IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 16 March 2023, accepted 30 March 2023, date of publication 3 April 2023, date of current version 7 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264216

SURVEY

Missing Traffic Data Imputation for Artificial Intelligence in Intelligent Transportation Systems: Review of Methods, Limitations, and Challenges

ROBIN KUOK CHEONG CHAN[1], **JOANNE MUN-YEE LIM**[1], **(Member, IEEE), AND RAJENDRAN PARTHIBAN**[2], **(Senior Member, IEEE)**

[1]School of Engineering, Monash University Malaysia, Bandar Sunway, Selangor 47500, Malaysia
[2]Department of Electrical and Computer Systems Engineering, Faculty of Engineering, Monash University, Clayton, VIC 3800, Australia

Corresponding author: Joanne Mun-Yee Lim (Joanne.Lim@monash.edu)

This work was supported by the Malaysian Ministry of Higher Education (MOHE), Fundamental Research Grant Scheme (FRGS), through the purview of Monash University Malaysia, under Grant FRGS/1/2019/TK08/MUSM/03/1.

ABSTRACT Missing data in Intelligent Transportation Systems (ITS) could lead to possible errors in the analyses of traffic data. Applying Artificial Intelligence (AI) in these circumstances can mitigate such problems. Past works focused only on specific data imputation methods, such as tensor factorization or a specific neural network model. While there are review papers covering singular topics regarding missing data, there are none in the field of traffic, to the best of our knowledge, that introduces the process of missing data collection and the viability of the traffic data collected while also broadly covering the popularly used models of recent years. This has led to non-uniformity of the terms used in missing data imputation, limited research in areas where datasets are not available, and a narrowed view of the methods used for data imputation. Hence, this paper aims to standardize the terms used in missing data classifications, look into the limitations of using available public or private datasets for urban traffic research, and discuss popular statistical and data-driven methods used by recent AI and ITS papers. It was found that tensor decomposition-based methods are the most popular for missing data imputation, followed by Generative Adversarial Networks and Graph Neural Networks, all of which rely on a large training dataset. Meanwhile, Probability Principle Component Analysis (PPCA) methods provide valuable insights via traffic analysis and are used for real-time traffic imputation. This paper also highlights the need for more efficient and reliable methods for traffic data collection, such as online APIs.

INDEX TERMS Intelligent transportation systems, artificial intelligence, communication system operations and management, reviews.

I. INTRODUCTION

Missing data is a prevalent problem in many fields of study, and Intelligent Transportation Systems (ITS) is one of them. As vehicles on the road continue to increase yearly, the importance of improving the existing ITS framework continues to grow as well. Hence, there has been much research in the field of traffic modeling, prediction, and routing, among others. All this research can be done thanks to the availability of traffic data or access to traffic data collection tools. However, these traffic data could be missing, possibly due to a sensor malfunction or connection errors between the sensor and the system. Hence, these missing data pose a major obstacle in the various traffic research as they would introduce errors or biases in the results if not handled appropriately.

Historically, such missing data are handled via historical averaging, deletion-based methods, and other relatively basic statistical methods [1]. However, these methods tend to result in other problems, such as incorrect data size or unnatural data patterns due to deleted data. Hence, researchers started to investigate missing traffic data imputation using better methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Jjun Cheng.

Over the years, there have been studies proposing various missing traffic data imputation methods, as shown by the many reviews from more than ten years ago [1] to even recent times [2], showing how crucial missing data imputation is to the future of a well-developed ITS. Recent reviews such as [3], [4], [5], and [6] tend to focus on a single aspect of missing traffic data imputation and the methods related to it, providing in-depth details in those areas, making them very suitable when trying to investigate the improvements made as well as to provide more detailed explanations regarding the reviewed methods alongside the authors' insight. However, focusing on a single aspect can lead to a lack of reviews on the other aspects of missing traffic data in the field of Intelligent Transportation Systems (ITS) besides the popular methods used in recent years, such as the limitations, possible challenges regarding data collection, and discussions related to parameters and statistical methods in which future researchers could use or investigate.

Besides that, while there are many traffic studies that have studied different kinds of missing data, the classifications of the types of missing data tend to be somewhat vague outside of random missing data, which by itself technically has three different classifications on its own. For example, the definition of block missing in [7] coincides with the definition of what is generally known as fiber missing. This paper intends to define and classify these different missing data types into three categories for the purpose of easing future traffic research.

Additionally, this paper aims to introduce the different data collection methods and their feasibility when researching a detailed urban network.

Finally, the paper reviews the few popular methods many researchers employ when dealing with missing traffic data to provide a general idea of the popularly selected model used as the base (e.g., Deep Neural Network, tensor decomposition, etc.) as well as investigate the other parameters applied to the research. Topics such as whether rural or urban road networks were used, the classification of the missing data the proposed research aimed to solve, as well as other possible limitations, are discussed in this paper.

To clarify, the objective of this paper is threefold: i) to provide a generalized classification for the different types of missing data, to allow for better identification, ii) to introduce the popular data collection method and their weaknesses when researching detailed urban road networks, while providing another avenue of data collection which is used less, iii) to review the popular missing data imputation methods, their common design choices, and their future potential.

The paper is organized as follows: Section II discusses the literature reviews and research gap. Section III discusses traffic data retrieval and the type of missing data faced by researchers. Section IV reviews the various popular methods from the statistical and data-driven models. Section V discusses the popular design choices used in conjunction with the base data imputation model. Section VI covers the challenges and limitations. Finally, Section VII concludes the paper, and Section 8 is the acknowledgment of contribution.

II. SIMILAR WORKS

Missing data as a whole has been studied extensively over the years, and it follows that there are reviews done with this in mind. There have been review works done in recent years that cover the topic of missing data imputation extensively. Reference [8] has reviewed missing data imputation techniques from 2006 to 2017, while [9] has reviewed techniques from 2010 to 2021. These two reviews have split the missing data imputation techniques into two types — statistical and machine learning-based methods — and looked into the distribution of studies done for each of the techniques and the evaluation methods considered. It is interesting to note that both [8] and [9] have classified missing data as only as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In the field of traffic, and likely for time series or spatial datasets, there are more than just these classifications of missing data types, as explained later on.

Other missing data review papers, such as [10], have made a comparison between different missing traffic data imputation methods, namely prediction, interpolation, and statistical learning methods, and concluded that the PPCA-based (Probability Principle Component Analysis) methods perform the best overall in terms of accuracy and computational complexity. In addition, [11] has compared the performance of variations of other existing statistical methods such as linear regressions, Predicting Mean Matching (PMM), and mean imputation, while also comparing regression tree-based methods such as Classification and Regression Trees (CART) and Random Forest. The conclusion is that the random forest implementation performed the best.

Looking into reviews done more specifically in the field of traffic, [2] provided a summary of the methods of traffic data collection, splitting them into fixed and mobile types, as well as explained the classification of various missing data types along with traffic imputation methods. Meanwhile, [5] reviewed temporal data imputation methods specifically, providing a more in-depth analysis of the state-of-the-art data imputation methods that utilized only the temporal aspect of traffic, covering their application conditions and limitations, as well as providing a list of popular public datasets [4] focused on traffic state estimation in urban road networks, of which there are missing data for segments due to the unavailability of traffic detectors due to installation costs as well as faulty detectors, with a focus on methods that fuses multiple sources of data into their models.

In these existing works, it is noted several times that while random missing data has been tested quite often, research that simulates non-random missing data due to situations such as faulty detectors is significantly less. Also, the authors would like to note that many public datasets, such as PeMS [12], are freeway traffic datasets, which do not equate to an urban traffic environment, as also mentioned by [5] and [4]. Certain

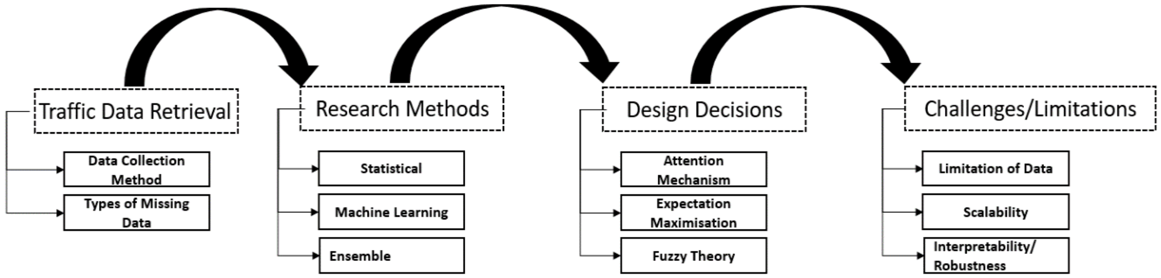

FIGURE 1. Flowchart on the steps taken in the review.

studies may make use of road segments in an urban environment [11], but a few individual roads are not representative of urban traffic as a whole. In reality, urban networks are more likely in need of such research, and this paper aims to review recent papers whose work covers urban networks.

