Intelligent Systems with Applications 20 (2023) 200268

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Review

# A survey on traffic flow prediction and classification

Bernardo Gomes [a], José Coelho [a,b], Helena Aidos [a,*]

[a] LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
[b] ESHTE, Estoril Higher Institute for Tourism and Hotel Studies, Cascais, Portugal

ARTICLE INFO

*Keywords:*
Road traffic
Prediction
Classification
Europe traffic flow

ABSTRACT

As cities continue to grow and the number of vehicles on the road increases, traffic congestion and pollution have become major issues. Fortunately, significant efforts have been made in recent decades to alleviate these problems through research and the development of Intelligent Transportation Systems (ITS). Governments are now utilizing advanced ITS technologies to better understand traffic patterns and make informed decisions on how to manage traffic. In this paper, we will explore the state-of-the-art methods employed in ITS for predicting traffic flow and speed, as well as classifying different traffic situations. We will also examine the preprocessing techniques used in these tasks, along with the metrics used to evaluate the results. By understanding the latest advancements in ITS, we can work towards creating more efficient and sustainable transportation systems that benefit both individuals and society as a whole.

## 1. Introduction

Data science has emerged as a crucial field in recent decades, enabling the extraction of knowledge, pattern detection, and data-driven insights for informed decision-making. The rapid growth of population and vehicles (More et al., 2016) originated several other problems (time spent in traffic, health issues related to stress, increase in fuel consumption, air and noise pollution and the number of accidents) creating an urgent need for intelligent vehicular systems that can efficiently manage and control traffic (Fitters et al., 2021, Priambodo & Ahmad, 2018). These systems are essential not only for city administrations but also for individual commuters (Sinha et al., 2020). Intelligent Transportation Systems (ITS) play a pivotal role in revolutionizing the transportation industry by employing traffic data for effective decision-making and traffic control (Wang et al., 2019). ITS encompasses various components, including traffic forecasting (or estimation), optimization techniques, and real-time information dissemination to improve traffic conditions and minimize travel delays (Alam et al., 2017). By providing drivers with current and projected traffic conditions, ITS enables informed route planning, considering potential delays and travel times for different routes within a city (Sinha et al., 2020). Consequently, ITS employs diverse prediction and classification models for traffic forecasting and management.

Traffic prediction has been a subject of extensive research since the late 1970s (Vázquez et al., 2020), while the identification and classification of traffic patterns are crucial for modern traffic management facilitated by ITS (Krishnakumari et al., 2017). The existing literature encompasses numerous studies focusing on traffic flow estimation, prediction, and classification, with researchers striving to develop enhanced control strategies to mitigate the escalating traffic issues of the past few decades (Wang et al., 2019, de Medrano & Aznarte, 2020, Müngen & Çetın Tas, 2021, Fitters et al., 2021, Wang & Thulasiraman, 2019, Ji et al., 2020).

Hence, several important concepts arise. Creating and predicting general traffic indicators, such as traffic flow, density, and mean speed, is crucial for effective traffic control and congestion prevention (Mena-Oreja & Gozalvez, 2021). Traffic flow represents the number of vehicles passing through a reference point per unit of time, while traffic density refers to the number of vehicles in a specific road section at a given moment. Mean speed indicates the average speed of vehicles in a particular road section (Mena-Oreja & Gozalvez, 2021). These indicators serve as the foundation for predicting traffic flow and facilitating traffic control. Prediction depends on factors like the specific road and time, considering spatiotemporal aspects. Traffic patterns vary during weekdays and weekends due to factors such as work schedules, school days, weather conditions, holidays, road networks, and public transportation quality. Accurate traffic flow prediction involves capturing and utilizing these patterns.

Moreover, prediction methods are utilized for both short-term and long-term traffic flow forecasting. Short-term prediction aids in route

* Corresponding author.
*E-mail address:* haidos@ciencias.ulisboa.pt (H. Aidos).

https://doi.org/10.1016/j.iswa.2023.200268
Received 12 May 2023; Received in revised form 5 August 2023; Accepted 10 August 2023
Available online 18 August 2023

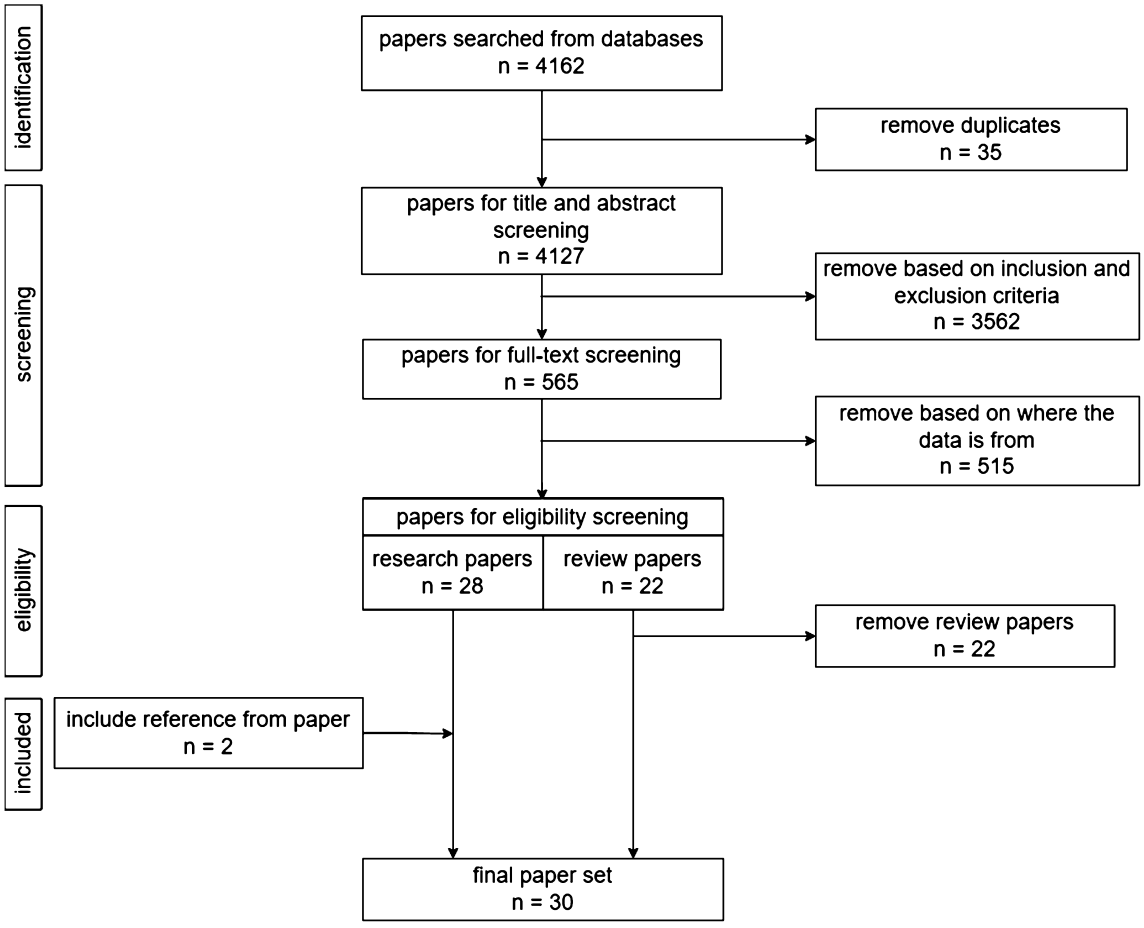


**Fig. 1.** Flowchart of the paper screening procedure.

management, trip duration estimation, and traffic signal synchronization, improving traffic conditions and congestion avoidance. Machine learning techniques, including parametric models, deep learning models, and genetic programming, are commonly employed for short-term traffic flow prediction. Long-term prediction focuses on forecasting traffic flow for the following day or days. Additionally, classification methods also play a crucial role in traffic management by categorizing roads or road segments based on their traffic conditions, allowing the determination of congestion levels. Classification techniques involve clustering and categorizing roads with similar characteristics and congestion levels. The number of congestion categories can vary, ranging from binary classifications (e.g., congested and not-congested) to more detailed classifications with multiple congestion levels.

Therefore, accurate traffic prediction and classification are essential for efficient traffic management and control. However, existing surveys and reviews predominantly concentrate on specific types of models. For example, Shi et al. (2019) compare only hybrid deep learning models for traffic flow prediction, while Wang (2021) review graph neural networks and compare them with Convolutional Neural Networks for traffic forecasting. In contrast, by describing and analyzing state-of-the-art models utilized for traffic prediction and classification in Europe over the past five years, this paper provides a comprehensive understanding of the whole process involved in traffic prediction and classification. This encompasses the data types used for traffic prediction and classification, data preprocessing techniques, prediction methods such as parametric models, deep learning models, and genetic programming, as well as classification models including clustering and classification approaches. Furthermore, we discuss the evaluation metrics employed to assess the performance of prediction and classification models. By adopting a broader perspective, the present study fills a gap in the existing literature that mainly focuses on specific model types.

The structure of the paper is as follows: Section 2 outlines the methodology employed to identify relevant articles about traffic prediction and classification. Section 3 presents the results of the literature search conducted in Section 2 regarding 5 different scopes: Type of data, Data preprocessing strategies, prediction and classification tasks, and performance evaluation metrics. Section 4 provides a comprehensive discussion of each of these scopes and section 5 presents the main conclusions drawn from the study.

