vol. 37, no. 9, pp. 1936–1969, 2023, doi: 10.1080/13658816.2023.2234434.
- [60] D. Abdulla, G. Ramu, and N. Mamatha, "A survey on citywide traffic estimation techniques," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, Aug. 2018, pp. 3313–3318, doi: 10.1109/ICECDS.2017.8390072.
- [61] A. Navarro-Espinoza *et al.*, "Traffic flow prediction for smart traffic lights using machine learning algorithms," *Technologies*, vol. 10, no. 1, 2022, doi: 10.3390/technologies10010005.
- [62] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, "Deep learning versus conventional methods for missing data imputation: A review and comparative study," *Expert Systems with Applications*, vol. 227, 2023, doi: 10.1016/j.eswa.2023.120201.
- [63] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021, doi: 10.1109/TNNLS.2020.2978386.
- [64] X. Yao, Y. Gao, D. Zhu, E. Manley, J. Wang, and Y. Liu, "Spatial origin-destination flow imputation using graph convolutional networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7474–7484, 2021, doi: 10.1109/TITS.2020.3003310.
- [65] W. Liang *et al.*, "Spatial-temporal aware inductive graph neural network for C-ITS data recovery," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8431–8442, 2023, doi: 10.1109/TITS.2022.3156266.
- [66] B. W. McDonald, L. Hall, and X. P. Zhang, "Who is most likely to offend in my store now? statistical steps towards retail crime prevention with Auror," *ANZIAM Journal*, vol. 57, p. 289, 2017, doi: 10.21914/anziamj.v57i0.10507.
- [67] A. Almulih *et al.*, "Ensemble learning based on hybrid deep learning model for heart disease early prediction," *Diagnostics*, vol. 12, no. 12, 2022, doi: 10.3390/diagnostics12123215.
- [68] H. Dong and S. B. Tsai, "An empirical study on application of machine learning and neural network in English learning," *Mathematical Problems in Engineering*, vol. 2021, 2021, doi: 10.1155/2021/8444858.
- [69] S. Wang, W. Li, S. Hou, J. Guan, and J. Yao, "STA-GAN: A spatio-temporal attention generative adversarial network for missing value imputation in satellite data," *Remote Sensing*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010088.
- [70] L. Liu, F. Wang, H. Liu, S. Zhu, and Y. Wang, "HD-Net: A hybrid dynamic spatio-temporal network for traffic flow prediction," *IET Intelligent Transport Systems*, vol. 18, no. 4, pp. 672–690, 2024, doi: 10.1049/itr2.12462.
- [71] H. Ma, X. Qin, Y. Jia, and J. Zhou, "Dynamic spatio-temporal graph fusion convolutional network for urban traffic prediction," *Applied Sciences (Switzerland)*, vol. 13, no. 16, 2023, doi: 10.3390/app13169304.
- [72] Y. Sun, "A fuzzy multi-objective routing model for managing hazardous materials door-to-door transportation in the road-rail multimodal network with uncertain demand and improved service level," *IEEE Access*, vol. 8, pp. 172808–172828, 2020, doi: 10.1109/ACCESS.2020.3025315.
- [73] C. Luo, C. Tan, X. Wang, and Y. Zheng, "An evolving recurrent interval type-2 intuitionistic fuzzy neural network for online learning and time series prediction," *Applied Soft Computing Journal*, vol. 78, pp. 150–163, 2019, doi: 10.1016/j.asoc.2019.02.032.
- [74] D. Ma, B. Sheng, X. Ma, dan S. Jin, "Fuzzy hybrid framework with dynamic weights for short-term traffic flow prediction by mining spatio-temporal correlations," *IET Intelligent Transport System*, vol. 14, no. 2, 2020, doi: 10.1049/iet-its.2019.0287.
- [75] M. Xu, T. Liu, S. P. Zhong, and Y. Jiang, "Urban smart public transport studies: a review and prospect," *Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology*, vol. 22, no. 2, pp. 91–108, 2022, doi: 10.16097/j.cnki.1009-6744.2022.02.009.
- [76] I. Zrigui, S. Khouilji, A. Ennassiri, S. Bourekkadi, and M. L. Kerkeb, "Reducing carbon footprint with real-time transport planning and big data analytics," in *E3S Web of Conferences*, 2023, vol. 412, doi: 10.1051/e3sconf/202341201082.
- [77] R. Kumar, J. Mendes Moreira, and J. Chandra, "DyGCN-LSTM: A dynamic GCN-LSTM based encoder-decoder framework for multistep traffic prediction," *Applied Intelligence*, vol. 53, no. 21, pp. 25388–25411, 2023, doi: 10.1007/s10489-023-04871-3.
- [78] W. Fang, Y. Shao, P. E. D. Love, T. Hartmann, and W. Liu, "Detecting anomalies and de-noising monitoring data from sensors: A smart data approach," *Advanced Engineering Informatics*, vol. 55, 2023, doi: 10.1016/j.aei.2022.101870.
- [79] S. Kumar, S. C. Sharma, and R. Kumar, "Wireless sensor network based real-time pedestrian detection and classification for intelligent transportation system," *International Journal of Mathematical, Engineering and Management Sciences*, vol. 8, no. 2, pp. 194–212, 2023, doi: 10.33889/IJMEMS.2023.8.1.012.
- [80] O. Jakšić, Z. Jakšić, K. Guha, A. G. Silva, and N. M. Laskar, "Comparing artificial neural network algorithms for prediction of higher heating value for different types of biomass," *Soft Computing*, vol. 27, no. 9, pp. 5933–5950, 2023, doi: 10.1007/s00500-022-07641-4.
- [81] D. Tzika-Kostopoulou and E. Nathanail, "Exploring the big data usage in transport modelling," in *Advances in Intelligent Systems and Computing*, 2021, vol. 1278, pp. 1117–1126, doi: 10.1007/978-3-030-61075-3\_107.
- [82] Y. Lin and S. Geertman, "Can social media play a role in urban planning? A literature review," in *Lecture Notes in Geoinformation and Cartography*, 2019, pp. 69–84, doi: 10.1007/978-3-030-19424-6\_5.
- [83] T. Yuan, W. Da Rocha Neto, C. E. Rothenberg, K. Obraczka, C. Barakat, and T. Turletti, "Machine learning for next-generation intelligent transportation systems: A survey," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 4, 2022, doi: 10.1002/ett.4427.
- [84] C. Ma, M. Zhao, and Y. Zhao, "An overview of Hadoop applications in transportation big data," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 10, no. 5, pp. 900–917, 2023, doi: 10.1016/j.jtte.2023.05.003.
- [85] K. Yao and S. Chen, "Resilience-based adaptive traffic signal strategy against disruption at single intersection," *Journal of Transportation Engineering, Part A: Systems*, vol. 148, no. 5, 2022, doi: 10.1061/jtepbs.0000671.
- [86] T. Afrin and N. Yodo, "A survey of road traffic congestion measures towards a sustainable and resilient transportation system," *Sustainability (Switzerland)*, vol. 12, no. 11, 2020, doi: 10.3390/su12114660.
- [87] X. Xu, M. Lin, X. Luo, and Z. Xu, "HRST-LR: A hessian regularization spatio-temporal low rank algorithm for traffic data imputation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 10, 2023, doi: 10.1109/TITS.2023.3279321.