Also, data acquisition can be difficult, depending on the country. As can be seen in later sections, many urban network datasets utilize taxi GPS datasets either from the public or private sector. It should be noted, however, that such methods may not be available for all countries and locations of interest, which would result in certain ambiguities when it comes to the viability of the various proposed methods in said locations.

FIGURE 1 summarises the flow of the following sections. This paper first discusses the various common traffic data collection and categorizes missing data types into three categories, namely random missing, fibers missing, and block missing, before going into commonly used data pre-processing methods. Secondly, the paper then looks into popular research methods, broadly categorized into statistical, machine learning, and ensemble methods. This paper reviews the type of road networks used in the various recent research and the types of missing data scenarios tested, as well as investigates the commonly used fundamental and auxiliary methods proposed. Notable design decisions are then mentioned in the following section. Finally, this paper also discusses the potential limitation of previously available datasets and how future researchers should investigate a more flexible yet easily accessible source of traffic data, along with emphasizing a focus on the scalability of models and their interpretability and robustness, besides purely accuracy-focused models.

III. TRAFFIC DATA RETRIEVAL

For any form of traffic management effort to succeed, the acquisition of traffic data is essential. Only by utilizing these data can the ITS process, learn, predict, and resolve the traffic issues it oversees. While some literature review includes reviews on public datasets [5], they focus on looking at the effect these datasets would have on the models rather than cover the different methods of traffic data retrieval. The aim of this section is to provide insight into the different methods of traffic data collection, as well as provide a definition of the type of missing data, as well as possible data pre-processing methods that could be used to augment a limited dataset. Besides that, Various factors need to be considered when handling traffic data: 3.1) Data collection method, 3.2) Types of missing data, and 3.3) Data Pre-processing

A. DATA COLLECTION METHOD

Acquisition of traffic data can be made via several methods. The most used methods would be through the access of public datasets or publicly available sensor data, such as those discussed below. Another method that is rarely seen being used in studies is the usage of online traffic API services, which is also discussed.

1) SENSORS AND CAMERAS

Sensors such as induction loop detectors were employed by ([13], [14], [15], [16], [17]) to collect real-time traffic data. The main reason for its frequent usage is that induction loop detectors perform well in vehicle counting in high and low-volume traffic under different weather conditions.

Besides that, studies such as [7], [18], [19], and [20] make use of street cameras and vehicle identification software in order to capture traffic data. Using cameras has the advantage of being able to analyze certain traffic parameters more accurately, such as the traffic flow, average gaps between vehicles during different traffic hours, as well as traffic accidents and other such events.

The drawback to such methods is that the user is limited to where the sensors and cameras are placed, making research into other areas or even larger urban networks not possible.

2) ONLINE SERVICES

An Application Programming Interface (API) is a software intermediary which enables the communication between two applications. Using APIs, an application can send a request to a server and receive a reply in the form of an output of the data interpreted by the corresponding server. By utilizing these applications, users can obtain information almost immediately. This is especially useful when an application requires real-time data, such as various GPS applications such as Google Maps or Waze. Examples of such services are Google Maps [21], Bing Maps [22], HERE Traffic API [23], and TomTom Traffic API [24]. Despite the flexibility of obtaining traffic data using these services, there is hardly any literature with regards to missing traffic data imputation that makes use of it. This could be due to the location of interest, having other available sources of data, or the difficulty of collecting data over a period using the API service. However, it should be emphasized that as traffic research grows in technology and knowledge, so should their simulations, and [25] has shown that online traffic data can be a good indicator of traffic speed.

3) PUBLIC DATASETS

Public data extensively employed in various literature includes PeMS, Department of Transportation, Induction Loop, Portal FHWA, AMAP, and KEEL [26]. PeMS, or the California Transportation Agency Performance Measurement System, emerged as the most widely employed public dataset in [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], and [39]. The main reason for its popularity is that PeMS can provide an easy-to-access source of historical and real-time traffic data, which is readily available over the internet, containing a series of built-in analytical capabilities to support various users. The source is available in [12].

Due to the readily available transportation statistics and real-time data over the Department of Transportation websites, datasets from the Department of Transportation were employed by in [40], [41], [42], [43], [44], and [45]. Other public datasets were also employed in recent literature, including Portal FHWA ([46], [47], [48]), AMAP ([49]), and KEEL ([29]). Portal FHWA and AMAP provide online access in China to collect real-time traffic data.

There has also been literature that compiled other available public datasets, such as [5], and has shown that these datasets tend towards freeways, highways, expressways, or limited signalized intersections. As shown here, there is a lack of public datasets with regard to urban traffic networks on a wider scale, as well as fewer public datasets outside of America and China, with few specific datasets in countries like Spain and England. This limits traffic studies for larger or more detailed urban networks and for areas located outside these few locations of interest.

B. MISSING DATA

The following subsections provide a standardized categorization of the types of missing data commonly experienced and how these missing data are manifested from the different types of data collection methods.

1) TYPES OF MISSING DATA

Existing missing data imputation researches have differing classifications for similar types of missing data. For example, [50] describes random missing data, along with two other types, namely univariate missing data and multivariate missing data. In other papers, such as [4], [51], and [52], univariate and multivariate would be named fiber and block or panel missing data, respectively. Other papers might have also given overlapping or different names for similar kinds of missing data, such as continuous missing data to represent fiber missing data [53].

For the sake of unification, these missing data types should be defined and generalized in order to help simplify the direction of future research. The general idea of the three categories is as mentioned below visualized in FIGURE 1:

Random Missing Data: Missing Data is caused by sporadic errors in the transmission of which there is little to no correlation known between the data loss and other variables. Results in missing data at random points in the dataset.

Fiber Missing Data: Missing data is caused by a sudden, temporary failure in connectivity or in the data-capturing device, resulting in long periods of missing data. Results in missing data for a length of time.

Block Missing Data: Missing data caused by the absence of a detector in the area of interest (i.e., A rarely used arterial road that does not justify the installation of a loop detector [4] or all sensors are not in operation for some reason). Results in complete missing data for the entire length of time over a long period or complete missing data from all sources of information for the same time horizon. This is seen in datasets with multiple sources of data.

While random missing data can be further broken down into three more types — Namely, Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) — simulations are usually done in an MCAR situation, such as [54]. Reference [55] has also stated that MNAR is generally not considered as well. Hence, for most research, MCAR is the general test case, followed by fiber and block missing.

It is important to note that block missing data imputation is not researched much, probably due to the significant lack of data as well as some research deeming that the areas with these levels of missing data do not contribute much to the overall traffic state.

2) MISSING DATA IN DATA COLLECTION METHODS

It can be said that all three types of missing data can occur for all the missing data collection methods mentioned above. However, the impact of such cases may differ depending on the method.