## 2. Method

The purpose of this research is to give an overview and discuss the several methods used in the last five years to tackle the problem of traffic prediction and classification. In order to achieve this we followed a screening and selection procedure based on 3 steps: identification, screening and eligibility. An overview of the method is shown in Fig. 1.

### 2.1. Identification

In order to find all relevant studies in the identification step, we first organised keywords into two different groups, those related to traffic and those related to traffic indicators and machine learning. We identified different synonyms for each group (as can be seen in detail in Table A.4) which were used to search for papers in four technological libraries: *ACM Digital Library*, *Web of Science*, *IEEE Xplore* and *Scopus*. The keywords search field covered only the title. At the end of this step, duplicates were removed.

**Table 1**
Summary of the type of data used for prediction and classification of traffic flow. For each specific type of data, the table reports references, and the percentage of papers (relative to the total number of papers) using that specific type of data.

| Type of Data | References | % of studies |
|---|---|---|
| Historical Data | Wang et al. (2019), Alam et al. (2017), Krishnakumari et al. (2017), de Medrano and Aznarte (2020), Müngen and Çetın Tas (2021), Fitters et al. (2021), Wang and Thulasiraman (2019), Ji et al. (2020), Priambodo and Ahmad (2018), Izhar et al. (2020), Culita et al. (2020), Agafonov (2020), Mystakidis and Tjortjis (2020), Di et al. (2019), Silva and Martins (2020), Loumiotis et al. (2018), Kunde et al. (2017), Chu et al. (2021), Laharotte et al. (2017), Toshniwal et al. (2020), Spławińska (2017), Ekárt et al. (2020), Offor et al. (2019) | 76.7% |
| Simulator Data | Vázquez et al. (2020), Mena-Oreja and Gozalvez (2021), Zambrano-Martinez et al. (2017), Offor et al. (2019), Sarlas and Kouvelas (2019) | 16.7% |
| Real-time Data | Sinha et al. (2020), More et al. (2016), Loumiotis et al. (2018), Kalamaras et al. (2018) | 13.3% |
| Floating Car Data | Mena-Oreja and Gozalvez (2021), Laharotte et al. (2017), Vázquez et al. (2020) | 10.0% |

### 2.2. Screening

The identification step resulted in 4127 papers; however, many of them were not relevant to our work (not related to traffic data, not related to the Computer Science field, etc.). Therefore for the screening step, we removed papers based on inclusion and exclusion criteria defined by the following criteria:

c1: Only related to traffic data. For this, we defined a set of keywords related to not relevant areas of study (e.g., network traffic) and used those to refine the search queries (all the keywords can be seen in Table A.5).
c2: Published in the last five years (2017 to 2021).
c3: Final and written in English.
c4: Related to subject areas relevant to our work such as Computer Science, Engineering, Mathematics, Decision Sciences, and Social Sciences, among a few others.
c5: Uses traffic data from Europe.
c6: Only sensor data from stationary sensors around urban centres or traffic-congested areas, such as freeways.

Applying the first four criteria was made using academic libraries filters and applied over title and abstract, resulting in 565 papers. These were fully screened and as a result of using the two last criteria (c5 and c6), 50 papers resulted. The last criterion eliminated articles that were based on video-captured data, since our main focus was on sensor data which typically comes in the format of tabular data.

### 2.3. Eligibility

In the last step, the resulting papers were separated between research and review articles, where 22 were selected as reviews, systematic reviews and surveys, and 28 as research articles. After analysing all the referenced papers in the resulting 28, we manually added two more that we found relevant to our research, resulting in a set of 30 papers (with 21 being relevant to prediction models, and 9 to classification models).

## 3. Results

The most relevant research found by the methodology adopted in this manuscript will provide an understanding of traffic prediction and classification methods found in the literature. Thus, this section is divided into several main relevant topics: types of data, data preprocessing techniques, prediction methods, classification methods, and performance metrics to analyse and compare results.

### 3.1. Type of data

Table 1 presents an overview of the type of data used in each of the analysed papers. As can be seen, there are several types of data that can be used to classify and predict traffic flow: historical data, real-time data, simulation-generated data, floating car data or video footage of traffic. Nowadays, video captured footage is being used more and more due to advances in deep learning methods, computer vision, and autonomous driving. Although worth mentioning, we eliminated any article that used video-captured data, as explained in the methodology section 2.

From the studies found by our query, historical data is the type of data most used in the literature being identified in around 77% of the studies. Simulator data was identified in only around 17% of the studies and real-time data was identified in around 13% of the studies. Finally, our query identified 10% of the studies that use floating car data.

Historical data refers to data collected over a month or several months in a city or on some roads in a city. Usually, this data is provided by the government and collected by stationary sensors around the city. This type of data can be used to predict and classify data in the short and long term. For instance, we can use data from the past week to predict traffic flow on the next day or use the data from one hour ago to predict traffic flow in the next hour. According to Culita et al. (2020), this type of data can be used as historical and real-time data.

Real-time data is data collected, as well as historical data, from sensors located in a city in real-time and used to predict and classify traffic flow for the short term. Meaning that we can only make good predictions or classifications of traffic in the next few minutes or hours.

Another type of data is data generated with a simulator, which is a widely used type of data to complement historical data when there are few historical data and to better train the models used. The simulators generate traffic data, based on historical data, according to the user specifications. Some of the simulators mentioned in the literature are Aimsun (Vázquez et al., 2020) and SUMO (Zambrano-Martinez et al., 2017).

And finally, Floating Car Data (FCD), which is typically time-stamped geolocalization and speed data directly collected by moving vehicles, in contrast to traditional traffic data collected at a fixed location by a stationary device or observer. Modern vehicles are connected to a network and can provide this type of data. Furthermore, related data can also be generated using a simulator (Vázquez et al., 2020).

### 3.2. Data preprocessing strategies

Table 2 presents an overview of the data preprocessing techniques used in the papers identified by our query. As can be seen, 30% of the studies were identified as using techniques to handle missing values, either by imputation or by removal, and to handle outliers, and 30% of the studies presented the aggregate of the data in time intervals, such as aggregate average traffic speed over a 10-minute interval (e.g., Agafonov (2020)). Also, data normalization was identified in around 27% of the studies, while feature selection and extraction techniques were identified in around 17% and 7% of the studies, respectively. Finally, data discretization and eliminating redundancy from the dataset were identified in 3.3% of the studies.

The quality of the results presented by a prediction or a classification method depends largely on the type and quality of data preprocessing. If we want to obtain good results, we need to perform a good preprocessing of the raw data. According to Table 2, most of the preprocessing

**Table 2**
Categorization of the data preprocessing techniques found in the literature. For each specific preprocessing category, the table reports the type of task (prediction (P), classification (C), or both) where the preprocessing category was identified, the corresponding references, and the percentage of papers (relative to the total number of papers) using that specific preprocessing.

| Preprocessing Techniques | Type of task | References | % of studies |
|---|---|---|---|
| Handling missing values and outliers | both | Sinha et al. (2020), Vázquez et al. (2020), de Medrano and Aznarte (2020), Fitters et al. (2021), Agafonov (2020), Kunde et al. (2017), Toshniwal et al. (2020), Ekárt et al. (2020), Kalamaras et al. (2018) | 30.0% |
| Aggregation of data in time intervals | both | Agafonov (2020), de Medrano and Aznarte (2020), Fitters et al. (2021), Kunde et al. (2017), Mystakidis and Tjortjis (2020), Kalamaras et al. (2018), Mena-Oreja and Gozalvez (2021), Spławińska (2017), Toshniwal et al. (2020) | 30.0% |
| Data normalization | both | de Medrano and Aznarte (2020), Ji et al. (2020), Wang and Thulasiraman (2019), More et al. (2016), Izhar et al. (2020), Agafonov (2020), Kunde et al. (2017), Chu et al. (2021) | 26.7% |
| Feature Selection | both | Alam et al. (2017), Fitters et al. (2021), Mystakidis and Tjortjis (2020), Toshniwal et al. (2020), Izhar et al. (2020) | 16.7% |
| Feature Extraction | C | Krishnakumari et al. (2017), Zambrano-Martinez et al. (2017) | 6.7% |
| Data discretization | C | Mystakidis and Tjortjis (2020) | 3.3% |
| Eliminate redundancy | P | Di et al. (2019) | 3.3% |

categories appear in both prediction and classification tasks of traffic flow. However, feature extraction and data discretization appear only in studies related to the classification of traffic flow, while eliminating redundancy appears only in the prediction task.

In particular, in general, almost every dataset concerning traffic data contains a large number of missing values in some of the features (e.g., Kalamaras et al. (2018), Agafonov (2020), Sinha et al. (2020)). This may occur because of a malfunction of a data collector, the bad reading of a situation, or a sensor that is obstructed and cannot obtain a reading. Therefore, to eliminate these missing values, a decision needs to be made on how these data will be handled. For example, if there are many missing values in one feature, most of the time, the feature is eliminated and no longer considered. But there are other ways to handle missing values, one can fill in the missing values with some imputation technique, such as the median value of all the measured data in the last measuring window.

