Expert Systems With Applications 201 (2022) 117163

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Cellular traffic prediction with machine learning: A survey

Weiwei Jiang

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Keywords:
Cellular network
Clustering
Decomposition
Deep learning
Machine learning
Traffic prediction

ABSTRACT

Cellular networks are important for the success of modern communication systems, which support billions of mobile users and devices. Powered by artificial intelligence techniques, cellular networks are becoming increasingly smarter, and cellular traffic prediction is an important basis for realizing various applications that have originated from this trend. In this survey, we review the relevant studies on cellular traffic prediction and classify the prediction problems as the temporal and spatiotemporal prediction problems. The prediction models with artificial intelligence are categorized into statistical, machine learning, and deep learning models and then compared. Various applications based on cellular traffic prediction are summarized along with their current progress. The potential research directions are pointed out for future research. To the best of our knowledge, this paper is the first comprehensive survey on cellular traffic prediction.

1. Introduction

Cellular network is an important communication network, which provides call, message, and data services to the end users in the range covered by the base stations. Cellular networks have undergone a long history of evolution and development, as depicted in Fig. 1, with progressively increasing mobile communication services and data transmission rates. In the late 1970s, the first-generation cellular network (1G) was set up to provide voice communication service based on analog technology. It was expensive and had limitations related to network coverage accompanied by the low battery power of mobile phones, resulting in poor quality of service, *e.g.*, frequent call drops. The analog transmission system was upgraded to the digital transmission system, *i.e.*, from 1G to 2G, in the 1990s, which highly improved both the reliability and security of the service. Short message service (SMS) was also added in the global system for mobile communication of 2G. Both time division multiple access and code division multiple access (CDMA) were implemented in the 2G system. During the transition from 2G to 3G, 2.5G network using general packet radio service (GPRS) provided Internet communication service. Powered by the universal mobile telecommunication system, CDMA2020, and other technologies, the 3G network provided higher mobile Internet connection ability with a variety of service types, *e.g.*, web browsing, Email, image, and video transmission. Compared with the 3G network, the 4G networks, *e.g.*, worldwide interoperability for microwave access and long-term evolution (LTE), demonstrated speed improvement. Mobile broadband transmission services such as high quality audio or video streaming services were enabled in the 4G network.

As the next generation cellular network, which is still in its early stage of commercial deployment, the 5G network has three different application scenarios, namely, enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency communications (URLCC). The data transmission rate may not be the only criterion; the scenario-specific goals are also important, *e.g.*, the low battery consumption and improved connectivity. To achieve these goals, not only various communication technologies but also artificial intelligence technologies, which can be used for 5G network optimization, optimal resource allocation, 5G physical layer unified acceleration, end-to-end joint optimization of physical layer, etc. (You, Zhang, Tan, Jin, & Wu, 2019), are combined and deployed.

As predicted in the Cisco Annual Internet Report (2018–2023) White Paper, 5G devices and connections will constitute over 10% of global mobile devices and connections by 2023 (Cisco, 2021). More specifically, global mobile devices are expected to grow from 8.8 billion in 2018 to 13.1 billion by 2023, of which 1.4 billion will be 5G capable. Such a huge user group would generate high traffic with various services, and the network resource should be allocated and scheduled more efficiently for supporting the smarter cellular networks, with the help of effective traffic prediction with artificial intelligence techniques. The cellular network would increase the energy consumption, which can be alleviated by using the network adaptively, *e.g.*, smart base station sleeping strategies can be designed based on traffic prediction. Further applications based on cellular traffic prediction would be discussed in Section 4.

Although cellular traffic prediction has been considered in the literature, it remains a challenging concept. One of the challenges relates to the complex internal patterns hidden in the historical traffic data, which can only be described and learned with efficient prediction

E-mail address: jww@bupt.edu.cn.

https://doi.org/10.1016/j.eswa.2022.117163
Received 30 December 2021; Received in revised form 18 March 2022; Accepted 31 March 2022
Available online 9 April 2022

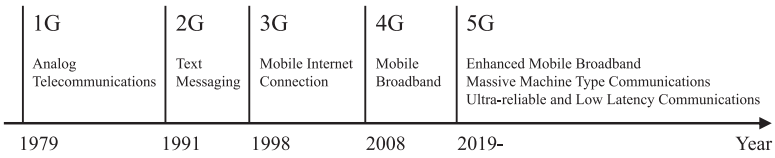

Fig. 1. Different generations of cellular networks.

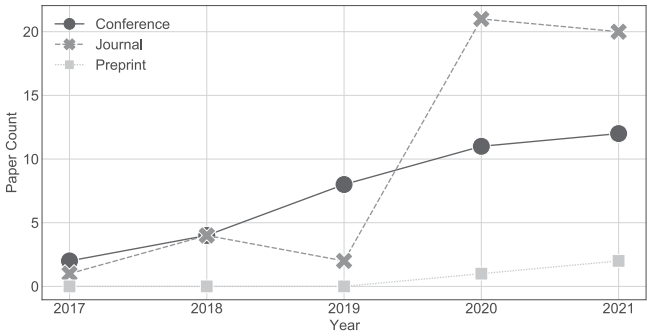

Fig. 2. Year and type of papers considered in this survey (updated until September 1, 2021).

models. Deep learning models have made great progress in this area during the past few years, although further efforts are still required. Another challenge is the practical deployment, which has often been neglected in current literature. The gap between designing a high-performance prediction model and deploying the model in a real-world system remains unaddressed. Inspired by these challenges, the potential research directions are also discussed in Section 5.

To the best of our knowledge, this is the first and the most comprehensive survey on cellular traffic prediction, in terms of both the range of literature reviewed and the depth of discussion. To collect relevant studies, two literature search approaches were adopted. The first approach involved searching the relevant keywords in the Google Scholar and Web of Science platforms. The keywords used included cellular, traffic prediction, and traffic forecasting. The other approach involved searching the articles that cite the open cellular traffic datasets, especially those that are already widely used, *e.g.*, Telecom Italia (Barlacchi et al., 2015). We filtered the identified articles with the purpose of using these datasets for traffic prediction manually.

In summary, 88 papers spanning from 2017 to 2021 were selected and are discussed in this study. These papers are categorized into three types—journal papers (including journals, transactions, magazines, and letters), conference papers (including conferences, symposiums, and workshops), and preprints (included for capturing the latest in-progress studies). The number of papers belonging to the different types and years is illustrated in Fig. 2, which shows that most of the selected papers were published in 2020 and 2021 and journal publications were the most common type of paper.

Based on the surveyed studies, the classification of the specific cellular prediction problems is first proposed. Four problem types are identified, namely, univariate temporal problem, univariate spatiotemporal problem, multivariate temporal problem, and multivariate spatiotemporal problem. The datasets used in the surveyed studies are also categorized into public, semi-public, private, and simulated datasets. The most comprehensive list of open cellular traffic datasets is also presented in this survey.

Then, the classification of the cellular prediction workflows and models is proposed, in which four workflows and three model types are identified. The four workflows include direct-prediction, classification-then-prediction, decomposition-then-prediction, and clustering-then-prediction, which use different data preprocessing techniques. The three model types are statistical, machine learning, and deep learning models, of which the deep learning models are currently emerging as the dominant solutions, as the frontiers of artificial intelligence. Both the general and domain-adapted metrics are summarized and discussed. Some auxiliary techniques that are helpful for prediction are also discussed.

The specific applications based on cellular traffic prediction are also summarized and listed, which include base station sleeping, admission control, resource allocation and scheduling, network dimensioning, network slicing, software defined networking (SDN), and mobile edge computing. Although these applications are far beyond the scope of traffic prediction, the latter can be used for designing a better solution for these applications. Finally, three potential research directions are identified for inspiring follow-up studies.

The major contributions of this survey are summarized as follows:

- The classification of cellular prediction problems is proposed and a comprehensive collection of eight open datasets is summarized and introduced in this survey.
- The classification of cellular prediction models and evaluation metrics is proposed, along with instrumental auxiliary techniques.
- The potential applications and directions for future research are identified, with the purpose of inspiring follow-up studies.

The rest of this survey is organized as follows. Different types of cellular traffic prediction problems and the latest collection of relevant open datasets are presented in Section 2. Various data preprocessing techniques and prediction models used in the surveyed studies are summarized and discussed in Section 3. Several scenarios in which cellular traffic prediction can be applied are presented in Section 4. Some potential directions for future research are pointed out in Section 5. The conclusion is drawn in Section 6.

2. Prediction problems and datasets

2.1. Prediction problems

In this section, we first categorize the cellular traffic problems into two main types – temporal prediction problem and spatiotemporal prediction problem – generated from two different scenarios, as shown in Fig. 3. In Fig. 3(a), the temporal prediction problem is depicted, where there is only one base station and only the traffic of the users or devices connected to this base station is considered. In this simplest type, only the temporal dependency within the historical traffic data would be used. In Fig. 3(b), the spatiotemporal prediction problem is depicted wherein the connected users have moved and connected from one base station to another base station, with the process of handover. In this problem of greater complexity, the traffic at multiple base stations or within multiple areas is considered along with their spatial dependencies in addition to the temporal dependencies. In some special cases of the spatiotemporal prediction problem, the objective is to predict the entire traffic distribution in a given area (Zhu & Wang, 2021) or only at the hotspots (Swedha & Gopi, 2021).

Both types of problems can be formulated as supervised learning problems by using moving windows to generate various input and output pairs, as shown in Fig. 4 for the temporal prediction problem. The collected traffic data are represented as a univariate time series and the prediction for the values in the future time steps is based on the historical data in the past time steps with a fixed length. The moving

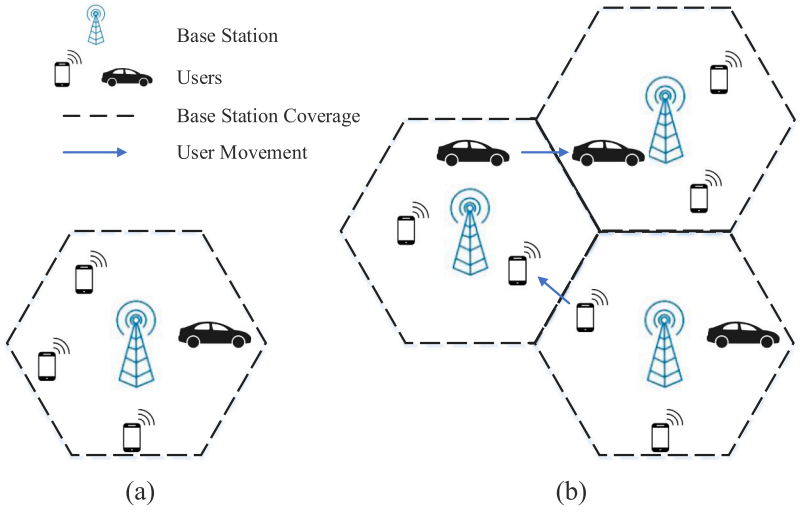

Fig. 3. Two scenarios for different prediction problems. (a) Temporal prediction. (b) Spatiotemporal prediction.

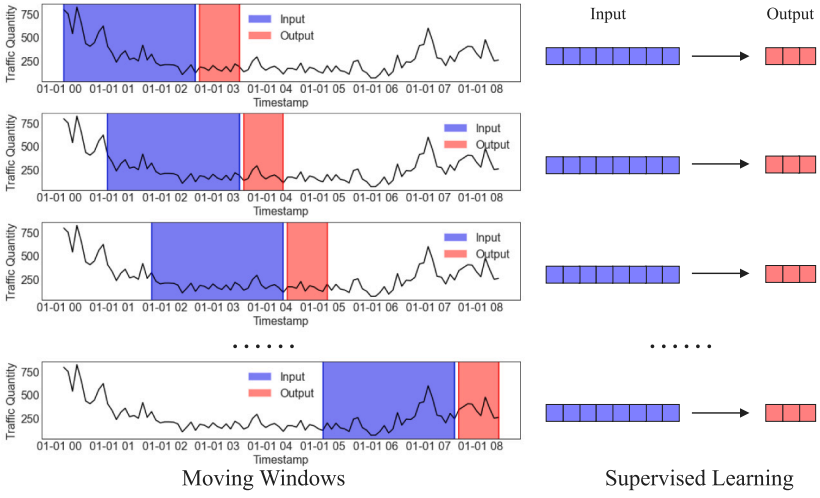

Fig. 4. Supervised learning problem formulation for cellular traffic prediction with moving windows.

windows shown in Fig. 4 are used to generate the input historical data as well as the prediction targets.

In a real-world cellular network, different metrics for measuring the traffic intensity can be used, including but not limited to SMS, call service, and Internet usage service (Barlacchi et al., 2015), physical resource block (PRB) utilization data (Nagib, Abou-Zeid, Hassanein, Sediq, & Boudreau, 2021; Zhao et al., 2020), and the number of connected users (Ferreira, Reis, Senna, & Sargento, 2021; Weerasinghe, Balapuwaduge, & Li, 2019). When only one metric or the aggregated traffic value is collected and used, it becomes a simple univariate prediction problem (or cell-level prediction problem used in previous studies). By adding more metrics and predicting them simultaneously, it becomes a multivariate prediction problem of greater complexity, in which case the service-wise or application-wise dependencies should be considered when the amount of traffic is measured for different services, applications, or users (the service-level, application-level, or user-level predictions used in previous studies are combined as a general multivariate prediction in this survey).

In most of the surveyed studies, the traffic data were collected by the base station or the cellular network operator by analyzing the operation log data and then aggregating the traffic with different time granularity levels. For example, the traffic demand was assessed by monitoring the GPRS tunneling protocol via dedicated probes deployed at the network gateway in Bega, Gramaglia, Fiore, Banchs, and Costa-Perez (2019). It is also widely assumed in the literature that the traffic data are used only at a single base station without transmission or can be transmitted to a central server efficiently, where the prediction is performed with sufficient computing resources. Some practical problems may arise, which have not been adequately researched until now. For example, the traffic log data were prioritized according to their contribution to the prediction accuracy and data of greater importance were transmitted from the base stations to the central server with higher priority in Yamada, Shinkuma, Sato, and Oki (2018). In only a few exceptional cases, the traffic data were collected from the user end by recruiting volunteers and installing appropriate traffic recording tools on their devices such as smartphones or vehicles. In most of the surveyed studies, only the cell-level or aggregated univariate prediction problems were analyzed, with only a few exceptions such as individual traffic prediction considered in Liu, Wu, Li and Wang (2021), in which the individual traffic was decomposed before prediction.