When it comes to monitoring systems such as sensors and cameras, the main reason tends to be equipment malfunction and electrical breakdowns, which leads to loss or damaged data [55]. Early detection leads to this being a case of fiber missing data and failure to do so causes it to devolve into block missing data. Public or private datasets which use similar monitoring systems would also be subjected to similar issues. However, as the data has already been collected in the past, it is trivial to ignore datasets with missing data and select the ones for which the dataset is complete. Online traffic APIs might face similar issues, but applications like HERE Traffic API [23] or Google Maps [21] would have more than one source of data to ensure the integrity of their data, such as floating car data (FCD) or probe vehicle data from a fleet of connected vehicles via GPS services or applications [25],

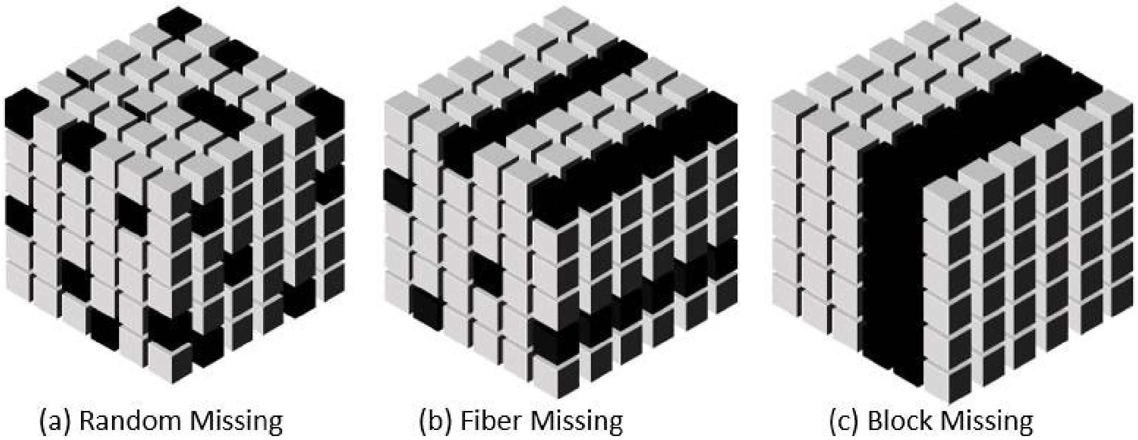

FIGURE 2. Visualization of different categories of missing data where the dark cubes represent the missing data. (a) Random missing data, (b) Fiber missing data, and (c) Block missing data.

although even then, there are times where missing data can occur with an online service.

Random missing data can be caused by sporadic errors due to aged electrical parts or packet drops during the transmission of data, causing data loss or corruption for an element in the dataset. It tends to be spread out and is not obviously affected by the environment.

C. DATA PRE-PROCESSING

Data retrieved could sometimes require additional processing to reduce possible errors or noise for training and prediction, such as smoothing, outlier detection [56], or removal. Large datasets may require some form of data compression for scalability. Research done by [57] has proposed a data denoising and compression method based on wavelet transformation, along with the construction of a data model.

In cases where there is a lack of traffic data, data augmentation is also considered a way to generate additional data for model training purposes, such as the one conducted by [58]. As traffic data are usually time-series data, [59] has conducted an empirical survey on various time-series data augmentation methods and their suitable use cases. While data augmentation is useful in generating additional datasets, it is important to use it cautiously to avoid distorting the dataset as a whole. Missing data is also considered one form of data pre-processing when it comes to traffic forecasting or routing models, but due to the focus of this paper, data pre-processing is treated as the process before the actual missing data imputation is done.

IV. RESEARCH METHODS

Past literature reviews a specific aspect of missing traffic data imputation, such as [3], [4], [5], and [6], usually focusing on the results but largely ignoring other aspects, such as the road networks or missing data types involved. This section reviews the popularly used methods, broadly categorized into two methods, along with looking into the type of road networks and missing data scenarios used in various literature.

There are generally two categories of missing data imputation methods — Statistical and machine learning. Statistical methods refer to the more classical methods of utilizing mathematical models and statistical theories to impute the data, whereas machine learning makes use of modern computational power and big data to better learn the non-linear, latent features and patterns in a dataset and attempt to learn and output the most likely result based on an input from a similar dataset.

A. STATISTICAL METHODS

Statistical methods analyze the available data and aim to develop a model that best represents the original dataset. Unlike machine learning, which makes use of big data to learn, it is less necessary for statistical methods to need such a large number of data at the cost of being less robust in general.

There are various ways to handle missing data, as mentioned by [60], such as deletion-based methods, learning methods utilizing complete and incomplete data, as well as imputation methods. Mean smoothing has also been used in studies such as [61]. On the other hand, deletion-based methods tend to be avoided as deleting data may result in bias in the estimates and decrease the quality of the dataset itself [62]. Note that deletion-based and mean smoothing represents the simplest methods and are usually not used in missing traffic imputation studies.

With regard to learning methods, predictive mean matching (PMM) based on multiple imputations by chained equations (MICE) has been looked into in [63]. A study done later on has then proceeded to compare variations of PMM methods, including MICE, Classification and Regression Trees (CART), Least Absolute Shrinkage and Selection Operator (LASSO), and random forest, with the result being the MissForest implementation of Random Forest being the best performer [11]. It is noteworthy that random forest is considered a machine learning algorithm, which shows why machine learning tends to be researched more compared to statistical methods, especially in recent years.

Instead, two of the most popular methods for missing data imputation would be Probability Principle Component Analysis (PPCA) and tensor decomposition. These methods are explained below:

1) PROBABILITY PRINCIPLE COMPONENT ANALYSIS (PPCA)

The most commonly used statistical method when it comes to data imputation is the PPCA-based (Probability Principle Component Analysis) model. PPCA is an extension of the Principal Component Analysis (PCA) method through the use of the expectation-maximization algorithm [64]. The resulting probability model results in the ability to better deal with missing data by treating the missing data as not-yet-observed missing data [65].

Recently, [3] has excellently reviewed spatiotemporal PPCA-based data imputation methods in an urban network setting for traffic flow data. As expected, the accuracy of the PPCA-based model changes depending on its field of view, i.e., whether it is a network, sub-network, or single-point imputation. Interestingly, if the view is too large, the result would drop, resulting in more inaccurate results. It was found that for a more realistic use case, the sub-network PPCA-based model worked the best for an urban road network as it is within a reasonable range of detectors.

Focusing on real-time missing traffic data imputation, [66] has proposed a PPCA-based minimum data imputation optimization method that ignores certain missing data points that it deems not required to be imputed, along with simplification of the spatial correlation between road segments on the map. However, not every country has a well-built traffic infrastructure that would provide clear road segment data, thus hampering the effectiveness of data imputation methods that requires the use of spatial data. Furthermore, although the effects may be small, missing data should be imputed to ensure the completion of the data set and to prevent possible bias in prediction results down the line.

Reference [65] also conducted a case study on the PPCA model for traffic analysis, data imputation, and flow prediction, and while the missing data rates tested were not large (1.4%, 4%, and 33% missing data rates), it was found that the PPCA did not show a large degradation in performance when comparing the Weighted Mean Absolute Percentage Error (WMAPE) between the 1.4% and 33% missing rates — around a 1% drop in accuracy from 1.4% to 33% — which means it is overall robust. However, the initial WMAPE itself is rather high at around 14.75%. Despite that, the case study also exhibits the strength of statistical methods, namely the ability to conduct traffic analysis via a breakdown of its principal component scores. While it seems that PPCA was not used much for missing traffic data imputation, it should not be ignored due to its analytical ability, which could contribute to the advancements of itself as well as other techniques.

Additionally, a comparison between MICE and PPCA was made for missing data imputation in the healthcare sector [67], and PPCA was found to have performed better as well, further explaining why this method is one of the more popular statistical methods.

2) TENSOR DECOMPOSITION AND FACTORIZATION

Tensor factorization and its derivatives have seen a significant rise in popularity when it comes to the field of missing data imputation, and missing traffic data is not an exception. This can be seen when comparing the reviewed literature between [8] and [9], noting the tensor factorization methods have shown a spike in use in [9] compared to [8]. In fact, tensor factorization can be considered both a statistical model and a machine learning model. However, tensor factorization is more interpretable compared to other machine learning models because it enables the extraction of a dataset's latent variables via tensor decomposition. Even papers that focus on traffic forecasts, such as [68], make use of tensor decomposition to deal with their missing data before moving on to their proposed model.