Another problem concerning the raw data set is that sometimes a value is wrongly calculated, the so-called outliers, and needs to be identified and replaced or removed (de Medrano & Aznarte, 2020, Kalamaras et al., 2018). For example, the speed of a vehicle cannot be negative. Just like what is done for the missing values, the wrong values can be either removed or replaced with the median value of all the measured data for that feature.

In both tasks, data normalization is an important step to improve performance and training stability of machine learning algorithms by ensuring that data presented in different scales are in the same scale, without distorting differences in the range of values (e.g., Chu et al. (2021), Ji et al. (2020), More et al. (2016)). Several normalization techniques can be applied to data, and the most common ones are linear scaling or z-score. The first one consists in convert the data to the range 0 and 1 (or sometimes -1 to +1), and it should be used if the data is approximately uniformly distributed across the range. While the latter scales the data to ensure that the feature distribution has a mean 0 and standard deviation 1, and it is useful when there are a few outliers, but not so extreme, in the feature distribution.

As explained in Section 3.1, traffic flow datasets are usually from stationary sensors around the city and are collected over a time period. All the information gathered from those sensors needs to be re-arranged to be fed to machine learning algorithms (either for prediction or classification tasks). Thus, it is typical to aggregate the information of the variables collected by the sensors in time intervals, for instance, Kalamaras et al. (2018) aggregated data in 5 minutes time intervals to forecast traffic speeds, while Toshniwal et al. (2020) aggregated the number of vehicles of the raw data in 5, 15, 30 or 60 minutes intervals to identify the factors that affect the traffic flow patterns in an urban area.

Moreover, when the data have a lot of features, feature selection can be performed to remove features that are not relevant to the work, allowing to reduce the computational cost of modelling and improving the performance of the model. In the context of traffic flow, feature selection can be found in the prediction task (e.g., Alam et al. (2017), Fitters et al. (2021)) as well as in the classification task (e.g., Mystakidis and Tjortjis (2020), Toshniwal et al. (2020)). In the literature it is possible to encounter several techniques for feature selection, being the most common ones based on filters and wrappers methods. The filters methods consist of selecting features based on statistics measures, such as information gain or chi-square test, which means the selection is independent of the learning algorithm to be used. While wrapper methods consist of selecting the features by considering a search problem, where different combinations of features are made, evaluated and compared with other combinations. The most common wrapper methods are forward selection (features are added iteratively to improve the performance of the learning algorithm) and backward elimination (features are eliminated iteratively, starting by the least significant, until the performance of the learning algorithm does not improve).

Many other preprocessing techniques were found in the literature, for instance, feature extraction techniques (Zambrano-Martinez et al., 2017) and data discretization (Mystakidis & Tjortjis, 2020) were found in classification tasks, and random undersampling (Izhar et al., 2020) to balance the data and remove bias in classification results. Some other preprocessing techniques found were more specific to traffic flow prediction tasks, such as organizing traffic variables (time mean speed, occupancy and mean length of the vehicles) into images with three channels (Mena-Oreja & Gozalvez, 2021) or splitting each city into regions (Ji et al., 2020).

Finally, data used with prediction and classification methods should be divided into train, test and sometimes validation sets. For prediction tasks, the data should be split to ensure the time dependency, so typically the first hours, days or months of the collected data are used to predict the following desired time period. While, for classification tasks, the data is typically divided into 80% for training and 20% for testing, or a cross-validation scheme is employed. This is valid if the classification task does not use clustering methods, otherwise, the split of the data may not be required since we do not have prior information on the data and hence, we are employing unsupervised methods. According to Appendix B, some papers do not provide such information which makes it difficult to assess the reliability and accuracy of the results.

### 3.3. Prediction task

In this section, we will discuss the different methods used for predicting traffic flow, which can be broadly categorized as deep learning models, parametric models and genetic programming. According to

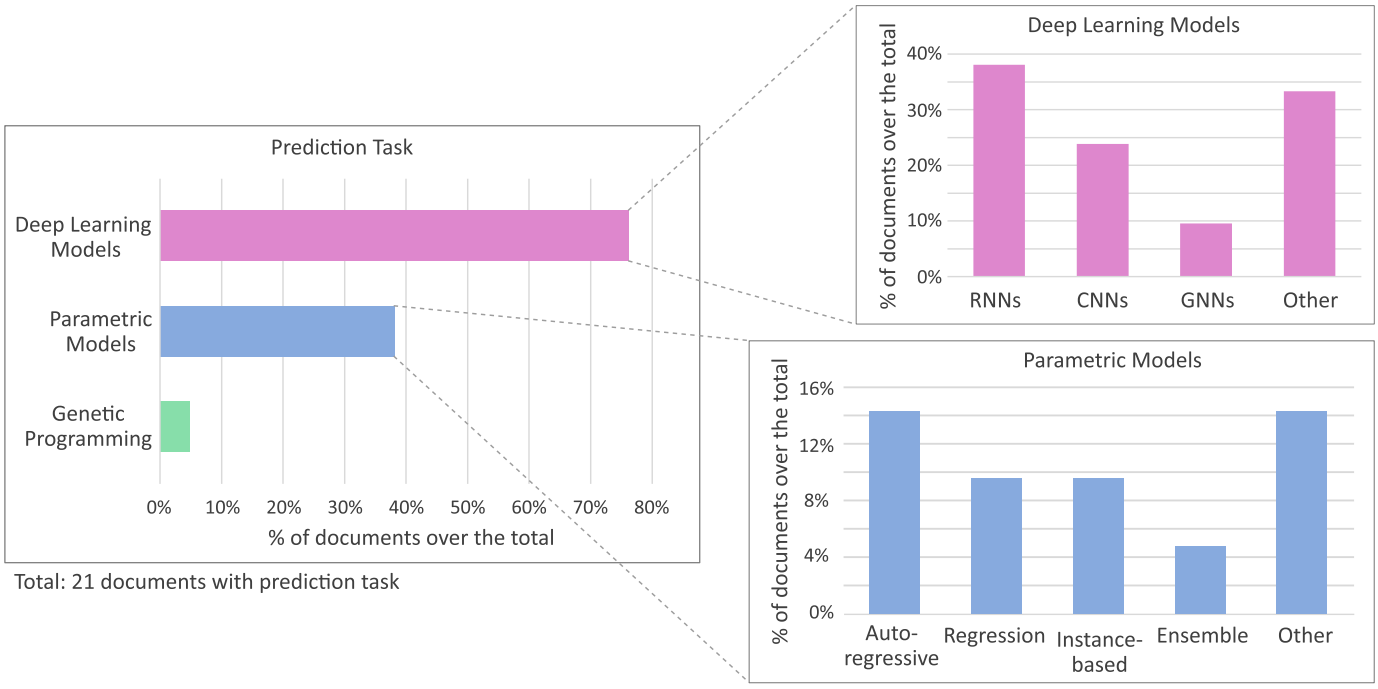


**Fig. 2.** Percentage of documents focused on prediction task categorized according to the type of model used (some documents can appear in more than one sub-category). RNNs: Recurrent Neural Networks, CNNs: Convolutional Neural Networks, GNNs: Graph Neural Networks.

Fig. 2, out of the 21 documents identified in the literature as doing prediction of traffic flow, 76% used deep learning models and less than 40% used parametric models. Only one study (around 5%) was identified as using genetic programming to predict traffic flow. Detailed information about the references that belong to each category of the prediction task can be found in Appendix B. In the following sections, we will present details of each one of these categories for the prediction task, including references and a brief explanation of the models.

#### 3.3.1. Deep learning models

The ability of deep learning models to capture complex patterns and relationships in data has led to their widespread adoption and popularity in traffic flow prediction. Several types of deep learning models have been identified in the literature, namely Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Graph Neural Network (GNN). According to Fig. 2, out of the 21 documents identified as prediction, 38% of the documents use RRNs, almost 24% of the documents use CNNs, 9.5% of the documents use GNNs, and 33% use other types of architectures, such as autoencoders.

*A. Recurrent Neural Network (RNN)* The most used methods usually utilize Recurrent Neural Networks (RNN) and their variants, as in Long-Short Term Memory (LSTM), to extract spatial relationships from the whole city by modelling citywide traffic. For instance, Wang and Thulasiraman (2019) used LSTM to predict traffic flow and speed. The authors used three prediction designs: one-to-one (train the model with the temporal information of one road to predict the traffic flow in the same road), many-to-one (train the model with temporal information of several roads to predict one road), and many-to-many (train the model with temporal information of several roads and predict multiple roads). On the other hand, Vázquez et al. (2020) implemented four different Deep Learning methods (LSTM, Gated Recurrent Unit (GRU), Spatiotemporal Recurrent Convolutional Network (SRCN) and a High-Order Graph Convolutional Long Short-Term Memory (HGC-LSTM)) to perform traffic forecasting in urban contexts, using floating car data to predict the average speed of the network road sections. In addition to using statistical models, Culita et al. (2020) used LSTM to predict traffic to avoid traffic congestion by adjusting the phase duration of the traffic lights that control each crossroad according to the real traffic conditions. Fitters et al. (2021) proposed an Outlier Enriched Long Short Term Memory (OE-LSTM) model to predict traffic flow, which employs a multi-step framework to detect outliers in the traffic flow and uses these outliers to learn the spatio-temporal correlations between different locations in the traffic network. Finally, Chu et al. (2021) used LSTM, GRU and Stacked Auto-Encoders (SAEs) to predict traffic flow in Finland.