Based on the abovementioned discussion, the different types of cellular prediction problems are summarized in Table 1. In all the cases, the timeline was divided into the time slots periodically with different levels of time granularity. For time period T, most of the studies used a value between five minutes to one hour, especially those that utilized the open datasets. Some extreme values do exist but are very rare; for example, a millisecond resolution was used to predict the

Table 1
Cellular traffic prediction problems classified in this survey.

		Base station number	
		One	Multiple
Traffic variable number	One	Univariate temporal prediction	Univariate spatiotemporal prediction
	Multiple	Multivariate temporal prediction	Multivariate spatiotemporal prediction

PRB utilization data (Nagib et al., 2021). Here, we consider the single-step prediction problem, which aims to predict the traffic in the next time slot. However, the following formulations can be easily extended to the multi-step prediction case, wherein the prediction target contains the traffic volume of more than one future time slot.

We start from the univariate temporal prediction, in which the traffic at the time step i is denoted as a single nonnegative value x_i. The historical data for N time steps were used as the inputs, denoted as a vector $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$. Then, the univariate temporal prediction problem was formulated as follows: find a function f that predicts $y = f(\mathcal{X})$ as the traffic variable x_{N+1} at time step $N+1$.

For the other formats of the prediction problem, the traffic at time step i can be denoted as a vector $\mathbf{x}_i$. For the univariate temporal prediction problem, the elements of $\mathbf{x}_i$ are the same variables collected from different spatial areas within a single time step, *e.g.*, data usages from different base stations. For the multivariate temporal prediction, the elements of $\mathbf{x}_i$ are different variables collected from the same spatial area within a single time step, *e.g.*, SMS, call, and data usages from the same base station. Finally, for the multivariate spatiotemporal prediction problem, the elements of $\mathbf{x}_i$ are different variables collected from different spatial areas. A special case commonly seen in the literature is that when different spatial areas exist in a regular grid, the traffic can be formatted as a matrix with the same size of the grid. We incorporate this special case by transforming and representing the matrix too as a vector. Given the traffic vectors $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in the historical time steps, the prediction problem is re-formulated as follows: find a function f that predicts $\mathbf{y} = f(\mathcal{X})$ as the traffic variable $\mathbf{x}_{N+1}$ at the time step $N+1$.

While historical traffic data are used as the main inputs, some external factors such as the point-of-interest data (Feng, Chen, Gao, Zeng, & Li, 2018; Kuber, Seskar, & Mandayam, 2021), weather situation (Kuber et al., 2021), and social events (Wang et al., 2018) may affect the traffic usage and are also used in the prediction process. User movement patterns are also considered as external factors, *e.g.*, the transition probability matrix between different base stations (Zhao, Jiang et al., 2020). We denote these external factors as ϵ and extend the problem as $\mathbf{y} = f(\mathcal{X}, \epsilon)$, where $\mathcal{X}$ represents the historical traffic vectors and ϵ represents the external factors.

2.2. Datasets

In this survey, the different datasets used for cellular traffic prediction were categorized into the following four types:

- Public datasets, which are real-world datasets, are widely used for academic purposes without any restrictions. These datasets can be obtained from multiple sources, *e.g.*, from the government under a requirement for public information disclosure, from network operators for hosting a data science competition or for academic research, or from researchers for reproducing their studies.
- Semi-public datasets, which are also real-world datasets and can be used in a data science competition or a research project after obtaining a signed document or agreement. We also plan to include the commercial data sources in this type, although this case was not observed in the surveyed studies.
- Private datasets are real-world datasets used in surveyed studies and were collected by researchers and used for their own research purpose, without being shared with the research community for multiple reasons, *e.g.*, data privacy.
- Simulated datasets, which are the datasets generated by computer programs instead of being collected from actual scenarios.

Table 2 summarizes the datasets as well as the associated studies. To encourage open research, the details of the publicly available datasets are summarized in Table 2 along with their download links. Some of these open datasets have already been widely used in the literature, *e.g.*, Telecom Italia (Barlacchi et al., 2015), whereas the others, especially those released more recently, have not been fully considered in relevant studies.

Telecom Italia (Barlacchi et al., 2015) is the most widely used open dataset in the literature on cellular prediction research, as shown in Table 2. This dataset was collected in the city of Milan, Italy, from November 1, 2013, to January 1, 2014. The entire spatial area was divided into 100×100 grids and each grid had an approximate size of 235×235 square meters. By analyzing the Call Detail Records (CDRs) generated by the Telecom Italia cellular network, different attributes were extracted for each grid every 10 min, which included SMSs, calls, and Internet usage data. Based on this dataset, both univariate and multivariate spatiotemporal prediction problems can be considered, depending on the number of attributes used. Some of the state-of-the-art solutions include convolutional neural networks (CNNs) (Chien & Huang, 2021; Gao et al., 2021; Shen et al., 2021; Zhan et al., 2021), Long Short-Term Memory (LSTM) (Kuber et al., 2021; Wu et al., 2021), convolutional LSTM networks (Zeng et al., 2021), Transformer (Liu, Li et al., 2021), and other complex deep learning models (Garrido et al., 2021; Lin, Chen et al., 2021).

The data in City Cellular Traffic Map (C2TM) (Chen et al., 2015) were collected from 13,269 base stations in a medium-sized city in China from August 19, 2012, to August 26, 2012. Each data record contains the base station id, a timestamp, number of mobile users, number of transferred packets, and number of transferred bytes every hour. The locations of these base stations are also provided in this dataset. The individuals can be identified by the international mobile subscriber identity; hence, only the aggregated user number is provided for data privacy protection. This dataset has been adopted in several studies and the state-of-the-art solution is based on LSTM (Swedha & Gopi, 2021).

The dataset LTE Network Traffic Data is publicly available in the Kaggle website, which is a platform for data science competitions. This dataset contains the 4G data usage within 57 cells in 24 h for one year, from October 23, 2017, to October 22, 2018. Because the locations of these 57 cells are not available, the cellular traffic prediction problem based on this dataset can only be categorized as the temporal type; some state-of-the-art solutions include LSTM (Alsaade & Hmoud Al-Adhaileh, 2021; Kurri et al., 2021) and a mix model combining LSTM with the adaptive neuro-fuzzy inference system (ANFIS) time series model (Aldhyani et al., 2020).

Cellular Traffic Analysis Data (Azari et al., 2019a) contains the traffic packets captured from the user side on several Android devices by using virtual private network tunneling. Each packet record contains the following attributes: packet arrival/departure time, source/destination IP addresses, communication protocol (e.g., UDP, TCP, SSL), and encrypted payload. For further use in traffic prediction, the packet payloads are aggregated under different time slots. Several models have been constructed based on this dataset, including auto-regressive integrated moving average (ARIMA) (Azari et al., 2019a, 2019b) and LSTM (Azari et al., 2019a, 2019b; He, Moayyedi et al., 2020).

Table 2
Datasets used in the surveyed studies.

Public datasets			
Type	Name	Link	Relevant studies
Spatial–temporal	Telecom Italia (Barlacchi et al., 2015)	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV	Ale, Zhang, King, and Guardiola (2021), Alvizu, Troia, Maier, and Pattavina (2017), Cai et al. (2020), Chen, Jiaze, Zhang, Sheng, and Cheng (2020), Chien and Huang (2021), Cui, Huang, Wu, and Zheng (2020), Gao et al. (2021), Garrido, Mekikis, Dalgkitsis, and Verikoukis (2021), Garroppo and Callegari (2020), Huang and Chen (2020), Huang, Chiang, and Li (2017), Kuber et al. (2021), Li, Wang, Wang and Zheng (2020), Lin et al. (2021), Liu, Li and Lu (2021), Santos et al. (2020), Shen, Zhang, Guo, and Zhang (2021), Sudhakaran, Venkatagiri, Taukari, Jeganathan, and Muthuchidambaranathan (2020), Wang et al. (2020), Wu, Chen, Zhou, Chen, and Zhang (2021), Yamada et al. (2018), Zeng, Sun, Chen, and Duan (2021), Zeng et al. (2020), Zhan et al. (2021), Zhang, Liu, Xie, Yang and Liu (2020), Zhang and Patras (2018), Zhang, Zhang, Qiao, Yuan, and Zhang (2019), Zhang, Zhang, Yuan and Zhang (2018) and Zhang, Dang, Shihada and Alouini (2021)
Spatial–temporal	City Cellular Traffic Map (C2TM) (Chen, Jin, Qiang, Hu, & Jiang, 2015)	https://github.com/caesar0301/city-cellular-traffic-map	Dommaraju et al. (2020), Mahdy, Abbas, Hassanein, Noureldin, and Abou-zeid (2020), Mejia, Ochoa-Zezzati, and Cruz-Mejia (2020), Swedha and Gopi (2021), Wang, Hu, Min, Zhao and Wang (2020), Zhang, Mozaffari, Saad, Bennis and Debbah (2018) and Zhang et al. (2020)
Temporal	LTE Network Traffic Data	https://www.kaggle.com/naebolo/predict-traffic-of-lte-network	Aldhyani, Alrasheedi, Alqarni, Alzahrani, and Bamhdi (2020), Alsaade and Hmoud Al-Adhaileh (2021) and Kurri, Raja, and Prakasam (2021)
Temporal	Cellular Traffic Analysis Data (Azari, Papapetrou, Denic, & Peters, 2019a)	https://github.com/AminAzari/cellular-traffic-analysis	Azari et al. (2019a), Azari, Papapetrou, Denic, and Peters (2019b) and He, Moayyedi, Dán, Koudouridis and Tengkvist (2020)
Temporal	China Unicom One Cell Data (Huang & Xiao, 2020)	https://github.com/JinScientist/traffic-data-5min	Huang and Xiao (2020)
Spatial–temporal	Shanghai Telecom Dataset (Guo et al., 2020)	http://sguangwang.com/TelecomDataset.html	Ale et al. (2021)
Spatial–temporal	CIKM21-MPGAT Data (Lin, Su, Tung and Hsu, 2021)	https://github.com/cylin-cmlab/CIKM21-MPGAT	Lin, Su et al. (2021)
Temporal	AIIA Data (Guo, Xia, Zhu, & Zhang, 2021)	https://github.com/Phil-Shawn/DMNN	Guo et al. (2021)
Semi-public datasets		Private datasets	Simulated datasets
Relevant studies		Relevant studies	Relevant studies
Deng et al. (2021) and Karimzadeh et al. (2021)		Abozariba, Naeem, Asaduzzaman, and Patwary (2020), Assem, Caglayan, Buda, and O'Sullivan (2018), Bega et al. (2019, 2020), Bejarano-Luque, Toril, Fernández-Navarro, Gijón, and Luna-Ramírez (2021), Clemente et al. (2019), Fang, Cheng, Wang, and Yang (2018), Feng et al. (2018), Ferreira et al. (2021), Gao, Wei, Zhou, and Lv (2019), Gijón, Toril, Luna-Ramírez, Marí-Altozano, and Ruiz-Avilés (2021), Gutterman, Grinshpun, Sharma, and Zussman (2019), Hachemi, Ghomari, Hadjadj-Aoul, and Rubino (2021), Li, Ma, Pan, Liu and You (2020), Liu, Wu et al. (2021), Nagib et al. (2021), Okic and Redondi (2019), Rago, Piro, Boggia, and Dini (2020), Shawel, Debella, Tesfaye, Tefera, and Woldegebreal (2020), Sun and Guo (2021), Tran, Hao, and Trinh (2020), Wang et al. (2017), Wang, Wang, Zhao, and Yue (2021), Wang, Zang and Cheng (2020), Wang et al. (2018), Xia, Wei, Gao, and Lv (2019), Xing, Lin, Gao, and Lu (2021), Yaghoubi, Catovic, Gusmao, Pieczkowski, and Boros (2021), Yu et al. (2020, 2020), Zhang, Zuo, Xu, Han and Zhang (2021), Zhao, Jiang et al. (2020), Zhao et al. (2020) and Zhu and Wang (2021)	He, Dán and Koudouridis (2020), Jiang, Deng, and Nallanathan (2021), Kirmaz, Michalopoulos, Balan, and Gerstacker (2020), Perveen, Abozariba, Patwary, and Aneiba (2021), Weerasinghe et al. (2019), Zeb et al. (2021) and Zhou, Zhao, and Chen (2020)

China Unicom One Cell Data (Huang & Xiao, 2020) were aggregated from the CDR data of the 4G network of a specific mobile operator, China Unicom, in 5-min time steps for 17 months, from January 1, 2016, to May 1, 2017. It can be used in a typical univariate temporal prediction problem with only one base station available. In the same study (Huang & Xiao, 2020), a neural network model with two hidden layers and based on conditional probability modeling between adjacent data windows was proposed as the state-of-the-art prediction model for this dataset.

Shanghai Telecom Dataset (Guo et al., 2020) was collected from 3233 base stations and 9481 mobile phones by Shanghai Telecom in Shanghai, China, from June 1 to November 30, 2014. Unlike other open datasets providing aggregated traffic information, this dataset provides the specific start time and end time for each user session and

the associated base station location. While this dataset was originally collected and used for mobile edge computing, it was also adopted for traffic prediction with Bayesian auto-regressive (AR) and Gaussian process (GP) models as the state-of-the-art solution (Ale et al., 2021).

The traffic data in CIKM21-MPGAT (Lin, Su et al., 2021) were collected from a large-scale cellular geographic system in Hsinchu City, Taiwan. Each traffic record contains the creation time, GPS location and location type (outdoor or indoor), and the international mobile station equipment identification (IMEI), which is a unique mobile phone identification number. The IMEI quantity was aggregated in 5-min time steps for six road intersections from January 1, 2020, to June 30, 2020. This dataset was released in 2021 and has not been widely used. By constructing a graph from these six road intersections, a multivariate and propagation graph attention network (MPGAT) model was proposed as the state-of-the-art prediction model in the same study (Lin, Su et al., 2021).