Papers such as [18], [19], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], and [79] are some of the recent state-of-the-art missing data imputation methods that have been proposed in the past three years that have utilized tensor factorization as a core part of their model. These tensor-based models performed well due to their being able to extract latent features from a traffic dataset and, through decomposition and completion, can fill in the missing blanks in an accurate manner. Via Bayesian Statistics ([69], [71], [74]), extending or modifying the existing tensor factorization methods ([18], [19], [70], [72], [75]), and even adding an additional pre-processing method ([76], [79]), the base tensor factorization method has shown a significant improvement in the field of missing traffic data imputation. This can also be seen as a majority of these models have been tested for robustness in imputing missing traffic data of rates ranging from 1% to 90% while still retaining a high level of accuracy when compared to their respective benchmarks. Besides that, out of the 13 papers mentioned, 11 of them ([18], [19], [69], [70], [73], [74], [75], [76], [77], [78], [79]) have also been tested on urban traffic networks, raising the evaluation on their robustness as urban traffic tends to be significantly more complicated than freeways/highways/expressways.

However, tensor factorization methods are largely dependent on their dataset and would be unable to perform similarly if the same trained model is tested in another location without retraining [52]. Besides that, tensor decomposition tends not to scale well with larger datasets [80].

B. MACHINE LEARNING

Data-driven models in machine learning methods utilize the availability of data and learn the best weights to obtain the most optimum result for a certain model, as compared to classical statistical methods, which require prior knowledge to derive an appropriate mathematical expression from a given data trend. In general, the model is trained via a training set to 'learn' the optimum values to output given a certain set of inputs. This is a very powerful tool as it requires little to no supervision from the user, but at the same time, a certain understanding of the model may be lost. However, it could be understood that the underlying features of the dataset have been, in theory, mined via the model, allowing it to be more robust and accurate compared to traditional statistical methods.

Neural networks are the models which are the most synonymous with the term machine learning despite just being a subset of it. Regardless, the idea of neural network was introduced in [81] back in 1943 and only began gaining traction in recent years due to the improvement in computation technology. Now, it is being used in various fields, from classification, prediction, and identification to missing data imputation, among others. The following subsections cover the popular methods used in missing data imputation.

1) GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GAN) is a new model proposed in 2014 [82] utilizing a Generator and Discriminator model to train a network. To summarize, the generator continuously attempts to 'trick' the discriminator into that the generated data is the same as the trained dataset. This results in both being trained to generate better, more realistic data as well as more discriminatory testing, allowing the overall model to impute missing data more accurately or realistically, in theory. While not as popular as tensor methods, GAN is a fairly popular method in missing data imputation applications due to its nature of constantly training to create a better dataset to trick the discriminator. This can be seen by the recent papers focusing on GAN methods such as [80], [83], [84], [85], [86], and [87]. As with other methods, this research tends to focus on the Spatiotemporal features of the traffic data ([80], [84], [87]) when conducting traffic data imputation. Some utilize the Attention mechanism ([83], [84]). Besides that, [85] makes use of additional external factors such as weather and holiday factor. Interestingly enough, that research found that external factors excluding holidays do not influence the data imputation much for missing rates less than 40%. While researchers tend to test for high levels of missing data, it could be said that missing traffic data would not be that high. In this case, future researchers can focus on methods that improve the missing traffic data imputation at low missing rates with minimal concern that external factors might cause a large discrepancy in their performance. Another interesting GAN model was proposed by [86], whereby the generated result is once again used as an input into another generator, and the discriminator tries to discriminate between the first generated data and the double-generated data.

2) GRAPH NEURAL NETWORKS

Various real-world datasets are represented as graphs, such as from a social network or the internet itself, and traffic data is not an exception. Traffic networks are naturally represented as a graph, as it is a suitable form in which to visualize road connections and their related information. Realizing this, researchers have proposed the use of Graph Neural Networks (GNN).

Recently, [6] has done a comprehensive survey regarding GNN and has classified various GNN models into four categories — Recurrent GNN, convolutional GNN, graph autoencoders, and spatial-temporal GNN. Among these, we have found that convolutional GNNs are the more popular choice in recent times when it comes to traffic research, as shown by [52], [88], [89], and [90]. Convolutional GNNs, or Graph Convolutional Networks (GCN), utilize convolutional neural networks to embed graph information into a tensor, resulting in a uniform framework from irregular datasets [89].

While GNN and GCN are popular methods used in traffic studies, most recent research focuses on traffic forecasting, and not as many focus on missing data imputation. Some research, such as [88], treats missing data as part of the traffic prediction process instead of the focus of the problem. This could be useful as traffic actions tend to require real-time analyses and predictions. While it is good to design a robust traffic prediction model towards missing data, having a missing data imputation model should not be neglected as it can further enhance the already robust traffic prediction model. On the other hand, [52] is more focused on missing data imputation, proposing a model that uses a bidirectional recurrent network (RNN) to capture temporal patterns and GCN to capture spatial patterns. Meanwhile, [90] proposed a Graph neural network that makes use of the attention mechanism, as well as a temporal convolutional network instead of RNN as standard RNN, which suffers from various drawbacks such as being unable to hold memory for long, prone to vanishing or exploding gradients, and having low efficiency in parallel training and inference. While not exactly imputing missing traffic data itself, [89] combined GCN with a mapping function to impute missing spatial flow data. This is another important aspect of traffic data that the authors believe should be highlighted and received attention, as origin-destination flow data can be a vital addition to other traffic-related models that could use additional traffic features.

Besides GCN, there are also pieces of literature, such as [7] and [91], that make use of spatial-temporal GNN instead. In other words, instead of utilizing convolution for feature extraction and graph embedding, the research proposes other methods, such as the fusion of multiple data sources ([7], [91]) or attention mechanism, as well as multitask learning [91].

As the concept of GNN was introduced relatively early in 2005 [92], and GCN itself was only introduced even more recently in 2017 [93], there is still plenty of room for improvement, as can be seen by the recent literature mentioned above. As road networks differ depending on the location, it is imperative to find a model that is robust towards various forms of missing data and the structure of road networks. GNN may have a strong potential in this due to its deep-learning structure as compared to tensor decomposition methods which might be more transductive.

C. ENSEMBLE MODEL

A single model tends to have some forms of shortcomings along with its advantages. In this case, researchers have come up with the idea to combine multiple models to resolve each model's weaknesses and enhance their strengths further.

For example, [94] uses the very popular tensor decomposition but utilizes a Fuzzy Neural Network to further enhance the imputation accuracy by optimizing the weights of the tensor resolvers. Besides that, [95] combines GCN and tensor decomposition using graph Laplace for tensor completion.

Meanwhile, [20] designs a framework combining matrix modeling and factorization and conducting matrix decomposition before using a dendrite neural network to fuse the information to obtain the final imputed data. Besides being another ensemble model, the proposed neural network model was recently proposed by [96], of which the code is provided in their paper. This could be another good avenue for researchers to look into as it expands upon the existing neuron structure to further resemble the human nervous system.

As shown, these ensemble models make use of already established methods while modifying them to work together to obtain an even greater result. However, it should be noted that using more models would inadvertently increase computation time, which may result in the inability to function in a real-time scenario.

D. OVERVIEW OF RESEARCH METHODS

Information regarding the base method used, the referenced papers, the type of road network, the method of data acquisition (i.e., Public, private, or manually collected dataset), and the type of missing data tested was summarised in TABLE 1.

Regarding the road network, data acquisition, and missing types of columns, the number of papers reviewed that fulfilled the criteria were counted, and the sum is shown in the table cells.

Other statistical methods include the MICE implementation [63] and Gaussian Processes [97], [98] which are less popular methods but were nonetheless researched relatively recently and showed good performance when benchmarked against established methods.

From TABLE 1, it can be seen that many of the reviewed literature were conducted in an urban network setting. This is because urban networks are the most volatile as well as the busiest, making them the most in need of the support of intelligent transportation systems. However, a deeper look into the datasets used shows that many of the datasets are the same set of data, such as the Guangzhou urban traffic speed dataset, or datasets related to public transport, such as Taxi traffic data. Studies such as [97] and [98] made use of crowdsourced data from Google Maps' Location Sharing function, which could allow more flexibility in the location chosen at the cost of access to specific traffic data variables due to certain information being hidden due to user privacy and security [98]. These datasets are either limited in their location or in their comprehensiveness, as other countries do not have the same traffic patterns or road networks as America or China. Neither do taxis represent the entire state of the network at any time. This shows that researchers need to conduct simulations based on a larger variety of locations and utilize datasets that better represent the state of the traffic. Besides that, not many researchers acquire their traffic data manually but rather utilize public datasets or datasets from the private sector. This is understandable but is also a form of limitation, as discussed later.