*B. Convolutional Neural Networks* Convolutional neural networks (CNN) are typically used to process images and assign importance to different aspects of an image in order to differentiate one from the other. However, one can apply CNNs in different contexts as long as the input is organized in the format of an image. For instance, de Medrano and Aznarte (2020) used the Convo-Recurrent Attentional Neural Network (CRANN) model, which combines neural modules (temporal, spatial, and dense) to exploit the various components identified in a spatio-temporal series: seasonality, trend, inertia and spatial relations. Also, Mena-Oreja and Gozalvez (2021) proposed an error-recurrent convolutional neural network (eRCNN) to predict in the short term the three fundamental traffic variables (traffic density, traffic flow, and space mean speed) using Floating Car Data (FCD). The FCD was organized to be an image with different channels (like an RGB image). On the other hand, Di et al. (2019) proposed a spatiotemporal traffic prediction model, namely CPM-ConvLSTM, consisting of three steps (congestion propagation pattern graph construction, spatial matrix construction, and congestion level prediction) to make a short-term prediction of the congestion level for each segment of the road.

*C. Graph Neural Networks* Regarding non-Euclidean structured data, such as spatial networks, some studies use Graph Convolution Networks (GCNs) to capture spatial patterns. For instance, Agafonov (2020) used a Graph Convolutional Neural Network for traffic flow prediction taking into account daily and weekly patterns of traffic flow distributions. While Vázquez et al. (2020) used a High-Order Graph Convolutional Long Short-Term Memory Neural Network (HGC-LSTM), which applies a CNN to the network graph encoded as a matrix.

*D. Other Neural Networks* Besides CNNs, GCNs, and RNNs (and its variants), other neural networks were identified in the literature to tackle the problem of traffic flow prediction, such as General Regression Neural Network (GRNN), or Autoencoders. For instance, Loumiotis et al. (2018) used a GRNN for short-term traffic prediction, taking advantage of the fast learning ability and the convergence to the optimal surface. To demonstrate the appropriateness of GRNN, the authors compared the result with a Multi-Layer Perceptron (MLP) and a group method for data handling (GMDH) neural network. On the other hand, Ji et al. (2020) proposed a spatiotemporal deep learning model for the spatiotemporal potential energy fields (similar to water flow driven by the gravity field), consisting of a temporal component to model the temporal correlation and a spatial component to model spatial dependencies. While Müngen and Çetın Tas (2021) used a GRNN, which consists of three independent components with the same structure. Each component considers the recent time series, the daily-periodic time series, and the weekly-periodic time series with different patterns in traffic data, respectively. Also, Kunde et al. (2017) implemented a Feed-Forward Neural Network (FFNN) to predict traffic for future temporal horizons of 5, 10, 15, 30 and 45 minutes and used four different input settings (only historical values of the target sensor to predict a future value, only historical values from nearest neighbours excluding the target sensor, only historical values from nearest neighbours including the target sensor, and historical values from all sensors). On the other hand, Priambodo and Ahmad (2018) proposed a neural network that uses backpropagation to predict traffic speed by investigating the spatial and temporal correlation on neighbouring roads.

### 3.3.2. Parametric models

Parametric models, including both traditional statistical models and ensemble methods, have also been widely used in traffic flow prediction. We identified several categories of parametric models, according to Fig. 2, out of the 21 documents identified with the prediction task, 14% of the documents used auto-regressive models, 9.5% used regression, 9.5% used instance-based models, almost 5% of the documents used ensemble methods, and 14% used another parametric model, such as support vector machines or Bayesian Kriging model.

*A. Auto-regressive models* This kind of model focuses on predicting the future based on data from the past, typically it uses a linear combination of the past values of the variables. Extensive studies exist on traffic flow prediction, and the most used approaches were statistical models, such as Auto-Regressive Integrated Moving Average (ARIMA) (Sinha et al., 2020, Culita et al., 2020). More specifically, Culita et al. (2020) applied ARIMA and LSTM to predict real urban traffic from the city of Bucharest and compare and discuss the applicability of both methods to that scenario. While Sinha et al. (2020) applied ARIMA to predict future traffic density on specific roads of Slovenia. Also, Space-Time ARIMA (STARIMA) model was used by Kalamaras et al. (2018) to predict traffic flow by considering multiple traffic-related features from one road that can influence other roads of the network.

*B. Regression models* Regression models are used to model the relationship between the target variable (in this context, traffic flow) and one or more independent variables. There are several types of regression models, namely linear regression, logistic regression, and polynomial regression, among many others. For instance, Alam et al. (2017) applied Linear Regression, Sequential Minimal Optimisation (SMO) Regression, and M5 Base Regression Tree and Regression Trees to make predictions of traffic flow in the city of Porto, Portugal. Silva and Martins (2020) applied several different machine learning models, including multiple regression, to predict traffic in the city of Braga, Portugal, using data collected by a fleet of buses from the local public transport company.

*C. Instance-based models* This is a family of algorithms that do not try to model any distribution for the training data but instead compare new instances with the ones seen in the training data. The most used algorithm in this family is the k-nearest neighbour (k-NN). For instance, Silva and Martins (2020) used k-NN to make predictions on the city of Braga, Portugal, by averaging the predictions to the closest values for the input values.

*D. Ensemble methods* Ensemble methods are used to combine multiple models to produce a more robust and improved result. There are several ensemble methods in the literature, such as voting, averaging, and boosting, among others. In the context of traffic flow prediction, our search identified the application of random forest to data collected in the city of Braga (Silva & Martins, 2020).

*E. Other parametric methods* Finally, our search query identified other parametric models such as support vector machines and Kriging-based models. For instance, Müngen and Çetın Tas (2021) used Support Vector Machine models to predict traffic flow one hour ahead. While Offor et al. (2019) used a Linear Predictor (Kriging algorithm) and a Multi-Model Bayesian Kriging model to predict urban traffic, which is able to represent congested regions and interactions in upstream and downstream areas.

### 3.3.3. Genetic programming

Finally, Ekárt et al. (2020) proposed the GENetic Programming with Transfer LEarning (GENTLE) algorithm to generate Single Source Transfer Learning (SSTL) and Multiple Source Transfer Learning (MSTL) models. The resulting algorithm uses knowledge from other road segments to predict vehicle flow through a junction where traffic data are unavailable.

## 3.4. Classification task

According to Fig. 3, out of 9 documents identified in the literature as doing classification of traffic flow, 55% of studies used clustering methods, while the remaining 45% used classification techniques. Detailed information about the references that belong to each category of the classification task can be found in Appendix B. In the following sections, we will present details, references and brief explanations of the models for each one of these categories.

### 3.4.1. Clustering methods

Clustering methods are used to capture the structure of the data by grouping data points with similar characteristics. These methods can be used to classify each data point into a specific group, given a set of data points. These models evaluate each part of the city or road in terms of traffic and at what hours and days the roads are congested or not congested. This can be done by classifying each situation into different traffic levels, as many as necessary. Several types of clustering methods have been identified, namely partitional, hierarchical, and other types. According to Fig. 3, out of 9 documents identified with classification task, more than 44% of the documents used partitional methods (e.g., k-means), 33% used hierarchical clustering (e.g., average linkage), and 22% used another type of clustering methods, such as density-based methods. For instance, Zambrano-Martinez et al. (2017) used the k-means clustering algorithm to divide all road segments into four clusters (categories) of traffic level, which were: increasing traffic, decreasing traffic, constant traffic, and unique traffic. Laharotte et al. (2017) used a generative probabilistic model, called Latent Dirichlet Allocation (LDA), associated with a clustering indicator, named perplexity, to categorize as recurrent or non-recurrent state results directly from the confrontation of the perplexity to its associated threshold (perplexity values above the threshold are classified as non-recurrent).

Moreover, Toshniwal et al. (2020) used the DBSCAN algorithm to classify and divide days into two clusters, i.e., weekdays and weekends. Furthermore, the authors used Partition Around Medoids (PAM) and Agglomerative Nesting (AGNES) to cluster the working day profiles for