The AIIA data (Guo et al., 2021) originated from an open data science competition named "AIIA Home Network Competition: Network Traffic Forecasting", organized by the China Mobile and China Artificial Intelligence Industry Development Alliance. The hourly traffic data were recorded from January 1, 2017, to November 15, 2018, for three anonymous regions labeled A, B, and C. Because the regions are anonymous, the relevant spatial information was not available, making the prediction problem based on this dataset a temporal type, and a state-of-the-art dynamic modification neural network model was used for prediction (Guo et al., 2021).

3. Data preprocessing and prediction models

In the previous section, different types of prediction problems were formulated mathematically. In practice, various data preprocessing steps are needed before the core prediction task. In this section, we summarize the frequently used data preprocessing techniques as well as the most common prediction models. As a reference, all abbreviations used for these techniques and models are summarized in Table 3.

3.1. Data preprocessing

In this survey, four types of general prediction workflows were considered – direct-prediction, classification-then-prediction, decomposition-then-prediction, and clustering-then-prediction – which required the use of different data preprocessing techniques.

3.1.1. Direct-prediction

In most surveyed studies, the input historical data and prediction target were already formatted in a proper format, *e.g.*, as a time series or a series of input vectors. In these cases, only some general data preprocessing techniques were required, such as those for data scaling through data standardization or min–max normalization. In some cases, the data collection process was not perfect and the missing data problem inevitably occurred in the collected dataset, which required the application of data imputation techniques. Some simple options include the forward filling or moving average method; more complex techniques were also considered, *e.g.*, the Bayesian Gaussian tensor decomposition was used to impute the missing observations in Deng et al. (2021).

The examples of data normalization and data imputation with a single time series are shown in Figs. 5(a) and 5(b), respectively. In Fig. 5(a), only the data range is changed with data normalization and the pattern is the same before and after the transformation. In Fig. 5(b), the missing range in the original time series is filled with a simple linear imputation method so that the time series is non-null for all time steps.

3.1.2. Classification-then-prediction

The general process of the classification-then-prediction workflow is shown in Fig. 6, in which separate traffic data for different applications or services are collected from the same source, *e.g.*, the raw data packets collected from the same base station or user device. Deep packet inspection techniques are often used to extract the details of the data transmitted, which becomes the basis for traffic classification. Machine learning and deep learning models are used to classify the data packets into specific applications or corresponding services, *e.g.*, Email, text message, video streaming, audio chat, or video call. Then, the traffic quantities are aggregated separately. Multiple prediction models are built to predict the future traffic for individual applications. This workflow has been used in Azari et al. (2019a, 2019b) and Rago et al. (2020).

The benefits of traffic classification before prediction are two-fold. The first benefit is that in the follow-up prediction process, the internal patterns in the traffic for individual applications are more constant and obvious than the mixed and aggregated total traffic, making it easier for the prediction models to achieve better performance. Another benefit is that by extracting a detailed observation of the data usages from different applications, it is possible to design the corresponding measurements, *e.g.*, lowering the quality of video streaming when additional transmission bandwidth is needed by the more important applications.

Another type of traffic classification was proposed and used in Clemente et al. (2019). Based on the cell features, a Naive Bayes classifier is used to determine if the traffic from a single cell presents a predictable or non-predictable pattern, *i.e.*, a potential low or high prediction error. Only those classified as predictable are used in the prediction process later. This kind of classification is beneficial in reducing the potential cost wasted in training a prediction model with preliminary verification.

3.1.3. Decomposition-then-prediction

The general process of the decomposition-then-prediction workflow is shown in Fig. 7, in which a single input traffic time series is first decomposed into multiple components. Then, each component is predicted using a separate model. The final prediction result is the combination of the outputs of these separate models. Unlike traffic classification, the components have no physical meaning, *i.e.*, they are not generated by one or several specific applications or services. These components are extracted with the purpose of achieving an easier and better prediction possibility; for example, the seasonal component in Fig. 7 would be much easier for a statistical model to fit than the original input traffic time series.

Several decomposition techniques have been used in the surveyed studies, *e.g.*, tensor completion (Liu, Wu et al., 2021), discrete wavelet transform (Li, Ma et al., 2020), and Fourier analysis (Wang, Zhou et al., 2020). In Li, Ma et al. (2020), the user traffic time series was decomposed into two components with discrete wavelet transform, namely, a high-frequency component and a low-frequency component. In Wang et al. (2018), the cellular traffic was decomposed into three components, namely, a seasonal component with a periodic pattern, a trend component, and a residual component. Furthermore, different prediction models can be used by considering the component patterns. In Wang, Zhou et al. (2020), the cellular traffic series was decomposed into three components and further modeled with three different models. The large periodic component was extracted with Fourier analysis and modeled as a summation of finite periodic signals, whose parameters were estimated with a least square method. The small random component was predicted using an LSTM model. Finally, a Gaussian process regression (GPR) model was used to learn the residual component.

Table 3
Abbreviations used in Section 3.

Abbreviation	Full name	Abbreviation	Full name
AHW	Additive Holt–Winters	LR	Linear Regression
ANN	Artificial Neural Network	LSTM	Long Short-Term Memory
AR	Auto-Regressive	MA	Moving Average
ARIMA	Auto-Regressive Integrated Moving Average	MDRNN (Wang et al., 2021)	Multidimensional Recurrent Neural Network
ARMA	Auto-Regressive Moving Average	ML	Machine Learning
BGCP	Bayesian Gaussian Tensor Decomposition	MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network	MLR	Multiple Linear Regression
ConvLSTM	Convolutional LSTM	MPGAT (Lin, Su et al., 2021)	Multivariate and Propagation Graph Attention Network
D-Conv	Dilated Convolution Network	NARNN	Nonlinear Autoregressive Neural Network
D-SARIMA	Double Seasonal ARIMA	NMLS	Normalized Least Mean Square
DMNN (Guo et al., 2021)	Dynamic Modification Neural Network	NN	Neural Network
DSSM	Deep Space State Model	RF	Random Forest
ECMCRR-MPDNL (Dommaraju et al., 2020)	Expected Conditional Maximization Clustering and Ruzicka Regression-based Multilayer Perceptron Deep Neural Learning	RL	Reinforcement Learning
EM	Expectation Maximization	RNN	Recurrent Neural Network
ES	Exponential Smoothing	ReLU	Rectified Linear Unit
ETS	Error-Trend-Seasonal	S2S	Sequence to Sequence
FFNN	Feed Forward Neural Network	SAE	Stacked AutoEncoder
FFT	Fast Fourier Transform	SARIMA	Seasonal ARIMA
FedDA (Zhang, Dang et al., 2021)	Dual Attention-Based Federated Learning	SES	Simple Exponential Smoothing
GAN	Generative Adversarial Network	SL	Supervised Learning
GAT	Graph Attention Network	SMC	Sequential Monte Carlo
GBRT	Gradient Boosting Decision Tree	ST-GPKL (Cai et al., 2020)	Spatio-temporal Gaussian Process Kalman Filter
GCN	Graph Convolutional Network	STC	Sequential Tensor Completion
GLU	Gated Linear Units	STCNet (Zhang et al., 2019)	Spatial–Temporal Cross-domain Neural Network
GMM	Gaussian Mixture Model	STHGCN (Zhao, Qin et al., 2020)	Spatio-Temporal Hybrid Graph Convolutional Network
GNN	Graph Neural Network	STN	Spatio-Temporal Neural Network
GP	Gaussian Process	STaLSTMs (Gao et al., 2019)	Spatial–temporal Attention Mechanism based LSTM
GPR	Gaussian Process Regression	SVM	Support Vector Machine
GRU	Gated Recurrent Unit	SVR	Support Vector Regression
GS-STN (Wu et al., 2021)	Geographical and Semantic Spatial–temporal Network	TCN	Temporal Convolutional Network
HW	Holt–Winters	TNN	Tensor Nuclear Norm
HW-ExpS	Holt–Winters Exponential Smoothing	TWACNet (Shen et al., 2021)	Time-wise Attention Aided Convolutional Neural Network
KR	Kernel Ridge	WEM	Weighted Expectation Maximization
LMA	Local Moving Average	att-MCSTCNet (Zeng et al., 2021)	Attention-based Multi-component Spatiotemporal Cross-domain Neural Network

3.1.4. Clustering-then-prediction

The general process of the clustering-then-prediction workflow is shown in Fig. 8. Unlike the classified traffic data that are usually collected from the same source, the input traffic time series used for clustering is often collected from different sources, *e.g.*, from different base stations or cells. The purpose of clustering is to group the different series based on their implicit similarities so that a small number (*e.g.*, the pre-defined cluster number) of prediction models can be built without the computational burden of building a single model for each base station, which could be a large number depending on the size of the spatial area (*e.g.*, there are 13,269 base stations in the C2TM dataset). In contrast, with a greater number of similar time series in the same cluster, the corresponding prediction model would have a larger amount of data as the input, which is beneficial for improving the prediction performance and preventing the overfitting problem. It has been shown that clustering can significantly improve the prediction performance because it increases the consistency of the training data in the same cluster (Mahdy et al., 2020).

Different clustering methods have been used in the surveyed studies, *e.g.*, density-based hierarchical clustering (Xing et al., 2021), k-means clustering (Santos et al., 2020; Shawel et al., 2020), non-crisp fuzzy c-means clustering (Aldhyani et al., 2020), and iterative expectation

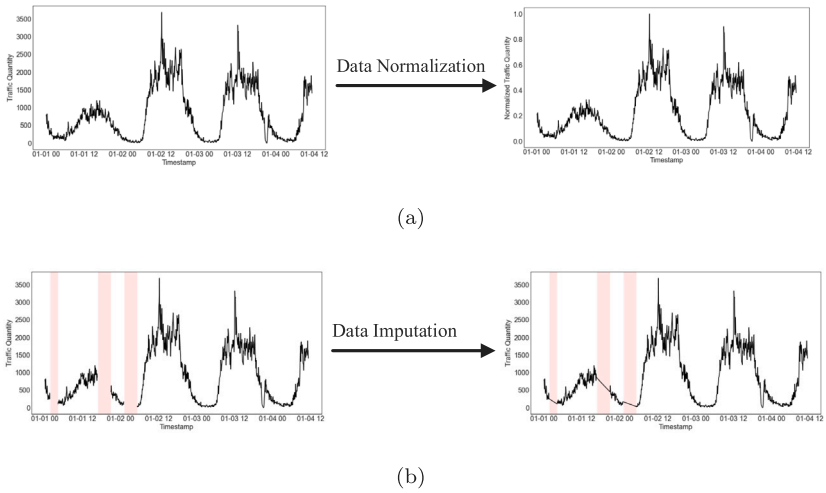

Fig. 5. (a) An example of data normalization. (b) An example of data imputation.

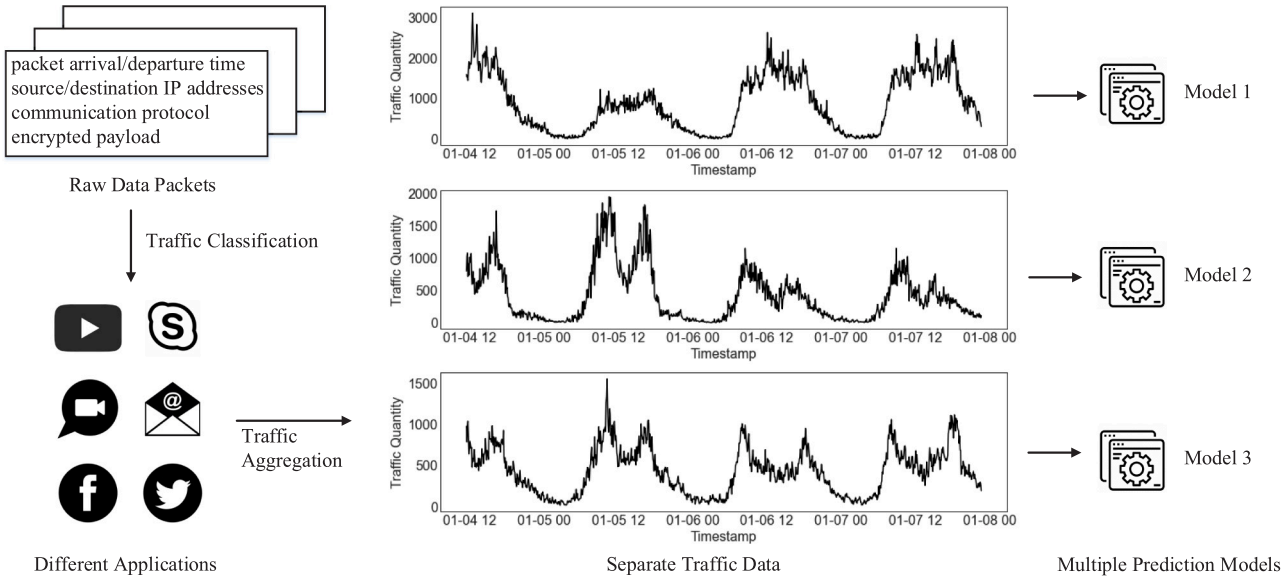

Fig. 6. General process of classification-then-prediction workflow.