Meanwhile, it can be seen that random missing types are almost always tested, followed by fiber missing and block missing. While some literature has mentioned block missing, by this paper's definition, they are fiber missing as it is only one source of data or the missing data period is not long enough. More research could be put into this particular missing type.

Additionally, a summary of the forecasted variables is shown in TABLE 2. The variable most studies focus on imputing is traffic speed, for obvious reasons, as it is the most direct traffic data used that tells the exact state of the traffic. This is followed by traffic flow, which could be due to the unavailability of the dataset for the area. It is interesting to note that the majority of the studies that impute traffic flow shown in TABLE 2 are those which use taxi GPS data, which could explain this situation as taxi GPS data may not have accurate traffic speeds logged in. However, GPS data does provide researchers with a more detailed view of the road network, which would help in proving the robustness of their work. Traffic volume sees fewer missing data imputation studies, likely due to data availability and the rather imprecise nature of traffic volume. However, traffic volume does give a good idea of the state of the traffic as well. Travel time and congestion levels are outlier studies but are also other parameters to keep in mind for future research.

Most studies focus on imputing only one traffic data, the exception being [7], which had done missing data imputation on both traffic speed and traffic volume, which leads to further proving their model's credibility.

E. POPULAR METHODS' ANALYSIS

TABLE 3 lists the advantages and disadvantages of the popular methods mentioned in TABLE 1. As a general guideline, future studies should take into account the accuracy, interpretability, as well as computational complexity into account when designing a model.

V. NOTABLE DESIGN DECISIONS

Section IV discussed the popular base models that were the focus of recent papers. This section discusses the popular design decisions that the literature tends to use to augment their base models. To the best of the authors' knowledge, past literature reviews do not look into the common mechanism

TABLE 1. Summary of reviewed literature regarding the number of research covering various topics such as the road network, data acquisition, and missing types.

Methods	Papers	Road Network		Data Acquisition		Missing Types		
		Freeway/ Road Segments	Urban Network	Public/Private Datasets	Self	Random	Fiber	Block
Principal Component Analysis	[65], [66]	0	2	0	2	1	1	0
Tensor Decomposition	[69], [70], [71], [72], [18], [73], [74], [75], [19], [76], [77], [78], [79]	4	11	11	3	12	7	1
Generative Adversarial Networks	[80], [83], [84], [85], [86], [87]	4	3	7	0	6	3	2
Graph Neural Networks	[88], [89], [52], [90], [7], [91]	3	3	4	2	6	3	3
Ensemble Models	[94], [95], [20], [53]	3	2	1	3	3	4	0
Other Statistical Methods	[63], [97], [98]	2	1	3	0	2	2	1

TABLE 2. Summary of forecasted variables for reviewed literature.

Methods	Papers	Forecasted Variables				
		Traffic Speed	Traffic Flow	Traffic Volume/Count	Travel Time	Congestion Level
Principal Component Analysis	[65], [66]	0	1	1	0	0
Tensor Decomposition	[69], [70], [71], [72], [18], [73], [74], [75], [19], [76], [77], [78], [79]	8	3	0	1	1
Generative Adversarial Networks	[80], [83], [84], [85], [86], [87]	3	2	1	0	0
Graph Neural Networks	[88], [89], [52], [90], [7], [91]	4	1	2	0	0
Ensemble Models	[94], [95], [20], [53]	1	2	1	0	0
Other Statistical Methods	[63], [97], [98]	2	0	1	0	0

used between different reviewed models and focus more on the overall quality of each individual model instead.

A. ATTENTION MECHANISM

The attention mechanism is widely used in many studies due to its optimization abilities, such as by [95] for weight optimization or extracting multiple features like in [51]. Besides those two, it can be seen that a few of the literature reviewed had also incorporated the attention mechanism into their model [52], [84], [90], [91].

It should be obvious that the attention mechanism is proving to be a very good mechanism to be added when dealing with feature extraction or weight optimization, and more research should take note of it should they require such functions. To that end, [99] has reviewed the state-of-the-art attention models proposed recently, as well as provided more in-depth points when making use of this mechanism along with their real-life applications.

B. EXPECTATION MAXIMIZATION

Probability Principal Component Analysis (PPCA) applies this algorithm to the base Principal Component Analysis (PCA) to derive a probabilistic formulation of the PCA. This is important as it allows for the application of Bayesian methods as an extension to the existing PCA [64], allowing for further improvements as well as analysis to be done, as shown by [100]. This trait can be used in other models as well to possibly provide deeper insight and data for machine learning.

C. FUZZY THEORY

Fuzzy theory introduces the concept of membership functions, which allows variables to be partially a part of a set instead of a single yes or no. This allows uncertain or imprecise data to be represented in a more flexible manner. While this has seen use in many fields for missing traffic data, it has

TABLE 3. Advantages and disadvantages of popular methods.

Methods	State of the Art	Research Gap
Principal Component Analysis	Allows traffic analysis to be done while also providing a good level of missing data imputation	The performance of missing traffic data falls behind the other popular methods.
Tensor Decomposition	Able to impute missing traffic data well. Allows some traffic analysis to be done by decomposing traffic data into latent variables	Faces scalability issues High computational costs when conducting missing data imputation due to requiring multiple iterations A large number of data is required to perform optimally.
Generative Adversarial Networks	Able to impute accurate and realistic traffic data due to the adversarial network design.	Neural network methods are difficult to interpret. Fine-tuning the parameters for the network can be exhaustive. A large number of data is required to perform optimally.
Graph Neural Networks	The structure of the network suits that of a traffic network, making certain operations more intuitive could be easier to visualize compared to other neural network-based methods. The graph structure makes GNN more scalable for traffic networks	A large number of data is required to perform optimally The structure of GNN limits the types of datasets it can work on, although traffic data should be fine due to the similar structure.
Ensemble Models	Makes use of the strengths of multiple models and minimizes or eliminates their weaknesses.	An increase in the number of models may inadvertently result in an increase in computational complexity, which may be detrimental to a real-time system such as ITS. Depending on the type of models involved, highly likely to require a large number of data to perform optimally.

seen minimal use, such as for [53], which makes use of fuzzy rough sets combined with a fuzzy neural network. Another study using fuzzy theory for missing data imputation is [101], which used a hybrid model combining fuzzy rough sets with fuzzy C-means. However, the study was conducted using a medical dataset.

Despite seeing minimal uses, the authors found the method worth mentioning as traffic data tends to be rather imprecise, due to many external variables. Fuzzy theory could potentially improve the performance of other models in a hybrid setting, as shown by the research above.

VI. CHALLENGES AND LIMITATIONS

This section covers certain challenges that existing literature faces and suggestions regarding the directions future researchers should take when undertaking their research. The focus of the challenges and limitations mentioned here are with regards to large-scale deployment in different areas, of which the common issue would be traffic data retrieval, as well as scalability problems and model interpretability as explained below:

A. LIMITATION OF DATA

It can be seen from TABLE 1 that most of the traffic data used came from publicly available datasets, while some are obtained via other special methods such as private institutions, while research that has attempted to collect the traffic data manually is fewer than those using existing datasets. While it could be said that successful simulations on these datasets would imply similar results in other datasets, researchers belonging to countries with limited public datasets available to them might still want to test for the model's validity in their own country and location of choice. In such situations, it should be noted that online traffic APIs, as mentioned in Section III-A2, could be used to collect the relevant traffic data as they can leverage the companies' existing infrastructures (i.e., Various sources of data) when collecting data, which ensures a level of reliability on top of the accessibility.

However, collecting a significant number of traffic data requires a significant amount of time, and as such, researchers might not have a lot of data points when compared to the available public datasets. This can be seen from some of the reviewed papers, such as [7], [19], [65], [66], [69], [91], and [20], whereby their datasets range from 14 days to 2 months. The quantity of data collected within this short period would prove difficult for models which rely on many data to be properly trained. This leads to the research question of how well a model can do when facing the problem of a limited amount of training data.