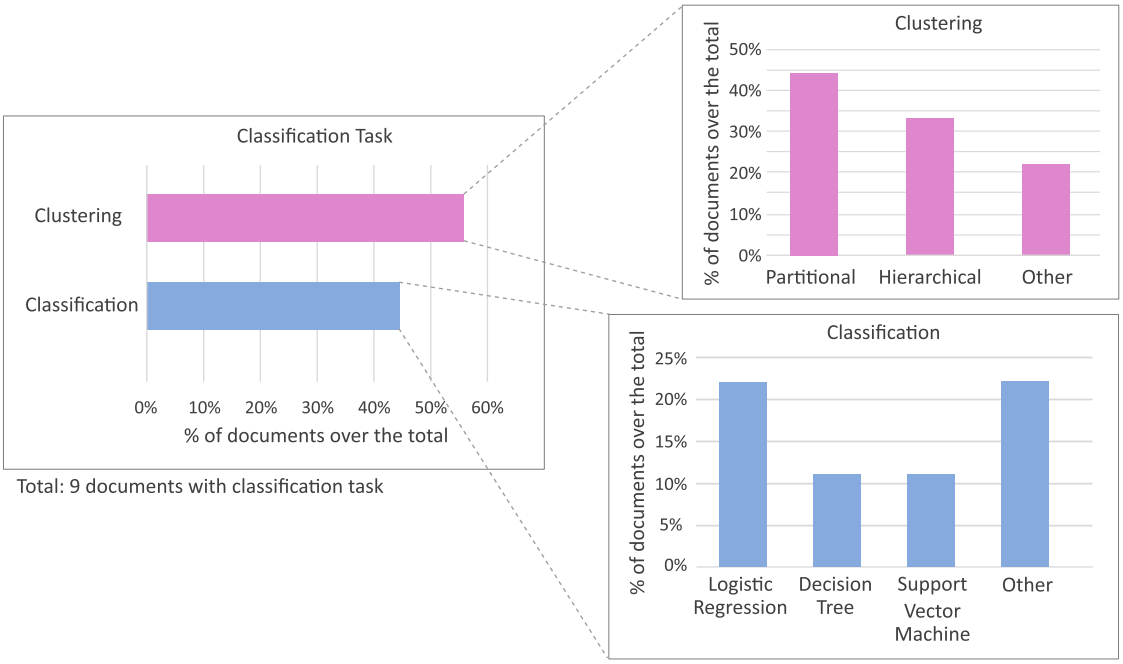


**Fig. 3.** Percentage of documents focused on classification task categorized according to the type of model used (some documents can appear in more than one sub-category).

roads remaining after preprocessing, obtaining six clusters in which all roads belonging to the same cluster have the same characteristics in terms of traffic. Spławińska (2017) used an agglomerative and k-means clustering algorithm to determine the new conversion factors suitable for freeways and expressways and directional analysis in heavy vehicle groups. Finally, Wang et al. (2019) used the affinity propagation (AP) clustering algorithm to provide drivers with information about the roads for both peak and non-peak hours.

#### 3.4.2. Classification methods

Classification methods allow the separation of data into a set of required classes by training the models with input data points together with the class label information. In Fig. 3, out of 9 documents identified with classification task, 22% used logistic regression to classify traffic flow, 3% used decision tree, and another 3% used support vector machine, while almost 7% used another type of classification algorithm, such as multinomial naive Bayes. For instance, Mystakidis and Tjortjis (2020) used decision tree classifiers to classify traffic into three categories: 0, which means that the road has low congestion, 1 for medium congestion, and 2 for high congestion. They also used these classification methods to predict the future traffic level. Izhar et al. (2020) used two classifiers: Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB). The authors used a hybrid feature-based label generation to predict traffic congestion. Based on the average speed and the number of vehicles, these two features can determine whether there is congestion or not. Also, Sarlas and Kouvelas (2019) created and computed six road traffic indicators to divide road intersections into two categories: high importance (critical set of nodes) and the complement set. In addition to the two indicators constructed based on the dynamic traffic data analysis, four graph theory indicators were built. The authors proposed a way to combine all these indicators and derived a generic ranking of nodes based on their criticality. The ranking aimed at classifying the signalized intersections into two groups, the high importance (critical set of nodes) and the complement set.

### 3.5. Performance evaluation metrics

After applying the suitable preprocessing techniques to the data, dividing the dataset into train and test sets (in some cases, train, test, and validation sets), and applying the methods explained in the previous subsection, it is crucial to evaluate the results and the performance of the methods. Hence, the results obtained by the approaches are compared to the real value of the data. Some state-of-the-art metrics are used to do so, but the metrics are different for prediction and classification methods.

From Table 3 we can see an summary of the performance evaluation metrics identified in the literature. For prediction models, evaluation metrics measure the error between predicted observations and real values. The most frequently used metrics are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), with 40%, 40% and 26.7% of the studies using the metric, respectively.

In detail, consider a set of real observations $Y = \{y_1, y_2, ..., y_n\}$ and a prediction model $\Phi$. The model $\Phi$ predicts a set of observations $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$, and we need to assess the performance of that model by comparing the observed values with the predicted ones. Thus, several metrics can be used for this purpose as presented in Table 3, and we will explain in more detail the most widely used in the literature. Hence, RMSE is a metric used to measure the average magnitude of the error, and is computed by:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \quad (1)$$

On the other hand, MAE measures the overall mean deviation of the predicted value and is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (2)$$

While MAPE is a measure of the prediction accuracy of a model and is given by:

**Table 3**
Summary of the evaluation metrics used to evaluate prediction, classification, and clustering methods. For each specific metric it is reported the references, and the percentage of papers (relative to the total number of papers) using that specific metric.

| Type | Evaluation Metrics | References | % of papers |
|---|---|---|---|
| **Prediction** | Root Mean Squared Error (RMSE) | Alam et al. (2017), Vázquez et al. (2020), de Medrano and Aznarte (2020), Ji et al. (2020), Priambodo and Ahmad (2018), Mena-Oreja and Gozalvez (2021), Culita et al. (2020), Agafonov (2020), Chu et al. (2021), Ekárt et al. (2020), Offor et al. (2019), Kalamaras et al. (2018) | 40.0% |
| | Mean Absolute Error (MAE) | Wang and Thulasiraman (2019), Alam et al. (2017), Vázquez et al. (2020), Müngen and Çetın Tas (2021), Mena-Oreja and Gozalvez (2021), Culita et al. (2020), Agafonov (2020), Di et al. (2019), Silva and Martins (2020), Loumiotis et al. (2018), Kunde et al. (2017), Chu et al. (2021) | 40.0% |
| | Mean Absolute Percentage Error (MAPE) | Müngen and Çetın Tas (2021), Fitters et al. (2021), Priambodo and Ahmad (2018), Mena-Oreja and Gozalvez (2021), Loumiotis et al. (2018), Chu et al. (2021), Agafonov (2020), Spławińska (2017) | 26.7% |
| | Mean Squared Error (MSE) | Wang and Thulasiraman (2019), Di et al. (2019), Silva and Martins (2020), More et al. (2016) | 13.3% |
| | Coefficient of determination $R^2$ | Wang and Thulasiraman (2019), Ji et al. (2020), Chu et al. (2021), Silva and Martins (2020) | 13.3% |
| | Explained Variance | Wang and Thulasiraman (2019), Chu et al. (2021), Silva and Martins (2020) | 10.0% |
| | Relative Absolute Error (RAE) | Alam et al. (2017) | 3.3% |
| | Root Relative Squared Error (RRSE) | Alam et al. (2017) | 3.3% |
| | Weighted Mean Absolute Percentage Error (WMAPE) | de Medrano and Aznarte (2020) | 3.3% |
| | Mean Absolute Deviation (MAD) | Priambodo and Ahmad (2018) | 3.3% |
| | Normalized Root Mean Squared Error (NRMSE) | Offor et al. (2019) | 3.3% |
| **Classification** | Accuracy | Krishnakumari et al. (2017), Izhar et al. (2020), Mystakidis and Tjortjis (2020), Loumiotis et al. (2018) | 13.3% |
| | Precision and Recall | Izhar et al. (2020) | 3.3% |
| | F1-score | Izhar et al. (2020) | 3.3% |
| | Detection Rate (DR) | Laharotte et al. (2017) | 3.3% |
| | False Alarm Rate (FAR) | Laharotte et al. (2017) | 3.3% |
| | Good Classification Rate (GCR) | Laharotte et al. (2017) | 3.3% |
| **Clustering** | Silhouette Score | Wang et al. (2019), Toshniwal et al. (2020) | 6.7% |
| | Dunn Index | Toshniwal et al. (2020) | 3.3% |

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%. \tag{3}$$

For classification models, the most widely used metrics in the literature are accuracy, precision and recall, and F1-score. However, in the context of traffic flow, the most frequently used evaluation metric is accuracy, with 13.3% of the studies using that metric, and precision and recall, and F1-measure only appear in one study (Izhar et al., 2020) (see Table 3). The prevalence of these types of metrics is small in this survey since our query only identified 9 out of 30 studies with the classification task.

In more detail, accuracy is the fraction of observations correctly classified by a model. While precision is the proportion of positive observations given by a model that are true positive observations, and recall is the proportion of real positive observations that are true positive observations given by a model, i.e., the ability of the classifier to find all the positive samples. Finally, the F1-score is the harmonic mean of precision and recall.

For clustering models, the most frequently used evaluation metrics are the silhouette score and Dunn index. According to Table 3, silhouette score is used in 6.7% of the studies found in our search, while the Dunn index is used in 3.3% of the studies. In detail, consider a set of data points $X = \{x_1, ..., x_n\}$ and a clustering model $\Gamma$. The model groups the data points with similar characteristics into clusters, and we need to assess the goodness of the model. Hence, the silhouette score measures the goodness of a clustering algorithm by considering how compact (intra-cluster distance) and separated (inter-cluster distance) the clusters are, and it is computed as the average of the silhouette index of each data point $x_i$ given by:

$$s(x_i) = \frac{b_i - a_i}{\max\{b_i, a_i\}}, \tag{4}$$

where $a_i$ is the average intra-cluster distance and $b_i$ is the average inter-cluster distance. On the other hand, the Dunn index identifies clusters that are compact and have small variance between the members of the cluster, and is given by:

$$D = \frac{\min_{1\le i\le c}\delta(x_i, x_j)}{\max_{1\le k\le c}\Delta(x_k)}, \tag{5}$$

where $\delta(x_i, x_j)$ is the inter-cluster distance between data points $x_i$ and $x_j$, and $\Delta(x_k)$ is the intra-cluster distance of a cluster.