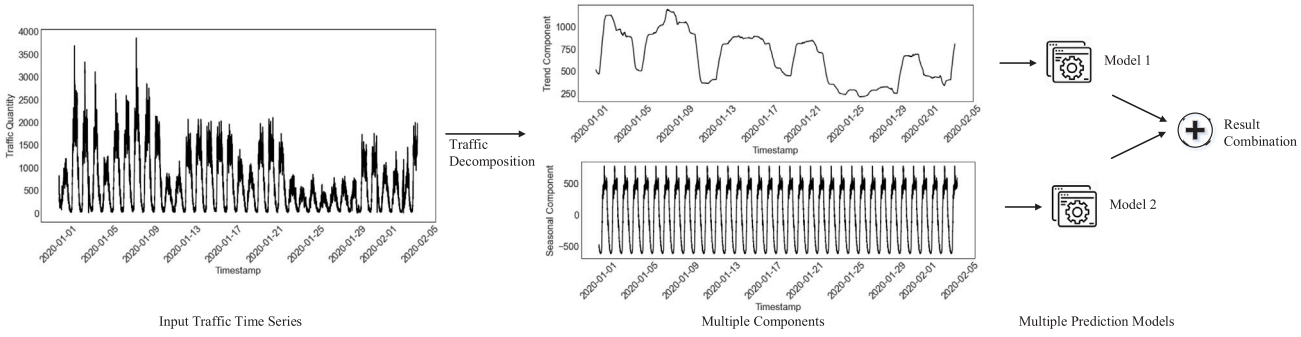

Fig. 7. General process of the decomposition-then-prediction workflow.

conditional maximization clustering (Dommaraju et al., 2020). Different clustering methods can be combined too. For example, in the clustering of different base stations, spatial clustering, which is based on their location, and time series clustering, which is based on their behavior, are both used in Mahdy et al. (2020). Besides the clustering method, the similarity metrics for the traffic series would also affect the clustering result. Some commonly used similarity metrics include the Euclidean distance and dynamic time warping. Other metrics are also proposed and used in the surveyed studies, *e.g.*, the Ruzicka similarity coefficient used in Dommaraju et al. (2020), which is a statistical measurement used to find the strength of the relationship between the variables.

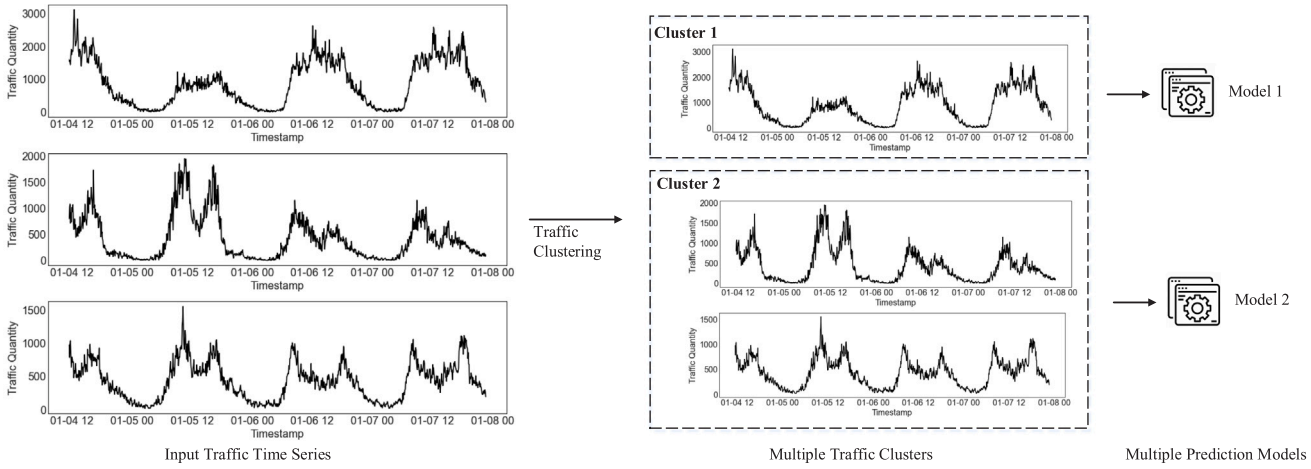

Fig. 8. General process of clustering-then-prediction workflow.

3.2. Evaluation metrics

Before discussing the different prediction models, the evaluation metrics for the prediction performance comparison are summarized and introduced here. Both the widely used general metrics and domain-adapted metrics designed for cellular traffic prediction are discussed.

3.2.1. General metrics

The major concern of the prediction performance is the error between the true observations and the predicted values, which can be measured by a series of regression evaluation metrics, *e.g.*, mean square error (MSE), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). If the true observations are denoted as $\mathbf{y}$ and the predicted values are denoted as $\hat{\mathbf{y}}$, these metrics are defined as

$$\begin{aligned} MSE(\mathbf{y},\hat{\mathbf{y}}) &= \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2, \\ RMSE(\mathbf{y},\hat{\mathbf{y}}) &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}, \\ MAE(\mathbf{y},\hat{\mathbf{y}}) &= \frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i|, \\ MAPE(\mathbf{y},\hat{\mathbf{y}}) &= \frac{1}{N}\sum_{i=1}^{N}|\frac{y_i-\hat{y}_i}{y_i}|\times 100\%, \end{aligned} \tag{1}$$

where N is the total number of observations used for the evaluation. A lower value would indicate a better prediction.

Another metric that is often used is the determination coefficient (*i.e.*, R^2), which is the square of the correlation coefficient. Using the above notations, it can be defined as $R^2 = \frac{SSR}{SST}$, where $\overline{y}$ is the mean value of the true observations, and $SSR = \sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2$ and $SST = \sum_{i=1}^{N}(y_i - \overline{y})^2$. R^2 ranges between 0 and 1, and a higher value implies a better prediction. Other regression metrics used include the average normalized root-mean square error (ANRMSE), mean squared logarithmic loss (MSLE) used in Sudhakaran et al. (2020), mean squared observed error (MSOE) with consideration of missing data in Deng et al. (2021), log loss used in Wang et al. (2017), and absolute cumulative error (ACE) used in Sun and Guo (2021).

Although most of the surveyed studies focus on specific traffic values, some only consider the trend of the predicted values, *e.g.*, the directional movement. In these cases, the classification evaluation metrics are used, *e.g.*, mean accuracy (MA) and false positive rate (FPR) used in Dommaraju et al. (2020).

Besides the prediction performance, another concern is the running time consumption of the prediction model, which includes the training time and prediction time separately for most machine learning and deep learning models. Several studies have considered the running time consumption. For example, the training time was evaluated in Shen et al. (2021) and the prediction time was evaluated in Dommaraju et al. (2020). The last concern is the computing device requirements for deploying the prediction models. However, because most of the reviewed studies are based on computer simulations, the practical device requirements are rarely mentioned.

3.2.2. Domain-adapted metrics

Although the above general evaluation metrics are widely adopted for temporal and spatiotemporal prediction problems, they fail to consider the specific requirements when being applied to cellular traffic prediction tasks and other applications (discussed in Section 4). It should be pointed out that not only the accuracy of the predictor matters, but also the usability of the predicted values in these applications. In the surveyed literature, only a few new evaluation metrics were designed by combining the domain knowledge; these are referred to as domain-adapted metrics here.

In Yu, Musumeci et al. (2020), cellular traffic prediction was used for network slice adjustment and migration for 5G radio access; thus, the prediction performance was further evaluated by the penalty caused by service degradation and the migrated traffic caused by slice migrations. Over-provisioning and service level agreement (SLA) violations caused by inaccurate traffic prediction were considered in Bega et al. (2019, 2020), and a customized loss function was designed by considering the penalty caused by the prediction error in terms of the monetary cost of the operator and using it to train the neural network model. To drive resource orchestration and scheduling mechanisms, a general and flexible context-aware loss function formulation was proposed in Garrido et al. (2021) by considering the penalties for both under-provisioning and over-provisioning. An impressive improvement of up to 61.3% was achieved by using the customized loss function for the base station traffic prediction when compared with the general metric of MSE.

3.3. Prediction models

The prediction models used in the surveyed studies were roughly categorized into three types, namely, statistical, machine learning, and deep learning models. Six papers used statistical models, 14 papers used machine learning models, and 68 papers used deep learning models. Therefore, this study focused more on the deep learning models. It is beyond the scope of this survey to discuss every single prediction model thoroughly; hence, only the common models are introduced and discussed. Nonetheless, the proposed/adopted models, baselines, and evaluation metrics are listed for each study for reference. Some studies have no baselines (*e.g.*, Cui et al. (2020)), as only the afterward applications with and without prediction are compared and all the comparisons demonstrate that the scheme with prediction is better than that without prediction.

Table 4
Summary of statistical models used in the studies included in this survey.

Study	Proposed/Adopted model	Baseline	Evaluation metrics
Huang and Xiao (2020)	Conditional Probability Estimation	Additive HW, Multiplicative HW, XGBoost, RNN	sMAPE
Tran et al. (2020)	ES	SARIMA	RMSE
Wang, Zang et al. (2020)	ARMA	N/A	MSE, MAE
Clemente et al. (2019)	HW	N/A	RMSE
Zhou et al. (2020)	Improved HW	OGD, LSTM, HW	R^2
Perveen et al. (2021)	SMC-based Particle Filtering	N/A	MSE

3.3.1. Statistical models

As presented in Table 4, the statistical models used in the surveyed studies are mainly the time series, probability estimation, and particle filtering models. Some representative models include the ARIMA and Holt–Winters (HW) model. As a univariate time series model, the prediction in the ARIMA model is based on the weighted linear combination of three types of components, namely, the autoregressive component, difference component, and moving average component. The order of these components can be chosen using the Akaike information criterion or Bayesian information criterion. Similar to ARIMA, the HW model is a widely used univariate time series prediction model, which is a combination of the following three components: simple exponential smoothing (ES), Holt's ES, and Winter's ES.

In addition to these two models, other statistical models have been used in the surveyed studies. Simple ES methods were used in Tran et al. (2020), with an error-trend-seasonal (ETS) framework that chooses the models based on likelihood-based information criterion or out-of-sample average mean square error minimization. A sequential Monte Carlo (SMC)-based particle filtering method was adopted in Perveen et al. (2021) for traffic demand prediction, with an objective to estimate the posterior density of the state variables from the observation variables.

Compared with the machine learning and deep learning models, statistical models require much less computation, making them attractive in mobile devices with limited storage and computation abilities. Statistical models also possess the theoretical advantage of choosing the suitable model parameter, instead of the trial-and-error process commonly seen in machine learning models, *e.g.*, in the grid search process for hyper-parameter tuning. However, most statistical models are based on a linear relationship between the input and output values, and they are mostly outperformed by the machine learning models with the ability to describe and learn the nonlinear relationship (Huang & Xiao, 2020).

3.3.2. Machine learning models

As presented in Table 5, the machine learning models used in the surveyed studies include random forest (RF), LightGBM (Ke et al., 2017), Gaussian progress regression (GPR), multiple linear regression (MLR), and Prophet (Taylor & Letham, 2018). These models are characterized by "shallow" structures when compared with the deep learning models. Most of them are tree-based, *e.g.*, RF and LightGBM, whereas the others may be based on GP or other mechanisms. In Xia et al. (2019), RF and LightGBM are combined for traffic prediction, in which RF is used to filter the redundant features and LightGBM is used as the final prediction model. Based on the weighted expectation maximization (WEM) algorithm, a Gaussian mixture model (GMM) was introduced for traffic prediction in Zhang, Mozaffari et al. (2018). Furthermore, a feature embedding kernel was proposed for the GP model in Sun and Guo (2021), which achieves a trade-off between the overall prediction accuracy and peak–trough accuracy. Machine learning models generally perform better than statistical models, but can achieve better prediction result than deep learning models in only a few cases (Cai et al., 2020; Liu, Wu et al., 2021).

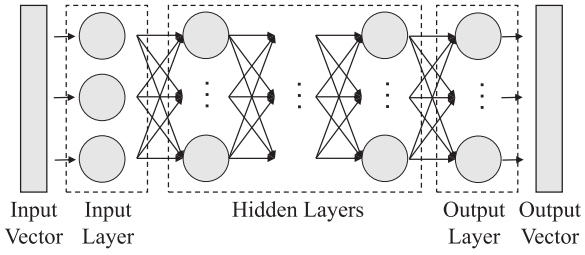

Fig. 9. Simple feed-forward neural network.

3.3.3. Deep learning models

As presented in Table 6, deep learning models used in the surveyed studies include feed-forward neural networks (FFNNs), CNNs, and recurrent neural networks (RNNs). Owing to their broad coverage, deep neural networks have been utilized for deep learning and have achieved great success in the past decade for a series of prediction problems, *e.g.*, vehicular traffic prediction (Jiang & Luo, 2021; Jiang & Zhang, 2018), Internet traffic prediction (Jiang, 2021c, 2021d), and stock market prediction (Jiang, 2021a). As indicated in Table 6, deep neural networks have demonstrated remarkable performance in cellular prediction problems, especially in the last three years. This can be mainly attributed to the accumulation of traffic data (for example, those discussed in Section 2) and improved computation resources (for example, parallel model training with graphics processing units). Other reasons include various input features, well-designed deep neural network structures, and other auxiliary technologies.

As shown in Fig. 9, the simple FFNN is the most basic deep neural network that is used for cellular traffic prediction. Similar models with different names have been used in the surveyed studies (Alvizu et al., 2017; Kirmaz et al., 2020), *e.g.*, artificial neural networks (ANNs) and multi-layer perceptron (MLP). FFNN can also be used as a network comprising densely connected layers for generating the final output in highly complex models; this aspect has been omitted from Table 6 as this type of usage is common. To facilitate the learning ability, additional hidden layers or various nonlinear activation functions can be used, *e.g.*, tanh, sigmoid, and ReLU functions. To use FFNNs for the spatial–temporal prediction problem, a flattening layer is needed to transform any format of the inputs into a vector, which may cause loss of spatial information in the inputs and downgrade the prediction performance. Several regularization techniques have been developed, *e.g.*, dropout and early stopping, to prevent the overfitting problem commonly seen in deep learning models.

FFNN represents a universal structure and can theoretically fit all possible mapping relationships between the input and output. However, it is not efficient in practice, owing to the use of a large number of model parameters. Originally designed for two-dimensional image processing problems, CNNs use small-size convolutional kernels and convolution operations to extract the local information from a limited receptive field instead of using a full connection. Different feature maps can be extracted with different convolutional kernels, and both the input and kernel sizes can be 1D vectors, 2D matrices, or 3D tensors, as depicted in Figs. 10(a), 10(b), and 10(c). The pooling layers can be further used to reduce the data size and location dependency of

Table 5
Summary of machine learning models used in the studies reviewed in this survey.