Models which relied on external features such as [84] and [85] might suffer a reduction in performance, but based on the experiments by [85], it would seem that external factors might not play such a big part unless the missing rate is greater than 40%, barring holiday factors, of which future researchers are encouraged to try to differentiate between holiday and non-holiday traffic datasets whenever possible.

As previously mentioned, manually gathering data takes a long time, and the collected data would be lacking in both quantity and quantity given that some research may be conducted under a time constraint and data collecting equipment may suffer from occasional breakdowns or failure in data transmission. While most, if not all, methods might have lowered performance, data-driven methods might suffer the most, depending on how limited the quantity of data is. However, there are methods such as data augmentation mentioned in Section III-C, which could be used with caution, as well as data generation methods via GAN, which could be researched.

B. SCALABILITY

Scalability is another challenge that researchers face. While tensor-based methods are popular, they also suffer a problem of scalability — as the size of the dataset increases, so does the compuational cost to conduct the traffic imputation. In times like those, a data-driven approach might be a better method. However, research into developing a scalable tensor decomposition missing traffic imputation method should not be ignored [79] has proposed a method that utilizes the tensor nuclear norm minimization scheme to model the inherent low-rank property of traffic data, breaking down the large tensor into smaller matrices, allowing for an overall more efficient computation while maintaining a similar level of accuracy. More research should be investigated to improve the accuracy further and reduce the computational cost. In [79], the comparison was made between existing tensor-based models but not with other models, such as the other machine learning models, so further testing can still be done.

C. MODEL INTERPRETABILITY AND ROBUSTNESS

Machine learning or data-driven models tend to have less interpretability than statistical methods, which is understandable given how they work. However, future researchers should take note of how much data is being used in their model to reduce their model's computational cost. A more interpretable model also helps researchers see which part of their model can be adjusted or trimmed, especially in situations with a limited dataset.

Besides that, while statistical methods have interpretability, they lack robustness in contrast to data-driven or machine learning, as the learned model is highly dependent on the dataset they were trained on and might require retraining. Tensor-based models, which are probably the most popular statistical method, also suffer from this issue, as mentioned by [52]. However, deep-learning methods may be more robust towards this issue, and this topic is constantly being researched, making the deep-learning model more computationally efficient for improved real-time utilization.

There is always a cost when selecting a model, and depending on the goal of the researcher, the most optimum model is selected. The idea is to design a model that is both interpretable for further understanding as well as robust to various changes in the dataset's environment while keeping the cost to a minimum. Ensemble models such as [94] managed to propose a robust, generalized model utilizing both Fuzzy Neural Network and tensor decomposition, which could be said to be an improvement but at the cost of computational complexity.

VII. CONCLUSION AND FUTURE WORK

This paper introduced, defined, and categorized the three different types of missing data, namely random missing, fiber missing, and block missing. This paper reviewed various popular state-of-the-art methods and their corresponding research in recent years based on the papers' focus and goals. It was found that tensor decomposition has been used a lot in recent years. However, tensor computations could lead to scalability problems and are dependent on the location of the training dataset. Generative Adversarial Networks (GAN) and Graph Neural Networks (GNN) were found to be similarly popular, as both are well used in data generation and traffic networks, respectively. Both GAN and GNN are relatively new models, with Graph Convolutional Networks (GCN) emerging as a branch of GNN. It is also shown that the attention mechanism and expectation-maximization algorithm are popularly used as auxiliary methods to help bolster the base model's missing traffic data imputation capabilities. In addition, this paper also discussed the limitations of popular datasets and collection methods. Various challenges related to the scalability and availability of data have been highlighted with different data collection methods. As traffic differs from location to location, even within a country, different countries would have different traffic patterns.

Moving forward, a traffic data-collecting initiative for improved traffic performance is encouraged for further analysis. Researchers also need to develop methods that are robust toward different locations' traffic patterns. The lack of data while keeping in mind the interpretability of the models is important. Methods such as PPCA have shown their strength in breaking down traffic analyses, which could help in further understanding the various traffic factors as well as determining what variables have the largest influence on the accuracy and robustness of the data imputation. Scalability of models to function well enough for real-time applications is also important. While tensor factorization tends to suffer from scalability and complexity issues, there are also studies done regarding the design of an online and quicker algorithm for it as well, making this another topic worth pursuing.

REFERENCES

[1] W. F. Velicer and S. M. Colby, "A comparison of missing-data procedures for ARIMA time-series analysis," *Educ. Psychol. Meas.*, vol. 65, no. 4, pp. 596–615, Aug. 2005.

[2] P. Wu, L. Xu, and Z. Huang, "Imputation methods used in missing traffic data: A literature review," in *Proc. Int. Symp. Intell. Comput. Appl.* vol. 1205. Singapore: Springer, 2020, pp. 662–677.

[3] E. Joelianto, M. F. Fathurrahman, H. Y. Sutarto, I. Semanjski, A. Putri, and S. Gautama, "Analysis of spatiotemporal data imputation methods for traffic flow data in urban networks," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 5, p. 310, May 2022.

[4] J. Xing, W. Wu, Q. Cheng, and R. Liu, "Traffic state estimation of urban road networks by multi-source data fusion: Review and new insights," *Phys. A, Stat. Mech. Appl.*, vol. 595, Jun. 2022, Art. no. 127079.

[5] 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, p. 2544, Jul. 2022.

[6] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2020.

[7] P. Wang, T. Zhang, Y. Zheng, and T. Hu, "A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation," *Int. J. Geograph. Inf. Sci.*, vol. 36, no. 6, pp. 1231–1257, Jun. 2022.

[8] W.-C. Lin and C.-F. Tsai, "Missing value imputation: A review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020.

[9] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)," *Informat. Med. Unlocked*, vol. 27, Jan. 2021, Art. no. 100799.

[10] Y. Li, Z. Li, and L. Li, "Missing traffic data: Comparison of imputation methods," *IET Intell. Transp. Syst.*, vol. 8, no. 1, pp. 51–57, Feb. 2014.

[11] M. A. Shafique, "Imputing missing data in hourly traffic counts," *Sensors*, vol. 22, no. 24, p. 9876, Dec. 2022.

[12] Caltrans. *PeMS Data Source*. Accessed: Oct. 15, 2020. [Online]. Available: https://dot.ca.gov/programs/traffic-operations/mpr/pems-source

[13] J. Z. Zhu, J. X. Cao, and Y. Zhu, "Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections," *Transp. Res. C, Emerg. Technol.*, vol. 47, pp. 139–154, Oct. 2014.

[14] J. Wang, W. Deng, and Y. Guo, "New Bayesian combination method for short-term traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 79–94, Jun. 2014.

[15] R. Tahmasbi and S. M. Hashemi, "Modeling and forecasting the urban volume using stochastic differential equations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 250–259, Feb. 2014.

[16] Y. Qi and S. Ishak, "A hidden Markov model for short term prediction of traffic conditions on freeways," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 95–111, Jun. 2014.

[17] K. Y. Chan, T. Dillon, E. Chang, and J. Singh, "Prediction of short-term traffic variables using intelligent swarm-based neural networks," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 1, pp. 263–274, Jan. 2013.

[18] X. Jia, X. Dong, M. Chen, and X. Yu, "Missing data imputation for traffic congestion data based on joint matrix factorization," *Knowl.-Based Syst.*, vol. 225, Aug. 2021, Art. no. 107114.

[19] T. Nie, G. Qin, and J. Sun, "Truncated tensor Schatten *p*-norm based approach for spatiotemporal traffic data imputation with complicated missing patterns," *Transp. Res. C, Emerg. Technol.*, vol. 141, Aug. 2022, Art. no. 103737.

[20] P. Wang, T. Hu, F. Gao, R. Wu, W. Guo, and X. Zhu, "A hybrid data-driven framework for spatiotemporal traffic flow data imputation," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16343–16352, Sep. 2022.

[21] Google. *Google Maps Platform*. Google Developers. Accessed: Sep. 4, 2019. [Online]. Available: https://developers.google.com/maps/documentation/

[22] Msdn.microsoft.com. *Bing Maps Traffic Coverage*. Accessed: Sep. 4, 2019. [Online]. Available: https://docs.microsoft.com/en-us/bingmaps/coverage/traffic-coverage

[23] HERE. *What is the Traffic API?—Traffic API—HERE Developer*. Accessed: Sep. 4, 2019. [Online]. Available: https://developer.here.com/documentation/traffic/topics/what-is.html

[24] TomTom Developer. *Homepage*. Accessed: Sep. 4, 2019. [Online]. Available: https://developer.tomtom.com/

[25] V. Verendel and S. Yeh, "Measuring traffic in cities through a large-scale online platform," *J. Big Data Anal. Transp.*, vol. 1, nos. 2–3, pp. 161–173, Dec. 2019.