## 4. Discussion

Based on our search and the related work found, we can identify the type of data used, prediction and classification models, the preprocessing techniques, and the performance evaluation metrics used in the state-of-the-art to classify and predict traffic flow. For each we now discuss findings and limitations. Additionally we can discuss on the limitations and future opportunities regarding this type of literature survey.

### 4.1. Type of data

Regarding the state-of-the-art presented in the referenced articles, we can see that the most used type of data is historical data, which is used to predict and classify traffic to obtain information and make decisions based on this information to better control and manage traffic in a particular city. But with the up-and-coming ITS, short-term traffic flow prediction and classification is becoming a crucial part of ITS and

is a widely investigated topic. To predict and classify traffic flow in the short term, the type of data used is real-time data or historical data used as real-time. Therefore, real-time data are becoming an essential type of data to control traffic in real-time better. For instance, better control traffic lights, make better suggestions on what route to take and better inform users of the traffic conditions and the expected traffic conditions in the next minutes or hours to better mitigate traffic congestion and traffic jams in large cities.

All the different classification and prediction methods can be applied to historical data. If the data were to be used in real-time, some of the more complex methods could not be used because when we use real-time data, we want good results at the same time as fast results. The computation time of all methods cannot be more than 5 or 10 minutes. This will imply that the amount of data used is not as big as the amount of data used when we want to make predictions and classifications in a larger time window. As discussed earlier, historical data can be used both as historical and real-time data, which is an enormous advantage of using historical data to create the best algorithm. For example, if we only consider the last couple of hours of traffic and make a prediction based only on that, we can provide real-time traffic information. Furthermore, we can also create hourly indicators based on all-week or monthly traffic.

Although Historical Data are the most used type of data, Floating car data (FCD) are the future. FCD is automatically collected from moving vehicles, but only modern cars can provide this data. One great advantage of this type of data is that, unlike stationary devices, such as traffic cameras, no additional hardware is required on the road network. FCD is used to determine the traffic speed on the road network. These data allow traffic congestion to be identified, travel times can be calculated, and traffic reports can be rapidly generated.

### 4.2. Data preprocessing strategies

Raw traffic data collected from real-world sources can be contaminated with errors and inconsistencies that can cause problems for machine learning models. As a result, preprocessing raw data is a crucial step in traffic flow prediction and classification. The accuracy and reliability of predictions depend on the quality of preprocessing performed on the data.

The main goal of data preprocessing is to convert raw data into a format that is clean and interpretable for algorithms. This involves several key steps, including changing variable data types, removing outliers and eliminating missing values from the dataset to improve results, and converting categorical and textual features into numerical values that can be better interpreted by machine learning models.

Despite the importance of data preprocessing, some research papers lack detailed descriptions of the techniques used to preprocess data, making it difficult to assess the reproducibility, reliability, and accuracy of their results. It is essential to carefully document preprocessing steps and ensure that data is cleaned and transformed accurately to obtain high-quality results in traffic flow prediction and classification.

### 4.3. Prediction and classification tasks

Traffic flow prediction has been extensively studied in the last few decades using statistical methods. However, deep learning-based approaches have become increasingly popular due to their improved accuracy and results. Despite their advantages, deep learning algorithms are black-box algorithms, making it difficult to interpret model predictions for decision-making in real-world transportation applications. The most commonly used deep learning-based model for traffic flow prediction is the LSTM network, which is well-suited for processing and predicting time-series data.

Regarding classification and clustering models, there is a lack of usage in traffic tasks. These models have not been widely used in the last 5 years, as predictive methods provide much more and better information than classifying traffic. While clustering and classifying traffic can inform users about current traffic conditions, predictive methods are necessary to anticipate and control traffic conditions in the future. Future research should investigate the development of hybrid models that combine predictive and clustering methods to provide a more comprehensive understanding of traffic patterns and trends. Additionally, new approaches should be explored to improve the interpretability of deep learning models for better decision-making in transportation applications.

### 4.4. Performance evaluation metrics

To assess the accuracy and validity of models, various performance evaluation metrics are employed. The purpose of these metrics is to determine whether the models are suitable predictors or classifiers and which model performs the best. One critical aspect to avoid in all models is overfitting, a condition that arises when a statistical model starts describing the random errors in the data instead of the relationships between variables. Overfitting is more likely to occur when the model is too complex, and it reduces the model's ability to generalize beyond the original dataset, leading to bias and making good predictions or classifications only for the training and test sets. Researchers should continue exploring methods to prevent overfitting, such as regularisation and cross-validation techniques, to increase the generalisability of the models to real-world scenarios.

### 4.5. Limitations and future opportunities

There are limitations that underscore the inherent challenges in conducting a comprehensive review of traffic flow prediction and classification literature. Predominantly, the emphasis on European data might have inadvertently omitted certain trends, patterns or techniques applicable in non-European regions. Geographic factors, infrastructure quality, population density, and traffic regulations all vary across regions, and might significantly affect the techniques developed or preferred in those areas. Expanding our scope to include non-European studies could also allow a more in-depth examination of data differentiation, which might reveal new avenues of investigation.

When considering the timeframe of our study, we acknowledge that an extension to the last 10 to 15 years might uncover different trends. Traffic technology and related prediction techniques have rapidly evolved over recent years, influenced by advances in artificial intelligence, machine learning, and sensor technology. A broader temporal scope could offer a clearer understanding of these developments and their implications for the field.

Lastly, our focus on sensor-acquired data could neglect the potential benefits and trends associated with other data types, such as those from images, videos, or sound. The recent advancements in fields like computer vision and auditory signal processing could significantly influence traffic flow prediction and classification techniques. However, the exploration of these alternative data types could lead us into a distinct, albeit related, research landscape that may require a different set of expertise and research approaches.

In essence, while our literature survey offers valuable insights into traffic flow prediction and classification, these limitations highlight the multifaceted and rapidly evolving nature of the field. Future studies could consider these aspects for a more extensive and diverse understanding of the subject matter.

## 5. Conclusions

This literature study has provided valuable insights into the field of traffic prediction and classification. General conclusions highlight the crucial need for correctly preprocessing the datasets before using machine learning models. Preprocessing techniques, including missing

value imputation, data normalisation, feature selection, and dataset reduction, are crucial in achieving reliable results. Neural networks, particularly LSTM and MLP, demonstrate superior accuracy compared to parametric models like ARIMA and Linear Regression. Combining models, such as CNN and LSTM, yields complex models such as CNN-LSTM. K-means, DBSCAN, and Agglomerative Clustering are the primary classification models for traffic flow. Performance evaluation metrics, such as MSE, RMSE, accuracy, precision, and F1 score, provide insight into model effectiveness. These models contribute to improved traffic management and informed decision making by governments.

This state-of-the-art analysis also highlighted the predominant use of historical data in predicting and classifying traffic, enabling informed decision-making for traffic management. However, the emergence of Intelligent Transportation Systems (ITS) emphasises the importance of short-term traffic flow prediction and the integration of real-time data. Real-time data, such as Floating Car Data (FCD), offer immense potential to improve traffic control and congestion mitigation, especially with the increasing availability of modern vehicles capable of providing such data without the need for additional infrastructure. While historical data exhibits advantages in its versatility for both historical and real-time use, the lack of detailed descriptions of data preprocessing techniques in some research papers poses challenges in assessing the reproducibility, reliability, and accuracy of results. Future studies should prioritise comprehensive documentation of preprocessing steps to ensure high-quality and reliable outcomes in traffic flow prediction and classification.

Moreover, deep learning-based approaches, particularly LSTM networks, have gained popularity because of their improved accuracy. However, their black-box nature limits interpretability for real-world decision-making in transportation applications. Future research should focus on improving the interpretability of deep learning models and exploring hybrid models that integrate predictive and clustering methods to gain a more holistic understanding of traffic patterns and trends. It is crucial for researchers to address issues related to overfitting and enhance the generalizability of models to real-world scenarios. Incorporating regularisation techniques and employing cross-validation methods can mitigate overfitting and improve the applicability of models in practical transportation settings.

Finally, while the inclusion of only European studies limited the scope of this study, it was necessary to account for significant cultural differences in traffic patterns. However, it is essential to acknowledge that a wealth of knowledge can be gained from studies conducted outside of Europe, and future research should aim to incorporate and compare methods used in a more diverse range of geographical contexts.

## List of acronyms

**ITS** Intelligent Transportation Systems
**FCD** Floating Car Data
**RNN** Recurrent Neural Network
**CNN** Convolutional Neural Network
**GNN** Graph Neural Network
**LSTM** Long-Short Term Memory
**GRU** Gated Recurrent Unit
**SAE** Stacked Auto-Encoder
**GCN** Graph Convolution Network
**MLP** Multi-Layer Perceptron
**GRNN** General Regression Neural Network
**ARIMA** Auto-Regressive Integrated Moving Average
**RMSE** Root Mean Squared Error
**MAE** Mean Absolute Error
**MAPE** Mean Absolute Percentage Error

## CRediT authorship contribution statement

**Bernardo Gomes:** Investigation, Methodology, Writing – original draft. **José Coelho:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Helena Aidos:** Conceptualization, Methodology, Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No data was used for the research described in the article.