Study	Proposed/Adopted model	Baseline	Evaluation metrics
Weerasinghe et al. (2019)	Tree-based ML model	N/A	RMSE
Yamada et al. (2018)	RF	N/A	RMSE
Liu, Wu et al. (2021)	Prophet+GPR	LSTM, GPR, Wavelet-PG, p-TNN-GP, p-TNN-GG, TNN, STC	RMSE, MAE, MAPE
Garroppo and Callegari (2020)	Improved NMLS	NMLS, SARIMA, DERIvative approach	MSE
Ale et al. (2021)	Bayesian AR and GP models	N/A	MSE, MAE, MAPE, RMSEP, PMCC
Cai et al. (2020)	ST-GPKL	LR, SVR, MA, AR, ARIMA, LSTM	MSE, MAE
Zhang, Mozaffari et al. (2018)	GMM	N/A	N/A
Zhang, Saad et al. (2020)	WEM	EM, k-means methods	MRE
Yu, Musumeci et al. (2020)	MLR	N/A	Service degradation penalty, migrated traffic
Okic and Redondi (2019)	MLR+NN+ARIMA	Last Observation, MLR, NN, ARIMA	NRMSE, Normalized Underestimated RMSE
Li, Ma et al. (2020)	Wavelet-Prophet-GPR	ARIMA, Wavelet-ARIMA	RMSE, MAPE
Xia et al. (2019)	RF+LightGBM	ARIMA, MLP, LR	MAPE, R^2
Sun and Guo (2021)	GP	GP, ARIMA	ACE
Abozariba et al. (2020)	ML Model	N/A	MAPE

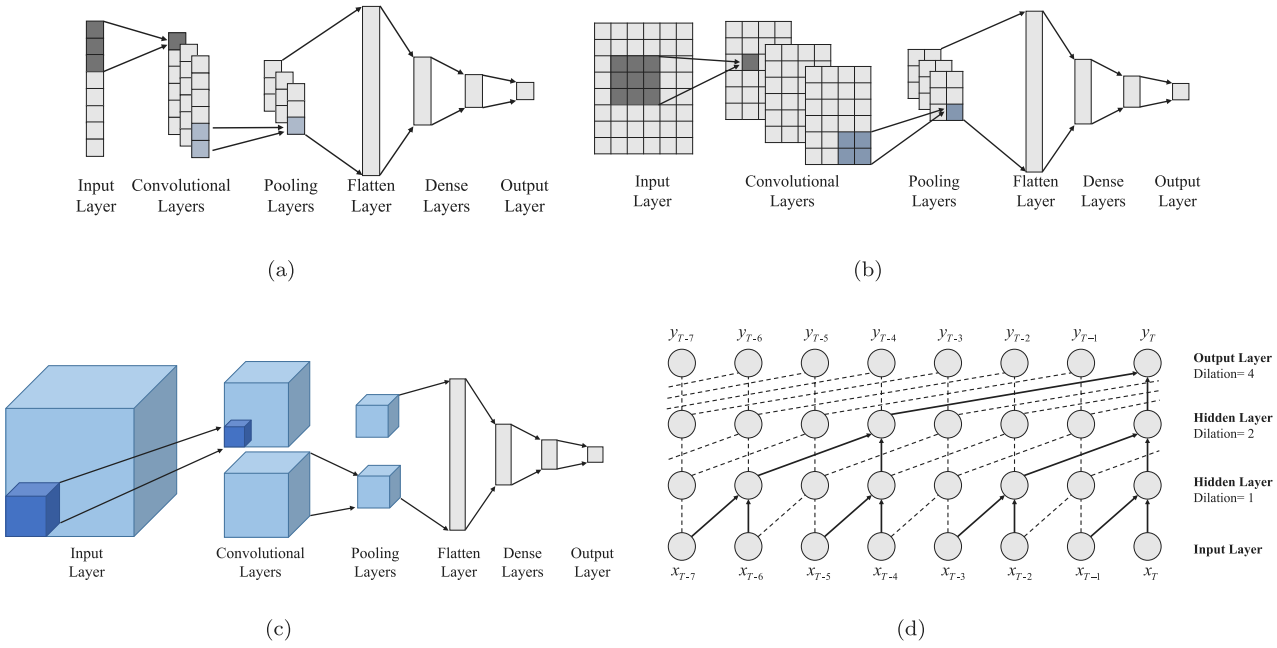

Fig. 10. Examples of (a) 1D CNN. (b) 2D CNN. (c) 3D CNN. (d) TCN.

the convolutional layers. The flattening layer and the follow-up dense layers are used to generate the required output format. Alternatively, a global average pooling layer can be used for generating the output.

Although CNNs are effective in solving several problems, they are not explicitly designed for time series prediction problems. An exception is the temporal convolutional network (TCN) (Bai, Kolter, & Koltun, 2018) shown in Fig. 10(d), which proposes the use of causal and dilated convolution for solving time series problems. Causal convolution was proposed to prevent future information usage. Dilated convolution was proposed for increasing the receptive field and is suitable for long sequences. In Fig. 10(d), the input is denoted as $\mathbf{X} = (x_1, x_2, \ldots, x_t)$ and the target is denoted as $\mathbf{Y} = (y_1, y_2, \ldots, y_t)$. A convolutional kernel of size K is denoted as $\mathbf{F} = (f_1, f_2, \ldots, f_K)$ and the dilatation rate used for the dilated convolution is d. Then, the convolution operation in the TCN can be formulated as follows:

$$(\mathbf{F} *_d \mathbf{X})_{(x_t)} = \sum_{k=1}^{K} f_k x_{t-(K-k)d}. \tag{2}$$

Unlike CNNs, RNNs (Rumelhart, Hinton, & Williams, 1986) can handle various types of sequence data, *e.g.*, natural languages, and time series. Compared with FFNN, hidden states are introduced in the RNN to remember the historical information. As shown in Figs. 11(a) and 11(b) for a basic RNN cell, the previous hidden state h_{t-1} is used to calculate the new hidden state h_t in a single cell as follows:

$$h_t = tanh(x_t \cdot w_{xh} + h_t \cdot w_{hh} + b_h), \tag{3}$$

where x_t is the input, w_{xh} and w_{hh} are the weight parameters, and b_h is the bias parameter. After t steps in a recurrent approach, the final output element can be calculated with the hidden state h_t as follows:

$$o_t = tanh(h_t \cdot w_{oh} + b_o), \tag{4}$$

where o_t is the output, w_{oh} is the weight parameter, and b_o is the bias parameter.

Although the concept of RNN is intuitive, it is not practical for several problems including the exploding and vanishing gradient problems. Hence, RNN variants such as the LSTM (Hochreiter & Schmidhuber, 1997) and gated recurrent unit (GRU) (Cho et al., 2014) were proposed, both of which are widely used in the surveyed studies. As shown in Fig. 11(c), LSTM introduces three gate mechanisms in a single cell to better control the stored information. LSTM also introduces a new variable as the cell state to maintain the historical information. The forget gate is used to decide the information to be discarded from this cell state as follows:

$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f), \tag{5}$$

Table 6
Summary of deep learning models used in the studies reviewed in this survey.

Study	Proposed/Adopted model	Baseline	Evaluation metrics
Zhu and Wang (2021)	LSTM	ARIMA	NRMSE
Swedha and Gopi (2021)	LSTM	N/A	Hotspot coordinate error
Nagib et al. (2021)	LSTM	Last observation, MA, SES	R^2, RMSE
Zhao, Jiang et al. (2020)	STGCN-HO (GCN+GLU)	HA, ARIMA, LSTM, GLU	RMSE, MAE, RRMSE
Ferreira et al. (2021)	N/A	RF, XGBoost, RF, GBRT, SVR, LSTM, FFNN, CNN	RMSE
Bega et al. (2019)	DeepCog (3D CNN)	Native method, GSAE-LSAE-LSTM (Wang et al., 2017), STN (Zhang & Patras, 2018)	Overprovisioning, SLA violations
Feng et al. (2018)	DeepTP (LSTM+attention)	ARIMA, SVR, RNN	NRMSE
Kuber et al. (2021)	LSTM	ARIMA, ES	RMSE, MAE
Wang et al. (2018)	GNN+RNN	NAIVE method, ARIMA, LSTM, HW	MAE, MARE
Li, Wang et al. (2020)	LA-ResNet (CNN+LSTM+attention)	ARIMA, 3D CNN, LSTM, GRU, CNN+RNN, MTL	RMSE, Accuracy
Chien and Huang (2021)	LSTCNN	CNN, CNN+LSTM, ConvLSTM, DenseNet	MAE, RMSE, MAPE
Zeng et al. (2021)	att-MCSTCNet (ConvLSTM+ConvGRU+attention)	LR, SVR, LSTM, DenseNet, STMNet, STNet, STCNet	RMSE, MAE, R^2
Zhan et al. (2021)	CNN	HA, ARIMA, SVR	RMSE, MAE, MAPE, Accuracy
Wang, Zhou et al. (2020)	LSTM-GPR	ARIMA, LSTM	RMSE, MAE
Zhang, Liu et al. (2020)	HSTNet (CNN+attention)	HA, ARIMA, LSTM, STDenseNet	MAE, RMSE
Wu et al. (2021)	GS-STN (LSTM+CNN+attention)	ARIMA, DNN, CNN-LSTM	NMAE, NRMSE
Zhang et al. (2019)	STCNet (ConvLSTM)	LR, SVR, LSTM, DenseNet (Zhang, Zhang et al., 2018), ST-Net, STM-Net	RMSE, MAE, R^2
Huang and Chen (2020)	CDRF (Huang et al., 2017)	N/A	N/A
Alvizu et al. (2017)	FFNN	N/A	MAE, MAPE, RMSE
Chen et al. (2020)	Transformer+attention	LSTM, Seq2Seq-Attn, DeepAR, DSSM	R, RMSE, sMAPE
Santos et al. (2020)	LSTM	GRU	RMSE
Shen et al. (2021)	TWACNet (attention+CNN)	STCNet, GRUNet, CT-CAM	RMSE, Training time
Zeng et al. (2020)	STC-N (CNN+ConvLSTM)	N/A	RMSE, MAE, R^2
Liu, Li et al. (2021)	ST-Tran (Transformer+attention)	HA, ARIMA, LSTM, STDenseNet, STACN, ConvLSTM	MAE, NRMSE, R^2
Sudhakaran et al. (2020)	DenseNet+ResNet	DenseNet, ResNet	RMSE, MSLE
Zhang, Dang et al. (2021)	FedDA (attention)	LASSO, SVR, LSTM, FedAvg, FedAtt	MSE, MAE
Garrido et al. (2021)	LSTM, 3D-CNN (Bega et al., 2019), STN (Zhang & Patras, 2018), ConvLSTM	N/A	Context-aware loss
Huang et al. (2017)	3D CNN+LSTM	ARIMA, ANN, RNN, 3D CNN	MA, MAE, RMSE
Gao et al. (2021)	CNN	ARIMA, Prophet	ANRMSE
Zhang, Zhang et al. (2018)	CNN	HA, ARIMA, LSTM	RMSE
Zhang and Patras (2018)	STN (ConvLSTM+3D CNN)	HW-ExpS, ARIMA, MLP, ConvLSTM, 3D CNN	NRMSE
Cui et al. (2020)	ConvLSTM	N/A	N/A
Lin, Chen et al. (2021)	GCN+LSTM	ARIMA, LSTM, ConvLSTM	RMSE, MAE, R^2
Mahdy et al. (2020)	GRU	ARIMA, SARIMA, SVM, MLP, DT, RF, XGBoost, LSTM	MSE,RMSE,MAE
Wang, Hu et al. (2020)	ctGAN-S2S	ARIMA, NARNN, LSTM, S2S	MSE, MAE
Dommaraju et al. (2020)	ECMCRR-MPDNL (FFNN)	STCNet (Zhang et al., 2019), GNN-D (Wang et al., 2018)	MA, FPR, Prediction time
Mejia et al. (2020)	3D CNN	ARIMA, LSTM	RMSE, MAE, MAPE, Accuracy
Kurri et al. (2021)	LSTM	ARIMA	MSE, MAE, R^2
Aldhyani et al. (2020)	LSTM+ANFIS	LSTM, ANFIS	MSE, RMSE, MAE, R
Alsaade and Hmoud Al-Adhaileh (2021)	SES-LSTM	LSTM	MSE, RMSE, NRMSE, R^2
Azari et al. (2019a)	LSTM	ARIMA	RMSE
Azari et al. (2019b)	LSTM	ARIMA, AR(1)	RMSE
He, Moayyedi et al. (2020)	LSTM	ARIMA	RMSE
Lin, Su et al. (2021)	MPGAT (GAT+TCN)	GWENT, MTGNN, STAWNET	MAPE
Guo et al. (2021)	DMNN (LSTM)	LSTM, BiLSTM, ConvLSTM, ConvBiLSTM	MAPE
Deng et al. (2021)	BGCP-RNN-ReLU	GRU, LSTM, RNN	MSOE
Karimzadeh et al. (2021)	HERITOR (LSTM+CNN)	A probabilistic model	MAE
Wang et al. (2017)	GSAE-LSAE-LSTM	ARIMA, SVR	MSE, MAE, Log loss
Gao et al. (2019)	STaLSTMs (LSTM+attention)	XGBoost, LSTM, attention-LSTM	MAPE, R^2
Bejarano-Luque et al. (2021)	CNN+LSTM	HW, SARIMA, LSTM, ConvLSTM, D-Conv	MAE
Zhao, Qin et al. (2020)	STHGCN (GCN+GRU)	HA, Prophet, STGCN, ASTGCN, DCRNN, Graph WaveNet	MSE, MAE
Bega et al. (2020)	3D CNN	SARIMA	Unserviced demand, Cost gains
Zhang, Zuo et al. (2021)	LMA-DeepAR	ETS, ARIMA, XGBOOST, LSTM, DeepAR	RMSLE
Shawel et al. (2020)	D-SARIMA+LSTM	LSTM, D-SARIMA	RMSE, MAE
Gijón et al. (2021)	N/A	SARIMA, AHW, RNN, RF, ANN-MLP, ANN-LSTM, SVR	MAE, MAPE
Fang et al. (2018)	GCN+LSTM	LSTM, GCN, SARIMA	MAE
Hachemi et al. (2021)	FFT+LSTM	ARIMA	MAPE
Rago et al. (2020)	AE	LSTM	MSE
Gutterman et al. (2019)	X-LSTM	LSTM, ARIMA	RMSE, MAE, MAPE
Assem et al. (2018)	ST-DenNetFus (CNN)	Naive, ARIMA, RNN, LSTM	RMSE, MAE
Yu, Li et al. (2020)	STEP (GCN+GRU)	ARIMA, LSTM, GNN	MARE, RMSE
Yaghoubi et al. (2021)	RF, LSTM	SVR, KR, DT	R^2
Xing et al. (2021)	TPBLN (LSTM)	XGBoost, LightGBM	R^2
Wang et al. (2021)	MDRNN	N/A	MAPE, MFR
Kirmaz et al. (2020)	MLP+CNN+LSTM	LR	MSE
He, Dán et al. (2020)	LSTM	N/A	N/A
Jiang et al. (2021)	LSTM	N/A	N/A
Zeb et al. (2021)	LSTM	N/A	RMSE, R^2