[26] I. Triguero, S. González, J. M. Moyano, S. García, J. Alcalá-Fdez, J. Luengo, A. Fernández, M. J. del Jesús, L. Sánchez, and F. Herrera, "KEEL 3.0: An open source software for multi-stage analysis in data mining," *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, p. 1238, 2017.

[27] F. Schimbinschi, L. Moreira-Matias, V. X. Nguyen, and J. Bailey, "Topology-regularized universal vector autoregression for traffic forecasting in large urban areas," *Expert Syst. Appl.*, vol. 82, pp. 301–316, Oct. 2017.

[28] Y. Rajabzadeh, A. H. Rezaie, and H. Amindavar, "Short-term traffic flow prediction using time-varying vasicek model," *Transp. Res. C, Emerg. Technol.*, vol. 74, pp. 168–181, Jan. 2017.

[29] X. Zhang, E. Onieva, A. Perallos, E. Osaba, and V. C. Lee, "Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 127–142, Jun. 2014.

[30] T. L. Pan, A. Sumalee, R. X. Zhong, and N. Indra-Payoong, "Short-term traffic state prediction based on temporal–spatial correlation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1242–1254, Sep. 2013.

[31] Y.-S. Jeong, Y.-J. Byon, M. Mendonca Castro-Neto, and S. M. Easa, "Supervised weighting-online learning algorithm for short-term traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1700–1707, Dec. 2013.

[32] X. Ling, X. Feng, Z. Chen, Y. Xu, and Z. Haifeng, "Short-term traffic flow prediction with optimized multi-kernel support vector machine," in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2017, pp. 294–300.

[33] M. Fouladgar, M. Parchami, R. Elmasri, and A. Ghaderi, "Scalable deep traffic flow neural networks for urban traffic congestion prediction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2251–2258.

[34] H. Tan, Y. Wu, B. Shen, P. J. Jin, and B. Ran, "Short-term traffic prediction based on dynamic tensor completion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 1–11, Aug. 2016.

[35] P. Lopez-Garcia, E. Onieva, E. Osaba, A. D. Masegosa, and A. Perallos, "A hybrid method for short-term traffic congestion forecasting using genetic algorithms and cross entropy," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 557–569, Feb. 2016.

[36] R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," in *Proc. 31st Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Nov. 2016, pp. 324–328.

[37] K. Abhishek and B. B. Misra, "Hybrid genetic algorithm and time delay neural network model for forecasting traffic flow," in *Proc. IEEE Int. Conf. Eng. Technol. (ICETECH)*, Mar. 2016, pp. 178–183.

[38] K. Abhishek, "Novel multi input parameter time delay neural network model for traffic flow prediction," in *Proc. Online Int. Conf. Green Eng. Technol. (IC-GET)*. Coimbatore, India: IEEE, May 2017, doi: 10.1109/GET.2016.7916799.

[39] Y. Tian and L. Pan, "Predicting short-term traffic flow by long short-term memory recurrent neural network," in *Proc. IEEE Int. Conf. Smart City*, Dec. 2015, pp. 153–158.

[40] I. M. Wagner-Muns, I. G. Guardiola, V. A. Samaranayke, and W. I. Kayani, "Traffic, Volume Forecasting," *Volumes*, vol. 19, no. 3, pp. 1–11, 2017.

[41] N. G. Polson and O. S. Vadim, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79 pp. 1–17, Dec. 2017.

[42] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 61–78, May 2015.

[43] Y. Hou, P. Edara, and C. Sun, "Traffic flow forecasting for urban work zones," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1761–1770, Aug. 2015.

[44] Z. Xiong, D. Rey, T. Mao, H. Liu, V. V. Dixit, and S. T. Waller, "A three-stage framework for motorway travel time prediction," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 816–821.

[45] J. Guo, W. Huang, and B. M. Williams, "Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 50–64, Jun. 2014.

[46] Y. Xu, Q.-J. Kong, R. Klette, and Y. Liu, "Accurate and interpretable Bayesian MARS for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2457–2469, Dec. 2014.

[47] Y. Xu, Q. J. Kong, and Y. Liu, "Short-term traffic volume prediction using classification and regression trees," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2013, pp. 493–498.

[48] Y. Xu, Q.-J. Kong, and Y. Liu, "A spatio-temporal multivariate adaptive regression splines approach for short-term freeway traffic volume prediction," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2013, pp. 217–222.

[49] Y.-Y. Chen, Y. Lv, Z. Li, and F.-Y. Wang, "Long short-term memory model for traffic congestion prediction with online open data," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 132–137.

[50] 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.

[51] X. Wu, M. Xu, J. Fang, and X. Wu, ''A multi-attention tensor completion network for spatiotemporal traffic data imputation,'' *IEEE Internet Things J.*, vol. 9, no. 20, pp. 1–11, Oct. 2022.

[52] Y. Liang, Z. Zhao, and L. Sun, ''Memory-augmented dynamic graph convolution networks for traffic data imputation with diverse missing patterns,'' *Transp. Res. C, Emerg. Technol.*, vol. 143, Oct. 2022, Art. no. 103826.

[53] 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,'' *J. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 439–454, Sep. 2021.

[54] 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,'' *Transp. Res. C, Emerg. Technol.*, vol. 118, Sep. 2020, Art. no. 102674.

[55] J. Du, M. Hu, and W. Zhang, ''Missing data problem in the monitoring system: A review,'' *IEEE Sensors J.*, vol. 20, no. 23, pp. 13984–13998, Dec. 2020.

[56] T. Mecheva, ''Outlier detection in traffic data set,'' in *Proc. AIP Conf.*, vol. 2449. Plovdiv, Bulgaria: AIP Publishing, p. 040014, Sep. 2022, doi: 10.1063/5.0093554.

[57] H. Dou and G. Wang, ''Data denoising and compression of intelligent transportation system based on two-dimensional discrete wavelet transform,'' *Int. J. Commun. Syst.*, vol. 34, no. 10, pp. 1–15, Jul. 2021.

[58] F. Sun, A. Dubey, and J. White, ''DxNAT—Deep neural networks for explaining non-recurring traffic congestion,'' in *Proc. IEEE Int. Conf. Big Data*, Jan. 2018, pp. 2141–2150.

[59] B. K. Iwana and S. Uchida, ''An empirical survey of data augmentation for time series classification with neural networks,'' *PLOS One* vol. 16, no. 7, Jul. 2021, Art. no. e0254841.

[60] N. Karmitsa, S. Taheri, A. Bagirov, and P. Makinen, ''Missing value imputation via clusterwise linear regression,'' *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1889–1901, Apr. 2022.

[61] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, ''LSTM-based traffic flow prediction with missing data,'' *Neurocomputing*, vol. 318, pp. 297–305, Nov. 2018.

[62] C. Velasco-Gallego and I. Lazakis, ''Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study,'' *Ocean Eng.*, vol. 218, Dec. 2020, Art. no. 108261.

[63] K. Henrickson, Y. Zou, and Y. Wang, ''Flexible and robust method for missing loop detector data imputation,'' *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2527, pp. 29–36, Nov. 2015.

[64] C. Bishop and M. Tipping, ''Probabilistic principal component analysis,'' *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.

[65] M. F. Fathurrahman, H. Y. Sutarto, and I. Semanjski, ''Urban network traffic analysis, data imputation, and flow prediction based on probabilistic PCA model of traffic volume data,'' in *Proc. 8th Int. Conf. Adv. Inform., Concepts, Theory Appl. (ICAICTA)*, Sep. 2021, pp. 1–6.

[66] A. Liu, C. Li, W. Yue, and X. Zhou, ''Real-time traffic prediction: A novel imputation optimization algorithm with missing data,'' in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[67] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya, ''MICE vs PPCA: Missing data imputation in healthcare,'' *Informat. Med. Unlocked*, vol. 17, Jan. 2019, Art. no. 100275.