## Acknowledgements

This work was supported by FCT through the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

## Appendix A. Method

**Table A.4**
Overview of the keywords used to search for relevant papers.

| Concept | Keywords |
|---|---|
| Traffic | "traffic" OR "urban traffic" OR "traffic flow" OR "traffic system" OR "traffic congestion" OR "traffic conditions" OR "traffic conflicts" OR "traffic density" OR "level of traffic stress" |
| Traffic Indicators and Machine Learning | "clustering" OR "cluster" OR "deep learning" OR "supervised learning" OR "unsupervised learning" OR "pattern mining" OR "data mining" OR "prediction" OR "classification" OR "forecasting" |

**Table A.5**
Overview of the keywords used to remove not relevant papers from the search.

| Concept | Keywords |
|---|---|
| NOTs | NOT "malware" AND NOT "network traffic" AND NOT "cloud computing" AND NOT "malicious" AND NOT "air traffic" AND NOT "cargo traffic" AND NOT "online traffic" AND NOT "internet traffic" AND NOT "botnet" AND NOT "ip flow" AND NOT "network flow" AND NOT "dns traffic" AND NOT "recognition" AND NOT "video traffic" AND NOT "streaming traffic" AND NOT "satellite network" AND NOT "video prediction" AND NOT "interconnect traffic" AND NOT "pavement" AND NOT "ip traffic" AND NOT "media traffic" AND NOT "object detection" AND NOT "encrypted traffic" AND NOT "bandwidth" AND NOT "traffic simulations" AND NOT "voice traffic" AND NOT "network management" AND NOT "datacenters" AND NOT "data center" AND NOT "self-driving cars" AND NOT "image segmentation" AND NOT "metro traffic" AND NOT "maritime traffic" AND NOT "traffic sign image" AND NOT "trajectory prediction" AND NOT "internet connections" AND NOT "image classification" AND NOT "classifying objects" AND NOT "collections of multimedia" AND NOT "anomaly detection" AND NOT "attack classification" AND NOT "cyber traffic" AND NOT "autonomous driving" AND NOT "traffic signs" AND NOT "sign identification" AND NOT "visual traffic" AND NOT "traffic signals" AND NOT "airspace" AND NOT "pedestrian" AND NOT "vehicle counting" AND NOT "anonymous traffic" AND NOT "traffic sound" AND NOT "object classification" AND NOT "mobile traffic" AND NOT "attack detection" AND NOT "vehicle classification" AND NOT "speed prediction" |

## Appendix B. Summary table

List of papers included in this survey (see Table B.6). For each paper, reported data include:

- the reference,

**Table B.6**
Summary table

| Citation | P or C? | Method Category | Type of Data | Dataset | Open Access? | Validation Scheme | pre-processing Techniques | Performance Metrics |
|---|---|---|---|---|---|---|---|---|
| Agafonov (2020) | P | DL | HD | 1760 road segments from Samara city, Russia, records for 60 days | - | 60% train + 20% control + 20% validation | MV, N, Agg | RMSE, MAE, MAPE |
| Alam et al. (2017) | P | PM | HD | data collected in Porto, Portugal, from 2013 to 2015 using 23 sensors every 5 min; using only 2014 | No (private to students of FEUP) | - | FS | RMSE, MAE, RAE, RRSE, CC |
| Chu et al. (2021) | P | DL | HD | Finish dataset with 4 lanes collected from January to August for 24 hours a day | - | split in train and test (no % provided) | N | RMSE, MAE, MAPE, $R^2$, EV |
| Culita et al. (2020) | P | DL, PM | HD | 3 road segments from Bucharest collected from 1 October to 31 January with 5 min time step | - | the last 5 or 10 samples from the whole measured data are the test set, the remaining is the train set | - | RMSE, MAE |
| de Medrano and Aznarte (2020) | P | DL | HD | traffic and weather data from 4 specific zones of Madrid from 2018 and 2019, collected with 30 sensors in each zone every 15 min | Yes | 10-cross-validation scheme without repetition for each zone | MV, N, Agg, O | RMSE, WMAPE, Bias |
| Di et al. (2019) | P | DL | HD | traffic congestion data from Helsinki, Finland, collected from 2018/09/01 to 2018/10/06 every 60 seconds corresponding to 553 road segments | Restricted | first 80% of the data (from the start date) as train set and the remaining as test set | ER | MSE, MAE |
| Ekárt et al. (2020) | P | GA | HD | traffic data from 4 junctions from Darmstadt, Germany, collected between 28th August to 1st October 2017, sampled every 15 min | Yes | first 3 weeks for training and final 2 weeks for testing | MV | RMSE |
| Fitters et al. (2021) | P | DL | HD | traffic data from intersections in city of The Hague, Netherlands, collected between 1st January 2015 to 31st May 2019 | Yes | - | MV, Agg, FS | MAPE |
| Izhar et al. (2020) | C | CLASS | HD | vehicle traffic datasets from the city of Aarhus, Denmark, collected in 2014 from February to June and August to September | Yes | 5-fold cross-validation with the method stratified sampling | N, FS, other | Acc, F1, Prec, Rec |
| Ji et al. (2020) | P | DL | HD | Dataset with trajectories of taxicab from 3 cities, Xi'an, Beijing and Porto, sampled every 3, 60, and 15 seconds respectively | Yes | previous 80% as training data and the rest as testing data | N, other | RMSE, $R^2$ |
| Kalamaras et al. (2018) | P | PM | RTD | Berlin dataset collected from 18/03/2012 to 31/03/2012 containing real vehicle speed measurements collected from several road points | Needs registration | train made with data from a specific day of the week and test data from the same day at another week | MV, Agg, O | RMSE |
| Krishnakumari et al. (2017) | C | CLASS | HD | Data from two heavily congested roads in The Netherlands collected in March 2015 | Yes | - | FE | Acc |
| Kunde et al. (2017) | P | DL | HD | data from city of Dresden, Germany, collected in July 2015 | Restricted | different offsets configurations to split the data in train and test | MV, N, Agg | MAE |
| Laharotte et al. (2017) | C | CLUS | HD, FCD | floating car data from city of Nantes, France, collected in September and November 2013 | - | one month to train and the last month to test | - | DR, FAR, GCR |
| Loumiotis et al. (2018) | P | DL | HD, RTD | data collected for 4 months by a vehicle detection system in Attica Tollway in Athens, Greece | - | 10-fold cross validation | other | MAE, MAPE, Acc |
| Mena-Oreja and Gozalvez (2021) | P | DL | SD, FCD | simulated data for a Spanish freeway between Alicante and Murcia for 9 full days of traffic | Yes | first seven days to train and the remaining two days for validation and testing (one day each) | Agg, other | RMSE, MAE, MAPE |
| More et al. (2016) | P | DL | RTD | data from Ireland road traffic control collected for five days | Yes | four days to train and one day to test | N | MSE |

(*continued on next page*)

**Table B.6** (*continued*)

| Citation | P or C? | Method Category | Type of Data | Dataset | Open Access? | Validation Scheme | pre-processing Techniques | Performance Metrics |
|---|---|---|---|---|---|---|---|---|
| Müngen and Çetın Tas (2021) | P | DL, PM | HD | traffic and weather data from Istanbul, Turkey, collected between 01/06/2020 and 01/01/2021 | Yes | 80% for training data and 20% for validation data | - | MAE, MAPE |
| Mystakidis and Tjortjis (2020) | C | CLASS | HD | traffic congestion data from the city of Thessaloniki, Greece, from August to October 2019 | Yes | 80% for training data and 20% for testing data | Agg, FS, D | Acc |
| Offor et al. (2019) | P | PM | HD, SD | real and simulated data from Santander, Spain | Yes | real data for training and simulated as ground truth for the prediction | - | RMSE, NRMSE |
| Priambodo and Ahmad (2018) | P | DL, PM | HD | traffic data from Aarhus, Denmark, collected between 13/02/2014 to 09/06/2014 | - | all data until 02/06/2014 as train data and 09/06/2014 as test data | other | RMSE, MAPE, MAD |
| Sarlas and Kouvelas (2019) | C | CLASS | SD | simulated data for the urban network of Barcelona, Spain, with the duration of 5 hours | - | - | - | Visualizations |
| Silva and Martins (2020) | P | DL, PM | HD | traffic data from a urban passenger transport company in Braga, Portugal | Restricted | - | - | MSE, MAE, $R^2$, EV, MeAE |
| Sinha et al. (2020) | P | PM | RTD | traffic data of 167 unique roads of the country Slovenia, collected for 7 days | Yes | - | MV | Visualizations |
| Spławińska (2017) | C | CLUS | HD | data from rural freeway and expressway of cross-sections of different regions of Poland, collected from 2010 to 2015 | Restricted | - | Agg | MAPE, Distance to centroids |
| Toshniwal et al. (2020) | C | CLUS | HD | traffic data from Aarhus, Denmark, collected between February 2014 to June 2014 | Restricted | - | MV, Agg, FS | Silh, Dunn |
| Vázquez et al. (2020) | P | DL | SD, FCD | simulated data for two urban networks of Barcelona, Spain (Camp Nou and Amara) | - | 5, 10 and 15 days of data to train and predicts 5, 10, 15, 20, 40 and 60 min | MV | RMSE, MAE |
| Wang and Thulasiraman (2019) | P | DL | HD | data set is collected from CityPulse from 01/03/2014 to 30/05/2014 | Yes | first 2 months for training and last month for testing | N | MSE, MAE, $R^2$, EV |
| Wang et al. (2019) | C | CLUS | HD | data set is collected from CityPulse from the city of Aarhus, Denmark, recorded every 5 min | Yes | - | - | Silh |
| Zambrano-Martinez et al. (2017) | C | CLUS | SD | simulated data for the city of Valencia | Restricted | - | FE | Visualizations |