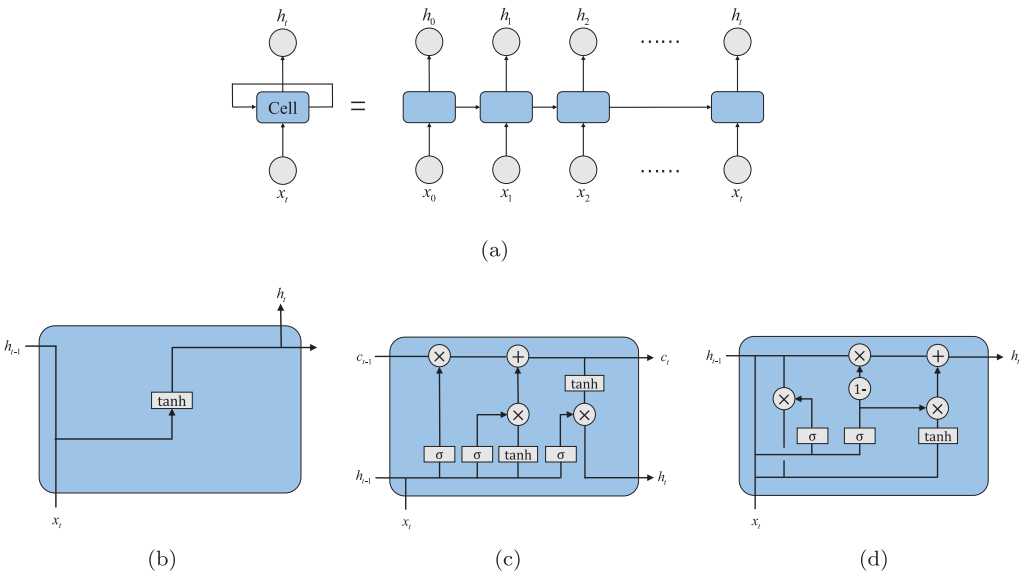

Fig. 11. (a) The general structure of RNN. (b) An example of RNN cell. (c) An example of LSTM cell. (d) An example of GRU cell.

where x_t is the input, h_{t-1} is the previous hidden state, c_{t-1} is the previous cell state, $\sigma(*)$ is the sigmoid activation function, W_{xf}, W_{hf}, and W_{cf} are the weight parameters, and b_f is the bias parameter.

Based on f_t, the input gate is used to decide the information to be stored in the cell state as follows:

$$\begin{aligned} c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t \\ i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ \tilde{c}_t &= tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \end{aligned} \tag{6}$$

where c_t is the cell state, $tanh(*)$ is the tanh activation function, W_{xi}, W_{hi}, W_{ci}, W_{xc}, and W_{hc} are the weight parameters, and b_i and b_c are the bias parameters.

Finally, the output is decided by the output gate depending on c_t as follows:

$$\begin{aligned} o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ h_t &= o_t * tanh(c_t), \end{aligned} \tag{7}$$

where o_t is the output, W_{xo}, W_{ho}, and W_{co} are the weight parameters, and b_o is the bias parameter.

As shown in Fig. 11(d), GRU is a simplified variant of LSTM and uses only two gate mechanisms, namely, the update and output gates. The cell state is also omitted in GRU. A similar process for updating the hidden state in GRU is formulated as follows:

$$\begin{aligned} z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \\ r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \\ \tilde{h}_t &= \sigma(W_{xo}x_t + W_{ho}r_t * h_{t-1} + b_o) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t, \end{aligned} \tag{8}$$

where z_t and r_t are the intermediate variables for calculation, W_* and b_* represent the weight and bias parameters, and the other parameters are the same as for LSTM.

These RNN variants are suitable for the temporal prediction problem with input time series but they are not well designed for the spatial–temporal prediction problem. Convolutional LSTM (ConvLSTM) (Shi et al., 2015) is an important extension of LSTM, which takes a series of two-dimensional matrices as the input. ConvLSTM has been used in the spatial–temporal prediction problem (Cui et al., 2020; Zeng et al., 2020; Zhang & Patras, 2018; Zhang et al., 2019) when the spatial area is divided into grids, as in the case of Telecom Italia (Barlacchi et al., 2015). It is not a difficult task to extend LSTM to ConvLSTM by replacing the variables from the vectors to the matrices and performing a simple matrix multiplication with the convolution operation $*$ as follows:

$$\begin{aligned} f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f), \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tilde{C}_t, \\ i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i), \\ \tilde{C}_t &= tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o), \\ H_t &= o_t \circ tanh(C_t), \end{aligned} \tag{9}$$

where $\circ$ is the Hadamard product and all the variables are as defined for LSTM but with different dimensions.

Another improvement in the LSTM is the combination with the attention mechanism (Vaswani et al., 2017), which aims to use all the hidden states to generate an output instead of the just the last one. The attention can be defined as the weighted average of the LSTM hidden states, with the dynamic weights learned from the data instead of the fixed weights. The attention mechanism originally proposed for computer vision tasks and full-attention models such as Transformer (Vaswani et al., 2017) are widely used nowadays. The attention mechanism and the Transformer networks have also been proven effective in cellular traffic prediction problems (Chen et al., 2020; Liu, Li et al., 2021; Zhang, Dang et al., 2021).

Compared with the neural networks with a single structure, the combination of different types of neural networks is more promising and achieves better performance. For example, the combinations of CNNs and RNNs have achieved desirable prediction result in a series of studies (Bejarano-Luque et al., 2021; Huang et al., 2017; Karimzadeh et al., 2021; Kirmaz et al., 2020; Li, Wang et al., 2020; Wu et al., 2021; Zeng et al., 2020). Only the commonly used deep neural networks are discussed in this paper, but it must be noted that additional options are also available for cellular prediction problems. For example, graph-based models (Jiang, 2021b) such as graph convolutional network (GCN) (Kipf & Welling, 2017) and graph attention network (GAT) (Veličković et al., 2018) have been used for spatial–temporal prediction problems; these exhibit an effective approach for exploiting the spatial dependency among multiple base stations and are not restricted by the grid division of the interested spatial area (Lin, Chen et al., 2021; Lin, Su et al., 2021; Yu, Li et al., 2020; Zhao, Qin et al., 2020).

To conclude this part, a comparison between statistical, machine learning, and deep learning models is made in Table 7, from different aspects. The major advantages of statistical models include their low requirement for both training data and computation resources and the high theoretical interpretability. The main drawback is the less

Table 7
A comparison between different kinds of prediction models.

Type	Data requirement	Computation requirement	Prediction performance	Theoretical interpretability	Scalability
Statistical	Low	Low	Low	High	Low
Machine learning	Medium	Medium	Medium	Medium	Medium
Deep learning	High	High	High	Low	High

accurate prediction performance, compared with machine learning and deep learning models. Another drawback is the low scalability, when statistical models are not suitable for large-scale problems and are difficult to be trained in parallel. The major advantages of deep learning models include the high accurate prediction performance and the high scalability, when deep learning models can be trained in a distributed and scalable approach, *e.g.*, with cloud computing. The major drawbacks include the high requirement for both data and computation resources and the low theoretical interpretability, when deep learning models are often criticized as black boxes. Machine learning models are in a middle zone between statistical and deep learning models and can achieve a better tradeoff between different requirements and the prediction performance, especially when the computation resources are limited.

3.4. Auxiliary technologies

In this part, the technologies that are helpful for cellular traffic prediction but cannot be used as predictors themselves are summarized as auxiliary technologies. These technologies, which include mobility prediction, data augmentation, transfer learning, meta learning, and federated learning have been used along with some of the prediction models in the surveyed studies. Compared with the significantly improved predictors, these auxiliary technologies are still in the early stage of use for cellular traffic prediction. However, these technologies have a higher possibility of achieving improved prediction performance and should receive greater focus in future studies.

3.4.1. Mobility prediction

Mobility prediction is useful for spatial–temporal traffic prediction problems when the user movement pattern is the cause for the changes in service usage among different base stations. Deep learning models have also proven to be the best choice for mobility prediction (Jiang & Zhang, 2018). Two different cases have been considered in the surveyed studies that have combined mobility prediction with cellular traffic prediction and proved the effectiveness of the approach. Individual trajectory prediction, which aims to predict the trace of a single user, was used for traffic flow prediction in Karimzadeh et al. (2021). Regional crowd mobility prediction, which aims to predict the change in crowd distribution, was used for cellular data rate prediction in Kirmaz et al. (2020).

3.4.2. Data augmentation

Data augmentation is a technique used to generate new data samples from existing data, thereby eliminating the issue of huge costs associated with data collection. This technique has been mainly used for image classification and recognition problems, and utilizes some simple operations, *e.g.*, flipping, rotation, and cropping. More sophisticated techniques have also been developed, *e.g.*, generative adversarial network (GAN) (Goodfellow et al., 2014) in which one generative neural network is responsible for generating samples created from random noises. Various data augmentation techniques were proposed for a time series and summarized in Wen et al. (2021). In the surveyed studies, GAN was used as the data augmentation technique for cellular traffic data (Wang, Hu et al., 2020) in the step before training the LSTM predictor. It was shown that while the data augmentation process increases the computation cost, the prediction gain is more notable.

3.4.3. Transfer learning

Transfer learning is another technique used for handling the data shortage situation. The machine learning or deep learning model is first trained for a single task using sufficient training data. Then, the trained model is used and fine-tuned in another similar or related task with limited training data, by leveraging the knowledge gained from the data-rich task. Transfer learning is widely used in the computer vision domain, where the CNNs pre-trained with ImageNet are further used in other computer vision tasks.

The data shortage problem appears in the cellular network when new base stations are built and the historical data is inadequate for training an effective model. This situation has been considered in Yaghoubi et al. (2021), where transfer learning is used between the existing base stations and newly built stations. Transfer learning is more effective as an approach based on cross-domain data fusion for cellular traffic prediction, especially when heterogeneous data are used as inputs (Zeng et al., 2020; Zhang et al., 2019). Different types of transfer learning have been used, including inter-cluster transfer learning, which aims to transfer knowledge among different clusters or spatial locations, and cross-domain transfer learning, which aims to transfer knowledge among different services or applications, *e.g.*, SMS, call, and Internet usages in the Telecom Italia dataset.

3.4.4. Meta learning

Meta learning is another approach of transferring knowledge among different tasks. It uses a set of meta-learners that can be combined adaptively in different situations. The meta-learners can be combined with other sophisticated techniques, *e.g.*, reinforcement learning (RL), to better adapt to the specific task. A meta-learning scheme is proposed in He, Moayyedi et al. (2020) for cellular traffic prediction; it consists of a set of predictors, each of which is trained for predicting a particular traffic type. Then, a deep RL based policy is used to choose the best-fit predictor dynamically by considering the recent prediction performance.

3.4.5. Federated learning

Federated learning was proposed to enable distributed model training when data usage is restricted to the local domain without data leakage concern. Only the model parameters or the information needed to update the model parameters, *e.g.*, gradients for updating neural networks, are transmitted and exchanged among the different participants. In cellular traffic prediction, federated learning can be used to avoid increase in traffic data transmission overhead while ensuring data privacy for different users or applications. A dual attention-based federated learning scheme was proposed in Zhang, Dang et al. (2021), which effectively collects the contributions of different client models in different base stations to a global model in the central server.

3.5. Open-source projects

At the end of this section, the open-source projects from the surveyed studies with their published years and links are listed in Table 8 to encourage open research. All four open-source projects are hosted in GitHub. Currently, only four out of the 89 studies have their source codes publicly accessible with no constraints. More open-source projects are encouraged by leveraging the open datasets summarized in this survey.

Table 8
Open-source projects from the studies reviewed in this survey.

Study	Year	Count
Huang et al. (2017)	2017	https://github.com/IPCLab/CDRF
Zhang, Zhang et al. (2018)	2018	https://github.com/chuanting/STDenseNet
Zhang et al. (2019)	2019	https://github.com/zctzzy/STCNet
Zhang, Dang et al. (2021)	2021	https://github.com/chuanting/FedDA

3.6. State-of-the-art performance

In this part, we summarize the state-of-the-art traffic prediction performance from the surveyed literature with those open datasets listed in Table 2, which can be used as baselines in follow-up studies when new prediction models are proposed and evaluated. Both the best model and baselines are listed for performance comparison. It should be mentioned that even for the surveyed studies using the same dataset, their results may not be comparable when different subsets, preprocessing techniques and evaluation settings are used. We show the representative tabular results for Telecom Italia (Barlacchi et al., 2015) and CIKM21-MPGAT Data (Lin, Su et al., 2021), in Tables 9 and 10, respectively. For those results in graphs or using other open datasets, the readers can check the relevant studies listed in Table 2. In Table 9, the predictions for SMS, call, and data usages are evaluated separately. In Table 10, different time intervals are considered. For both Telecom Italia (Barlacchi et al., 2015) and CIKM21-MPGAT Data (Lin, Su et al., 2021), deep learning models achieve the best performance, as shown in bold.

4. Application scenarios

Traffic prediction is not the ultimate objective. Enhanced network management and planning actions are needed for fully leveraging the predicted traffic results. Several applications from the surveyed studies are summarized in this section. First, we discuss the applications for cellular networks, *e.g.*, base station sleeping, admission control, and resource allocation and scheduling. Additional applications in broader scenarios, some of which are beyond the scope of cellular networks, are discussed next, *e.g.*, network dimensioning, network slicing, SDN, and mobile edge computing.