- [88] E. K. Xidias, I. E. Panagiotopoulos, and P. T. Zacharia, "An intelligent management system for relocating semi-autonomous shared vehicles," *Transportation Planning and Technology*, vol. 46, no. 1, pp. 93–118, 2023, doi: 10.1080/03081060.2022.2162052.

## BIOGRAPHIES OF AUTHORS






**Yohanes Pracoyo Widi Prasetyo**    received his M.T degree in engineering from Udayana University Bali, Indonesia. He is currently pursuing his Doctoral program in the Doctor of Engineering Science Program at Udayana University Bali, and currently serves as the vice dean of the Faculty of Engineering for academic affairs at 17 August 1945 Banyuwangi University, Indonesia. Has served as head of the Information Technology Center at the same campus. His research interests include big data analysis, traffic engineering, smart cities and intelligent transportation systems. He can be contacted at email: widiprasetyo@untag-banyuwangi.ac.id.






**Linawati**    Engineering Faculty, Udayana University, She got her Ph.D. and Master degree from School of Telecommunications and Electrical Engineering, The University of New South Wales, Sydney, Australia. Currently, she is a senior lecturer in Electrical Engineering Department, Faculty of Engineering, Udayana University. She has been a lecturer in Udayana University since 1991. Then she was appointed as a Postgraduate Program Director since September 2021 until May 2023. Now she is a Dean of Engineering Faculty and a secretary of IEEE Indonesia Section. During her professional work as an IEEE member, she has won many research fundings and has been published in International Conferences, and Journals. She chaired international conferences, many workshops, focus group discussions, seminars, and trainings which involved a wide range of professionals with many backgrounds. Her research interests are, but not limited to network performance, network security, applied technology, internet of things, and smart city. She can be contacted at email: linawati@unud.ac.id



**Dewa Made Wiharta**    has been a teaching staff at the Electrical Engineering Study Program since 1997. He completed his undergraduate study in 1996 at the Sepuluh Nopember Institute of Technology, Surabaya and obtained a master's degree from Gadjah Mada University, Yogyakarta in 2002. His doctorate degree was completed in 2015 at the Department of Electrical Engineering, Sepuluh Nopember Institute of Technology, Surabaya in the scientific field of Multimedia Telecommunications. Research interests include telecommunications in general, computer visuals, and robotics. Currently he is actively conducting research on image processing, IoT, and computer visuals for robotics together with several students. Dewa Made Wiharta was the coordinator of web development and information systems at Udayana University (2006-2008), and was an ICT consultant/expert for several agencies in Badung Regency and Denpasar City (2012-present). His current position is as head of the Udayana University Information Resources Unit, responsible for the ICT network and development of integrated information systems at Udayana University. He can be contacted at email: wiharta@unud.ac.id.



**Nyoman Putra Sastra**    received the B.Eng. and M.Eng. degrees in electrical engineering from Institut Teknologi Bandung (ITB), Indonesia, in 1998 and 2001, respectively. In 2015 he received his Ph.D. degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya. Since 2001, he joined Udayana University, Bali as a lecturer. His research interests include wireless multimedia sensor networks, IoT, unmanned vehicles, robotic, and multimedia signal processing. He is a Member of the IEEE. He can be contacted at email: putra.sastra@unud.ac.id.