- the aim of the paper (P = prediction; C = classification),
- categorization of the method (DL = deep learning, PM = parametric model, GA = genetic algorithm, CLASS = classification method, CLUS = clustering method),
- type of data (HD = historical data, SD = simulated data, RTD = real-time data, FCD = floating car data),
- a small description of the dataset,
- the type of data availability,
- validation scheme,
- preprocessing techniques (MV = handling missing values, N = data normalization, Agg = aggregation in time intervals, FS = feature selection, O = handling outliers, FE = feature extraction, D = data discretization, ER = eliminate redundancy, other = includes random undersampling or other task-specific technique),
- performance metrics (MSE = mean squared error, RMSE = root mean squared error, NRMSE = normalized root mean squared error, MAE = mean absolute error, MAPE = mean absolute percentage error, MAD = mean absolute deviation, MeAE = median absolute error, $R^2$ = coefficient of determination, EV = explained variance, CC = correlation coefficient, RAE = relative absolute error, RRSE = root relative squared error, WMAPE = weighted mean absolute percentage error, Acc = accuracy, F1 = F1-score, Prec = precision, Rec = recall, DR = detection rate, FAR = false alarm rate, GCR = good classification rate, Silh = silhouette score, Dunn = Dunn index).

## References


Agafonov, A. (2020). Traffic flow prediction using graph convolution neural networks. In *2020 10th international conference on information science and technology (ICIST)* (pp. 91–95).

Alam, I., Ahmed, M. F., Alam, M., Ulisses, J., Farid, D. M., Shatabda, S., & Rossetti, R. J. F. (2017). Pattern mining from historical traffic big data. In *2017 IEEE region 10 symposium (TENSYMP)* (pp. 1–5).

Chu, Q., Li, G., Zhou, R., & Ping, Z. (2021). Traffic flow prediction model based on LSTM with Finnish dataset. In *2021 6th international conference on intelligent computing and signal processing (ICSP)* (pp. 389–392).

Culita, J., Caramihai, S. I., Dumitrache, I., Moisescu, M. A., & Sacala, I. S. (2020). An hybrid approach for urban traffic prediction and control in smart cities. *Sensors*, *20*(24), 7209.

de Medrano, R., & Aznarte, J. L. (2020). A spatio-temporal attention-based spot-forecasting framework for urban traffic prediction. *Applied Soft Computing*, *96*, Article 106615.

Di, X., Xiao, Y., Zhu, C., Deng, Y., Zhao, Q., & Rao, W. (2019). Traffic congestion prediction by spatiotemporal propagation patterns. In *2019 20th IEEE international conference on mobile data management (MDM)* (pp. 298–303).

Ekárt, A., Patelli, A., Lush, V., & Ilie-Zudor, E. (2020). Genetic programming with transfer learning for urban traffic modelling and prediction. In *2020 IEEE congress on evolutionary computation (CEC)* (pp. 1–8).

Fitters, W., Cuzzocrea, A., & Hassani, M. (2021). Enhancing LSTM prediction of vehicle traffic flow data via outlier correlations. In *2021 IEEE 45th annual computers, software, and applications conference (COMPSAC)* (pp. 210–217).

Izhar, A., Quadri, S. M. K., & Rizvi, S. A. M. (2020). Hybrid feature based label generation approach for prediction of traffic congestion in smart cities. In *2020 3rd international conference on intelligent sustainable systems (ICISS)* (pp. 991–997).

Ji, J., Wang, J., Jiang, Z., Ma, J., & Zhang, H. (2020). Interpretable spatiotemporal deep learning model for traffic flow prediction based on potential energy fields. In *2020 IEEE international conference on data mining (ICDM)* (pp. 1076–1081).

Kalamaras, I., Drosou, A., Votis, K., Kehagias, D., & Tzovaras, D. (2018). *A multiobjective data mining approach for road traffic prediction.* Springer International Publishing (pp. 425–436).

Krishnakumari, P., Nguyen, T., Heydenrijk-Ottens, L., Vu, H., & Lint, J. (2017). Traffic congestion pattern classification using multiclass active shape models. *Transportation Research Record*, *2645*, 94–103.

Kunde, F., Hartenstein, A., Pieper, S., & Sauer, P. (2017). Traffic prediction using a deep learning paradigm. In *EDBT/ICDT workshops* (p. 4).

Laharotte, P.-A., Billot, R., & El Faouzi, N.-E. (2017). Detection of non-recurrent road traffic events based on clustering indicators. In *ESANN* (pp. 435–440).

Loumiotis, I., Demestichas, K., Adamopoulou, E., Kosmides, P., Asthenopoulos, V., & Sykas, E. (2018). Road traffic prediction using artificial neural networks. In *2018 SouthEastern European design automation, computer engineering, computer networks and society media conference (SEEDA_CECNSM)* (pp. 1–5).

Mena-Oreja, J., & Gozalvez, J. (2021). On the impact of floating car data and data fusion on the prediction of the traffic density, flow and speed using an error recurrent convolutional neural network. *IEEE Access*, *9*, 133710–133724.

More, R., Mugal, A., Rajgure, S., Adhao, R. B., & Pachghare, V. K. (2016). Road traffic prediction and congestion control using artificial neural networks. In *2016 international conference on computing, analytics and security trends (CAST)* (pp. 52–57).

Mystakidis, A., & Tjortjis, C. (2020). Big data mining for smart cities: Predicting traffic congestion using classification. In *2020 11th international conference on information, intelligence, systems and applications (IISA)* (pp. 1–8).

Müngen, A. A., & Çetın Tas, I. (2021). An investigation about traffic prediction by using ANN and SVM algorithms. In *2021 international conference on electrical, communication, and computer engineering (ICECCE)* (pp. 1–6).

Offor, K. J., Wang, P., & Mihaylova, L. (2019). Multimodel Bayesian Kriging for urban traffic state prediction. In *2019 sensor data fusion: Trends, solutions, applications (SDF)* (pp. 1–6).

Priambodo, B., & Ahmad, A. (2018). Traffic flow prediction model based on neighbouring roads using neural network and multiple regression. *Journal of Information and Communication Technology*, *17*(4), 513–535.

Sarlas, G., & Kouvelas, A. (2019). Analysis of urban traffic network vulnerability and classification of signalized intersections. In *2019 6th international conference on models and technologies for intelligent transportation systems (MTITS)* (pp. 1–6).

Shi, Y., Feng, H., Geng, X., Tang, X., & Wang, Y. (2019). A survey of hybrid deep learning methods for traffic flow prediction. In *Proceedings of the 2019 3rd international conference on advances in image processing, ICAIP 2019* (pp. 133–138). Association for Computing Machinery.

Silva, C., & Martins, F. (2020). *Traffic flow prediction using public transport and weather data: A medium sized city case study, chapter traffic flow prediction using public transport and weather data: A medium sized city case study*. Springer International Publishing (pp. 381–390).

Sinha, A., Puri, R., Balyan, U., Gupta, R., & Verma, A. (2020). Sustainable time series model for vehicular traffic trends prediction in metropolitan network. In *2020 6th international conference on signal processing and communication (ICSC)* (pp. 74–79).

Spławińska, M. (2017). The application of cluster analysis for division of the territory of Poland into homogenous groups in terms of traffic. *DEStech Transactions on Computer Science and Engineering*.

Toshniwal, D., Chaturvedi, N., Parida, M., Garg, A., Choudhary, C., & Choudhary, Y. (2020). Application of clustering algorithms for spatiotemporal analysis of urban traffic data. *Transportation Research Procedia*, *48*, 1046–1059.

Vázquez, J. J., Arjona, J., Linares, M., & Casanovas-Garcia, J. (2020). A comparison of deep learning methods for urban traffic forecasting using floating car data. *Transportation Research Procedia*, *47*, 195–202.

Wang, Y. (2021). Graph neural network in traffic forecasting: A review. In *2021 the 3rd international conference on robotics systems and automation engineering (RSAE), RSAE 2021* (pp. 34–39). Association for Computing Machinery.

Wang, Z., & Thulasiraman, P. (2019). Foreseeing congestion using LSTM on urban traffic flow clusters. In *2019 6th international conference on systems and informatics (ICSAI)* (pp. 768–774).

Wang, Z., Thulasiraman, P., & Thulasiram, R. (2019). A dynamic traffic awareness system for urban driving. In *2019 international conference on Internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData)* (pp. 945–952).

Zambrano-Martinez, J., Calafate, C., Soler, D., Cano, J.-C., & Manzoni, P. (2017). Analysis and classification of the vehicular traffic distribution in an urban area. In *Ad-hoc, mobile, and wireless networks* (pp. 121–134). Springer International Publishing.