4.1. Base station sleeping

One of the obvious applications of cellular traffic prediction involves the designing of sleeping strategies for the base station. When the traffic demand is low, some of the base stations can be closed or they may operate in a low-function status for energy saving without reducing the service quality for users. Precise cellular traffic prediction plays a core role in designing these strategies, which has been considered in Lin, Chen et al. (2021), Wu et al. (2021) and Zhu and Wang (2021).

4.2. Admission control

While the base station sleeping strategy is designed for the entire base station in the idle time period, differentiated admission control strategies can be designed and used in the busy time period. Admission control aims to allow or prevent specific users or applications from using the network resources, especially when the service ability of a base station is already in a saturated situation. Proactive admission control policies can be designed based on the traffic predictions, instead of taking actions only after the resources are already exhausted. Admission control based on cellular traffic prediction has been considered in Jiang et al. (2021) and Perveen et al. (2021). For example, a fully reconfigurable admission control framework via fuzzy-logic optimization was designed and validated in Perveen et al. (2021).

4.3. Resource allocation and scheduling

The allocation and scheduling of different resource types are considered in He, Dán et al. (2020), Weerasinghe et al. (2019) and Zhao, Qin et al. (2020), in which traffic prediction is used as the basis for designing the follow-up allocation or scheduling schemes. For example, dynamic preamble allocation was considered in Weerasinghe et al. (2019). This application was enabled through the cognitive radio network technique in Zhao, Qin et al. (2020), in which the wireless spectrum resource could be identified and allocated dynamically. A semi-persistent scheduler for downlink scheduling was proposed in He, Dán et al. (2020), which shows similar performance as the traditional proportional-fair scheduling in terms of throughput, fairness, and latency and greatly reduces the computational cost.

4.4. Network dimensioning

Network dimensioning is used for deploying a new network when the minimum capacity requirements for meeting the service agreement need to be determined. For this purpose, the peak-hour traffic, which is the traffic intensity at its peak, can be predicted and used. This application has been considered in Gijón et al. (2021), in which different supervised learning models have been proven more effective than the traditional time series models for peak-hour traffic prediction.

A special case of the on-demand or predictive unmanned aerial vehicle (UAV) networks was considered in Zhang, Mozaffari et al. (2018) and Zhang, Saad et al. (2020). UAVs can be used as temporary base stations, especially when ground base stations are not working, such as during natural disasters. UAVs can also be deployed under traffic overload to complement the ground cellular systems when the downlink data demand is predicted to approach its peak value.

4.5. Network slicing

The purpose of network slicing is to provide end-to-end network transmission, with logically isolated network infrastructures. Network slicing is an important ability in the 5G radio access network, where different slices are set up for providing different services, *e.g.*, eMBB, mMTC, and URLCC slices. Network slicing is often provided by the network operators for the network tenants, who provide services for end users through leased networks. Prediction-assisted network slicing has been considered in the surveyed studies for different scenarios (Abozariba et al., 2020; Cui et al., 2020; Ferreira et al., 2021; Gutterman et al., 2019; Yu, Musumeci et al., 2020; Zhou et al., 2020). For example, the allocation of PRBs for different slice requirements is considered in Gutterman et al. (2019). The 5G RAN slice adjustment and migration in wavelength-division multiplexing (WDM) metro-aggregation networks is considered in Yu, Musumeci et al. (2020), and network slicing in vehicular networks is considered in Cui et al. (2020). Cellular traffic prediction has been proven effective in providing considerable advantages in network slicing, *e.g.*, minimizing the resource allocation costs, maximizing radio resource utilization, and guaranteeing the SLAs for network tenants.

4.6. Software defined networking

SDN is a concept that separates the control and forwarding behaviors of the network, thereby reducing the reliance on hardware equipment and operability of the network. By adding the programming ability, the SDN controller is able to operate white-box switches and other network equipment. The entire network becomes more flexible and it is possible to design prediction-based resource scheduling. Different applications in SDN have been considered along with cellular traffic prediction, *e.g.*, virtual network function placement and resource scaling in Bega et al. (2020), and dynamic optical routing in Alvizu et al. (2017). A more efficient resource allocation scheme was proposed for SDN in Zhan et al. (2021), where the traffic trend of each cell could be predicted in the control domain and the communication loads could be balanced among different cells.

Table 9
State-of-the-art results for Telecom Italia (Barlacchi et al., 2015) with the spatial resolution of 100 × 100 grids and the time interval of 1 h.

Subset	Metric	ARIMA	LSTM	Conv-LSTM	MGCN-LSTM (Lin, Chen et al., 2021)
SMS	RMSE	189.00	144.22	144.86	**81.79**
	MAE	135.31	105.96	106.77	**58.74**
	R^2	0.889	0.917	0.932	**0.978**
Call	RMSE	177.67	164.06	118.42	**72.07**
	MAE	115.04	109.09	85.31	**48.44**
	R^2	0.917	0.937	0.967	**0.985**
Internet	RMSE	919.91	910.65	640.22	**494.95**
	MAE	579.67	611.68	431.53	**374.18**
	R^2	0.908	0.912	0.953	**0.972**

Table 10
State-of-the-art MAPE results for CIKM21-MPGAT Data (Lin, Su et al., 2021) with different time intervals.

Time interval	LSTM	GWNET	MTGNN	STAWNET	MPGAT (Lin, Su et al., 2021)
5 min	0.1592	0.1576	0.1629	0.1611	**0.1511**
10 min	0.1801	0.1781	0.1811	0.1823	**0.1720**
30 min	0.1984	0.1934	0.1951	0.1947	**0.1876**
60 min	0.2342	0.2194	0.2198	0.2180	**0.2149**

4.7. Mobile edge computing

Mobile edge computing is a new computing paradigm that aims to improve the computing and storage capacities at the edge of a cellular network system, reduce the content and service delivery delay, and further improve the quality of experience for end users. It has been proposed for smart cities and traffic prediction is used to facilitate multi-access edge computing (MEC) deployment and resource management (Ale et al., 2021). Mobile traffic offloading was further considered in Huang and Chen (2020), in which the maximum, average, and minimum traffic demands in the coming hour were predicted and the corresponding offloading strategies were designed in a MEC environment.

5. Research directions

In this section, several potential research directions are summarized for future studies. Some of these directions have already been addressed in the surveyed studies but only at a very preliminary level and a comprehensive consideration was therefore lacking. The remaining research directions have not yet been addressed but appear promising based on the insightful knowledge summarized in this survey.

One research direction is to make a fair and comprehensive evaluation of different models across different datasets, *e.g.*, the open datasets listed in Section 2. In most of the surveyed studies, only one dataset was used to evaluate the proposed or adopted prediction model. Although the Telecom Italia dataset has been widely used in the literature, different date ranges and data reprocessing techniques were chosen, preventing a fair comparison of the prediction performance among different studies. In addition, the evaluation metrics vary among the existing studies; the studies did not consider the training time consumption and did not employ metrics developed based on domain knowledge. A better approach is therefore proposed for future research, where multiple open datasets can be used with the same data split ratio and different models should be evaluated with both general as well as domain-specific metrics mentioned in Section 3.2.

The second direction is to combine traffic prediction with its deployment, which has often been neglected by previous studies. Although the problem of improving the prediction performance is crucial, the improved model is useless if it cannot be deployed in practice, *e.g.*, if the deep learning models requires high computational or storage costs. Additionally, the cost of setting up the central server for prediction should also be evaluated along with the historical traffic data collection and transmission costs, model training and maintenance costs, updating and notification cost for sending the predicted results to each base station, etc.

One possible solution is to deploy cellular traffic prediction programs with cloud computing, which supports elastic computing ability. When the prediction model is being trained, a higher computing ability is used. When the prediction model has been trained and deployed, a lower computing ability for prediction is required. Another promising direction is to deploy cellular traffic prediction programs with mobile edge computing. This idea has been considered in Zeb et al. (2021), in which the edge μ-box is used. It merits deeper exploration, especially with the real-world prototype systems.

The third direction is to deepen the combination between the cellular traffic prediction models and the auxiliary technologies listed in Section 3.4, which include mobility prediction, data augmentation, transfer learning, meta learning, and federated learning. The application of these techniques to cellular networks is in an early stage and several research gaps exist, which need to be addressed in the follow-up studies.

Other technologies worth mentioning include RL and blockchain. Most of the current studies have modeled the cellular traffic prediction problems using a supervised learning framework, as we did in Section 2; however, an explicit training stage is needed and the prediction process cannot be conducted without historical traffic data. Another limitation of supervised learning is that the prediction model should yield good results in terms of explicit evaluation metrics, which may not be optimal in specific domains. RL techniques are proposed to enable the prediction model to act as an agent and learn to predict by interacting with the environment (*e.g.*, the base station) and then receiving the reward or penalty from it (*e.g.*; a reward would be achieved in the case of an accurate prediction and a penalty in the case of a service under-provisioning or over-provisioning). This idea was considered in Huang and Chen (2020) and Wu et al. (2021) but is still in an early stage and should therefore be investigated in detail.

Blockchain is a distributed and decentralized linked data structure for data storage and retrieval, which can be used in a series of scenarios that require a secure and reliable distributed solution. Blockchain was proposed for use in the 4G LTE access network, in which each base station acts as a separate node as in a peer-to-peer network (Kurri et al., 2021). As a promising technique for improving the cellular systems, blockchain has the advantages of providing secure and transparent data service and preventing fraudulent subscribers. Blockchain can complement the traffic prediction technique used to provide congestion control ability in a blockchain-enabled cellular network.

6. Conclusion

A comprehensive survey of cellular traffic prediction was presented, and a classification of the reviewed problems and methods was performed, which showed that deep learning models were the dominant solutions in the surveyed studies. This research topic is still in the initial stage, with several interesting ideas worthy of exploration. Various applications based on cellular traffic prediction and the potential research directions were summarized to encourage future studies.

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.

References

Abozariba, R., Naeem, M. K., Asaduzzaman, M., & Patwary, M. (2020). Uncertainty-aware RAN slicing via machine learning predictions in next-generation networks. In *2020 IEEE 92nd vehicular technology conference (VTC2020-Fall)* (pp. 1–6). IEEE.

Aldhyani, T. H., Alrasheedi, M., Alqarni, A. A., Alzahrani, M. Y., & Bamhdi, A. M. (2020). Intelligent hybrid model to enhance time series models for predicting network traffic. *IEEE Access*, *8*, 130431–130451.

Ale, L., Zhang, N., King, S. A., & Guardiola, J. (2021). Spatio-temporal Bayesian learning for mobile edge computing resource planning in smart cities. *ACM Transactions on Internet Technology (TOIT)*, *21*(3), 1–21.

Alsaade, F. W., & Hmoud Al-Adhaileh, M. (2021). Cellular traffic prediction based on an intelligent model. *Mobile Information Systems*, *2021*.

Alvizu, R., Troia, S., Maier, G., & Pattavina, A. (2017). Matheuristic with machine-learning-based prediction for software-defined mobile metro-core networks. *Journal of Optical Communications and Networking*, *9*(9), D19–D30.

Assem, H., Caglayan, B., Buda, T. S., & O'Sullivan, D. (2018). St-dennetfus: A new deep learning approach for network demand prediction. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 222–237). Springer.

Azari, A., Papapetrou, P., Denic, S., & Peters, G. (2019a). Cellular traffic prediction and classification: A comparative evaluation of LSTM and ARIMA. In *International conference on discovery science* (pp. 129–144). Springer.

Azari, A., Papapetrou, P., Denic, S., & Peters, G. (2019b). User traffic prediction for proactive resource management: Learning-powered approaches. In *2019 IEEE global communications conference (GLOBECOM)* (pp. 1–6). IEEE.

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

Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., et al. (2015). A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific Data*, *2*(1), 1–15.

Bega, D., Gramaglia, M., Fiore, M., Banchs, A., & Costa-Perez, X. (2019). DeepCog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting. *IEEE Journal on Selected Areas in Communications*, *38*(2), 361–376.

Bega, D., Gramaglia, M., Perez, R., Fiore, M., Banchs, A., & Costa-Perez, X. (2020). AI-based autonomous control, management, and orchestration in 5G: From standards to algorithms. *IEEE Network*, *34*(6), 14–20.

Bejarano-Luque, J. L., Toril, M., Fernández-Navarro, M., Gijón, C., & Luna-Ramírez, S. (2021). A deep-learning model for estimating the impact of social events on traffic demand on a cell basis. *IEEE Access*, *9*, 71673–71686.

Cai, Y., Cheng, P., Ding, M., Chen, Y., Li, Y., & Vucetic, B. (2020). Spatiotemporal Gaussian process Kalman filter for mobile traffic prediction. In *2020 IEEE 31st annual international symposium on personal, indoor and mobile radio communications* (pp. 1–6). IEEE.

Chen, Z., Jiaze, E., Zhang, X., Sheng, H., & Cheng, X. (2020). Multi-task time series forecasting with shared attention. In *2020 international conference on data mining workshops (ICDMW)* (pp. 917–925). IEEE.

Chen, X., Jin, Y., Qiang, S., Hu, W., & Jiang, K. (2015). Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale. In *2015 IEEE international conference on communications (ICC)* (pp. 3585–3591). IEEE.

Chien, W.-C., & Huang, Y.-M. (2021). A lightweight model with spatial–temporal correlation for cellular traffic prediction in internet of things. *The Journal of Supercomputing*, 1–17.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Cisco (2021). Cisco annual internet report (2018–2023) white paper. Online; accessed 11 October 2021, https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

Clemente, D., Soares, G., Fernandes, D., Cortesao, R., Sebastiao, P., & Ferreira, L. S. (2019). Traffic forecast in mobile networks: Classification system using machine learning. In *2019 IEEE 90th vehicular technology conference (VTC2019-Fall)* (pp. 1–5). IEEE.

Cui, Y., Huang, X., Wu, D., & Zheng, H. (2020). Machine learning based resource allocation strategy for network slicing in vehicular networks. In *2020 IEEE/CIC international conference on communications in china (ICCC)* (pp. 454–459). IEEE.

Deng, T., Wan, M., Shi, K., Zhu, L., Wang, X., & Jiang, X. (2021). Short term prediction of wireless traffic based on tensor decomposition and recurrent neural network. *SN Applied Sciences*, *3*(9), 1–14.