[68] J. Liu, G. P. Ong, and X. Chen, ''GraphSAGE-based traffic speed forecasting for segment network with sparse data,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1755–1766, Mar. 2022.

[69] X. Chen, Z. He, and L. Sun, ''A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation,'' *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 73–84, Jan. 2019.

[70] M. Nouri, M. Reisi-Gahrooei, and M. Ilbeigi, ''A method for granular traffic data imputation based on PARATUCK2,'' *Transp. Res. Rec.*, vol. 2676, Oct. 2022, Art. no. 036119812210892.

[71] M. Lei, A. Labbe, Y. Wu, and L. Sun, ''Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and Kriging,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18962–18974, Oct. 2022.

[72] A. Baggag, S. Abbar, A. Sharma, T. Zanouda, A. Al-Homaid, A. Mohan, and J. Srivastava, ''Learning spatiotemporal latent factors of traffic via regularized tensor factorization: Imputing missing values and forecasting,'' *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2573–2587, Jun. 2021.

[73] M. Bhanu, J. Mendes-moreira, and J. Chandra, ''Embedding traffic network characteristics using tensor for improved traffic prediction,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 1–13, Jun. 2021.

[74] X. Chen, Z. He, Y. Chen, Y. Lu, and J. Wang, ''Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model,'' *Transp. Res. C, Emerg. Technol.*, vol. 104, pp. 66–77, Jul. 2019.

[75] A. Ben Said and A. Erradi, ''Spatiotemporal tensor completion for improved urban traffic imputation,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6836–6849, Jul. 2022.

[76] C. Gong and Y. Zhang, ''Urban traffic data imputation with detrending and tensor decomposition,'' *IEEE Access*, vol. 8, pp. 11124–11137, 2020.

[77] H. Yu, S. Liu, H. Jiang, and Y. Ren, ''Vehicle trajectory reconstruction using a tensor-based individual travel time matching method,'' in *Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, Jul. 2021, pp. 184–187.

[78] J. Li, L. Xu, R. Li, P. Wu, and Z. Huang, ''Deep spatial-temporal bidirectional residual optimisation based on tensor decomposition for traffic data imputation on urban road network,'' *Int. J. Speech Technol.*, vol. 52, no. 10, pp. 11363–11381, Aug. 2022.

[79] X. Chen, Y. Chen, N. Saunier, and L. Sun, ''Scalable low-rank tensor learning for spatiotemporal traffic data imputation,'' *Transp. Res. C, Emerg. Technol.*, vol. 129, Aug. 2021, Art. no. 103226.

[80] Y. Yuan, Y. Zhang, B. Wang, Y. Peng, Y. Hu, and B. Yin, ''STGAN: Spatio-temporal generative adversarial network for traffic data imputation,'' *IEEE Trans. Big Data*, vol. 9, no. 1, pp. 1–13, Feb. 2022.

[81] W. Mcculloch and W. Pitts, ''A logical calculus of ideas immanent in nervous activity,'' *Bull. Math. Biophys.*, vol. 5, pp. 127–147, Dec. 1943.

[82] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ''Generative adversarial networks,'' *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[83] 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 Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7919–7930, May 2021.

[84] B. Yang, Y. Kang, Y. Yuan, X. Huang, and H. Li, ''ST-LBAGAN: Spatio-temporal learnable bidirectional attention generative adversarial networks for missing traffic data imputation,'' *Knowl.-Based Syst.*, vol. 215, Mar. 2021, Art. no. 106705.

[85] B. Yang, Y. Kang, Y. Yuan, H. Li, and F. Wang, ''ST-FVGAN: Filling series traffic missing values with generative adversarial network,'' *Transp. Lett.*, vol. 14, no. 4, pp. 407–415, Apr. 2022.

[86] A. Kazemi and H. Meidani, ''IGANI: Iterative generative adversarial networks for imputation with application to traffic data,'' *IEEE Access*, vol. 9, pp. 112966–112977, 2021.

[87] D. Xu, H. Peng, C. Wei, X. Shang, and H. Li, ''Traffic state data imputation: An efficient generating method based on the graph aggregator,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13084–13093, Aug. 2022.

[88] Y. Chen and X. Chen, ''A novel reinforced dynamic graph convolutional network model with data imputation for network-wide traffic flow prediction,'' *Transp. Res. C, Emerg. Technol.*, vol. 143, Oct. 2022, Art. no. 103820.

[89] X. Yao, Y. Gao, D. Zhu, E. Manley, J. Wang, and Y. Liu, ''Spatial origin-destination flow imputation using graph convolutional networks,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7474–7484, Dec. 2021.

[90] W. Liang, Y. Li, K. Xie, D. Zhang, K.-C. Li, A. Souri, and K. Li, ''Spatial-temporal aware inductive graph neural network for C-ITS data recovery,'' *IEEE Trans. Intell. Transp. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TITS.2022.3156266.

[91] R. Liu, Y. Kan, S. Zhao, B. Cheng, Z. Ma, and W. Wu, ''Turning traffic volume imputation for persistent missing patterns with GNNs,'' *Int. J. Speech Technol.*, vol. 53, no. 1, pp. 491–508, Jan. 2023.

[92] M. Gori, G. Monfardini, and F. Scarselli, ''A new model for learning in graph domains,'' in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 2, Jul. 2005, pp. 729–734.

[93] T. N. Kipf and M. Welling, ''Semi-supervised classification with graph convolutional networks,'' in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[94] T. Zhang, D.-G. Zhang, H.-R. Yan, J.-N. Qiu, and J.-X. Gao, ''A new method of data missing estimation with FNN-based tensor heterogeneous ensemble learning for Internet of Vehicle,'' *Neurocomputing*, vol. 420, pp. 98–110, Jan. 2021.

[95] H. Dong, F. Ding, H. Tan, and H. Zhang, "Laplacian integration of graph convolutional network with tensor completion for traffic prediction with missing data in inter-city highway network," *Phys. A, Stat. Mech. Appl.*, vol. 586, Jan. 2022, Art. no. 126474.

[96] G. Liu and J. Wang, "Dendrite net: A white-box module for classification, regression, and system identification," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13774–13787, Dec. 2022.

[97] F. Rodrigues, K. Henrickson, and F. C. Pereira, "Multi-output Gaussian processes for crowdsourced traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 2, pp. 594–603, Feb. 2019.

[98] F. Rodrigues and F. C. Pereira, "Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data," *Transp. Res. C, Emerg. Technol.*, vol. 95, pp. 636–651, Oct. 2018.

[99] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[100] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 108–120, Sep. 2013.

[101] D. Li, H. Zhang, T. Li, A. Bouras, X. Yu, and T. Wang, "Hybrid missing value imputation algorithms using fuzzy C-means and vaguely quantified rough set," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 5, pp. 1396–1408, May 2022.

ROBIN KUOK CHEONG CHAN received the B.E. degree (Hons.) in electrical and computer systems from Monash University Malaysia, where he is currently pursuing the Ph.D. degree with the School of Engineering. His research interests include artificial intelligence, optimization, and data mining.

JOANNE MUN-YEE LIM (Member, IEEE) received the Ph.D. degree in engineering from Multimedia University. She is currently a Senior Lecturer with Monash University Malaysia. Her research interests include transportation, vehicular ad hoc networks (VANETs), artificial intelligence, optimization schemes, robotic design, and drones and applications. She is a Professional Engineer of the Board of Engineers Malaysia.

RAJENDRAN PARTHIBAN (Senior Member, IEEE) received the B.E. degree (Hons.) and the Ph.D. degree in optical networks from The University of Melbourne, Australia, in 1997 and 2004, respectively. In 2006, he joined as a Lecturer with the School of Engineering, Monash University Malaysia, where he was a Professor, in 2022, and has been the Deputy Head of School (Education), since August 2010. He moved to the Faculty of Engineering, Monash University, Australia, as the Associate Dean (Education), and has been a Professor, since August 2022. His research interests include optical networks, visible light communications, vehicular communication, the Internet of Things, and engineering education. He is a Senior Member of Optica [formerly known as the Optical Society of America (OSA)].