Dommaraju, V. S., Nathani, K., Tariq, U., Al-Turjman, F., Kallam, S., Patan, R., et al. (2020). ECMCRR-MPDNL for cellular network traffic prediction with big data. *IEEE Access*, *8*, 113419–113428.

Fang, L., Cheng, X., Wang, H., & Yang, L. (2018). Mobile demand forecasting via deep graph-sequence spatiotemporal modeling in cellular networks. *IEEE Internet of Things Journal*, *5*(4), 3091–3101.

Feng, J., Chen, X., Gao, R., Zeng, M., & Li, Y. (2018). Deeptp: An end-to-end neural network for mobile cellular traffic prediction. *IEEE Network*, *32*(6), 108–115.

Ferreira, D., Reis, A. B., Senna, C., & Sargento, S. (2021). A forecasting approach to improve control and management for 5G networks. *IEEE Transactions on Network and Service Management*, *18*(2), 1817–1831.

Gao, Y., Wei, X., Zhou, L., & Lv, H. (2019). A deep learning framework with spatial-temporal attention mechanism for cellular traffic prediction. In *2019 IEEE globecom workshops (GC Wkshps)* (pp. 1–6). IEEE.

Gao, Y., Zhang, M., Chen, J., Han, J., Li, D., & Qiu, R. (2021). Accurate load prediction algorithms assisted with machine learning for network traffic. In *2021 international wireless communications and mobile computing (IWCMC)* (pp. 1683–1688). IEEE.

Garrido, L. A., Mekikis, P.-V., Dalgkitsis, A., & Verikoukis, C. (2021). Context-aware traffic prediction: Loss function formulation for predicting traffic in 5G networks. In *ICC 2021-IEEE international conference on communications* (pp. 1–6). IEEE.

Garroppo, R. G., & Callegari, C. (2020). Prediction of mobile networks traffic: enhancement of the nmls technique. In *2020 IEEE 25th international workshop on computer aided modeling and design of communication links and networks (CAMAD)* (pp. 1–6). IEEE.

Gijón, C., Toril, M., Luna-Ramírez, S., Marí-Altozano, M. L., & Ruiz-Avilés, J. M. (2021). Long-term data traffic forecasting for network dimensioning in LTE with short time series. *Electronics*, *10*(10), 1151.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Guo, Y., Wang, S., Zhou, A., Xu, J., Yuan, J., & Hsu, C.-H. (2020). User allocation-aware edge cloud placement in mobile edge computing. *Software - Practice and Experience*, *50*(5), 489–502.

Guo, D., Xia, X., Zhu, L., & Zhang, Y. (2021). Dynamic modification neural network model for short-term traffic prediction. *Procedia Computer Science*, *187*, 134–139.

Gutterman, C., Grinshpun, E., Sharma, S., & Zussman, G. (2019). RAN resource usage prediction for a 5G slice broker. In *Proceedings of the twentieth ACM international symposium on mobile ad hoc networking and computing* (pp. 231–240).

Hachemi, M. L., Ghomari, A., Hadjadj-Aoul, Y., & Rubino, G. (2021). Mobile traffic forecasting using a combined FFT/LSTM strategy in SDN networks. In *2021 IEEE 22nd international conference on high performance switching and routing (HPSR)* (pp. 1–6). IEEE.

He, Q., Dán, G., & Koudouridis, G. P. (2020). Semi-persistent scheduling for 5G downlink based on short-term traffic prediction. In *GLOBECOM 2020-2020 IEEE global communications conference* (pp. 1–6). IEEE.

He, Q., Moayyedi, A., Dán, G., Koudouridis, G. P., & Tengkvist, P. (2020). A meta-learning scheme for adaptive short-term network traffic prediction. *IEEE Journal on Selected Areas in Communications*, *38*(10), 2271–2283.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, C.-W., & Chen, P.-C. (2020). Joint demand forecasting and DQN-based control for energy-aware mobile traffic offloading. *IEEE Access*, *8*, 66588–66597.

Huang, C.-W., Chiang, C.-T., & Li, Q. (2017). A study of deep learning networks on mobile traffic forecasting. In *2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)* (pp. 1–6). IEEE.

Huang, J., & Xiao, M. (2020). Mobile network traffic prediction based on seasonal adjacent windows sampling and conditional probability estimation. *IEEE Transactions on Big Data*.

Jiang, W. (2021a). Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, Article 115537.

Jiang, W. (2021b). Graph-based deep learning for communication networks: A survey. *Computer Communications*.

Jiang, W. (2021c). Internet traffic matrix prediction with convolutional LSTM neural network. *Internet Technology Letters*, Article e322.

Jiang, W. (2021d). Internet traffic prediction with deep neural networks. *Internet Technology Letters*, Article e314.

Jiang, N., Deng, Y., & Nallanathan, A. (2021). Traffic prediction and random access control optimization: Learning and non-learning-based approaches. *IEEE Communications Magazine*, *59*(3), 16–22.

Jiang, W., & Luo, J. (2021). Graph neural network for traffic forecasting: A survey. arXiv preprint arXiv:2101.11174.

Jiang, W., & Zhang, L. (2018). Geospatial data to images: A deep-learning framework for traffic forecasting. *Tsinghua Science and Technology*, *24*(1), 52–64.

Karimzadeh, M., Aebi, R., de Souza, A. M., Zhao, Z., Braun, T., Sargento, S., et al. (2021). Reinforcement learning-designed LSTM for trajectory and traffic flow prediction. In *2021 IEEE wireless communications and networking conference (WCNC)* (pp. 1–6). IEEE.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International conference on learning representations (ICLR '17)*.

Kirmaz, A., Michalopoulos, D. S., Balan, I., & Gerstacker, W. (2020). Mobile network traffic forecasting using artificial neural networks. In *2020 28th international symposium on modeling, analysis, and simulation of computer and telecommunication systems (MASCOTS)* (pp. 1–7). IEEE.

Kuber, T., Seskar, I., & Mandayam, N. (2021). Traffic prediction by augmenting cellular data with non-cellular attributes. In *2021 IEEE wireless communications and networking conference (WCNC)* (pp. 1–6). IEEE.

Kurri, V., Raja, V., & Prakasam, P. (2021). Cellular traffic prediction on blockchain-based mobile networks using LSTM model in 4G LTE network. *Peer-to-Peer Networking and Applications*, *14*(3), 1088–1105.

Li, Y., Ma, Z., Pan, Z., Liu, N., & You, X. (2020). Prophet model and Gaussian process regression based user traffic prediction in wireless networks. *Science China. Information Sciences*, *63*(4), 1–8.

Li, M., Wang, Y., Wang, Z., & Zheng, H. (2020). A deep learning method based on an attention mechanism for wireless network traffic prediction. *Ad Hoc Networks*, *107*, Article 102258.

Lin, J., Chen, Y., Zheng, H., Ding, M., Cheng, P., & Hanzo, L. (2021). A data-driven base station sleeping strategy based on traffic prediction. *IEEE Transactions on Network Science and Engineering*.

Lin, C.-Y., Su, H.-T., Tung, S.-L., & Hsu, W. (2021). Multivariate and propagation graph attention network for spatial-temporal prediction with outdoor cellular traffic. In *Proceedings of the 30th ACM international conference on information and knowledge management*.

Liu, Q., Li, J., & Lu, Z. (2021). ST-tran: Spatial-temporal transformer for cellular traffic prediction. *IEEE Communications Letters*.

Liu, C., Wu, T., Li, Z., & Wang, B. (2021). Individual traffic prediction in cellular networks based on tensor completion. *International Journal of Communication Systems*, Article e4952.

Mahdy, B., Abbas, H., Hassanein, H. S., Noureldin, A., & Abou-zeid, H. (2020). A clustering-driven approach to predict the traffic load of mobile networks for the analysis of base stations deployment. *Journal of Sensor and Actuator Networks*, *9*(4), 53.

Mejia, J., Ochoa-Zezzati, A., & Cruz-Mejia, O. (2020). Traffic forecasting on mobile networks using 3D convolutional layers. *Mobile Networks and Applications*, *25*, 2134–2140.

Nagib, A. M., Abou-Zeid, H., Hassanein, H. S., Sediq, A. B., & Boudreau, G. (2021). Deep learning-based forecasting of cellular network utilization at millisecond resolutions. In *ICC 2021-IEEE international conference on communications* (pp. 1–6). IEEE.

Okic, A., & Redondi, A. E. (2019). Forecasting mobile cellular traffic sampled at different frequencies. In *2019 12th IFIP wireless and mobile networking conference (WMNC)* (pp. 189–195). IEEE.

Perveen, A., Abozariba, R., Patwary, M., & Aneiba, A. (2021). Dynamic traffic forecasting and fuzzy-based optimized admission control in federated 5G-open RAN networks. *Neural Computing and Applications*, 1–19.

Rago, A., Piro, G., Boggia, G., & Dini, P. (2020). Multi-task learning at the mobile edge: An effective way to combine traffic classification and prediction. *IEEE Transactions on Vehicular Technology*, *69*(9), 10362–10374.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Santos, G. L., Rosati, P., Lynn, T., Kelner, J., Sadok, D., & Endo, P. T. (2020). Predicting short-term mobile internet traffic from internet activity using recurrent neural networks. arXiv preprint arXiv:2010.05741.

Shawel, B. S., Debella, T. T., Tesfaye, G., Tefera, Y. Y., & Woldegebreal, D. H. (2020). Hybrid prediction model for mobile data traffic: A cluster-level approach. In *2020 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Shen, W., Zhang, H., Guo, S., & Zhang, C. (2021). Time-wise attention aided convolutional neural network for data-driven cellular traffic prediction. *IEEE Wireless Communications Letters*.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, *28*.

Sudhakaran, S., Venkatagiri, A., Taukari, P. A., Jeganathan, A., & Muthuchidambaranathan, P. (2020). Metropolitan cellular traffic prediction using deep learning techniques. In *2020 IEEE international conference on communication, networks and satellite (Comnetsat)* (pp. 6–11). IEEE.

Sun, S. C., & Guo, W. (2021). Forecasting wireless demand with extreme values using feature embedding in gaussian processes. In *2021 IEEE 93rd vehicular technology conference (VTC2021-Spring)* (pp. 1–6). IEEE.

Swedha, S., & Gopi, E. (2021). LSTM network for hotspot prediction in traffic density of cellular network. In *Machine learning, deep learning and computational intelligence for wireless communication* (pp. 35–47). Springer.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45.

Tran, Q. T., Hao, L., & Trinh, Q. K. (2020). A comprehensive research on exponential smoothing methods in modeling and forecasting cellular traffic. *Concurrency Computations: Practice and Experience*, *32*(23), Article e5602.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *International conference on learning representations*.

Wang, Z., Hu, J., Min, G., Zhao, Z., & Wang, J. (2020). Data-augmentation-based cellular traffic prediction in edge-computing-enabled smart city. *IEEE Transactions on Industrial Informatics*, *17*(6), 4179–4187.

Wang, J., Tang, J., Xu, Z., Wang, Y., Xue, G., Zhang, X., et al. (2017). Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach. In *IEEE INFOCOM 2017-IEEE conference on computer communications* (pp. 1–9). IEEE.

Wang, H., Wang, L., Zhao, S., & Yue, X. (2021). Multi-dimensional prediction model for cell traffic in city scale. *International Journal of Pattern Recognition and Artificial Intelligence*, *35*(03), Article 2150010.

Wang, L.-N., Zang, C.-R., & Cheng, Y.-Y. (2020). The short-term prediction of the mobile communication traffic based on the product seasonal model. *SN Applied Sciences*, *2*(3), 1–9.

Wang, W., Zhou, C., He, H., Wu, W., Zhuang, W., & Shen, X. S. (2020). Cellular traffic load prediction with lstm and gaussian process regression. In *ICC 2020-2020 IEEE international conference on communications (ICC)* (pp. 1–6). IEEE.

Wang, X., Zhou, Z., Xiao, F., Xing, K., Yang, Z., Liu, Y., et al. (2018). Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing*, *18*(9), 2190–2202.

Weerasinghe, T. N., Balapuwaduge, I. A., & Li, F. Y. (2019). Supervised learning based arrival prediction and dynamic preamble allocation for bursty traffic. In *IEEE INFOCOM 2019-IEEE conference on computer communications workshops (INFOCOM WKSHPS)* (pp. 1–6). IEEE.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., et al. (2021). Time series data augmentation for deep learning: A survey. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 4653–4660). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2021/631, Survey Track.

Wu, Q., Chen, X., Zhou, Z., Chen, L., & Zhang, J. (2021). Deep reinforcement learning with spatio-temporal traffic forecasting for data-driven base station sleep control. *IEEE/ACM Transactions on Networking*, *29*(2), 935–948.

Xia, H., Wei, X., Gao, Y., & Lv, H. (2019). Traffic prediction based on ensemble machine learning strategies with bagging and lightgbm. In *2019 IEEE international conference on communications workshops (ICC workshops)* (pp. 1–6). IEEE.

Xing, X., Lin, Y., Gao, H., & Lu, Y. (2021). Wireless traffic prediction with series fluctuation pattern clustering. In *2021 IEEE international conference on communications workshops (ICC workshops)* (pp. 1–6). IEEE.

Yaghoubi, F., Catovic, A., Gusmao, A., Pieczkowski, J., & Boros, P. (2021). Traffic flow estimation using LTE radio frequency counters and machine learning. arXiv preprint arXiv:2101.09143.

Yamada, Y., Shinkuma, R., Sato, T., & Oki, E. (2018). Feature-selection based data prioritization in mobile traffic prediction using machine learning. In *2018 IEEE global communications conference (GLOBECOM)* (pp. 1–6). IEEE.

You, X., Zhang, C., Tan, X., Jin, S., & Wu, H. (2019). AI for 5G: research directions and paradigms. *Science China. Information Sciences*, *62*(2), 1–13.

Yu, L., Li, M., Jin, W., Guo, Y., Wang, Q., Yan, F., et al. (2020). Step: A spatio-temporal fine-granular user traffic prediction system for cellular networks. *IEEE Transactions on Mobile Computing*.

Yu, H., Musumeci, F., Zhang, J., Tornatore, M., Bai, L., & Ji, Y. (2020). Dynamic 5G RAN slice adjustment and migration based on traffic prediction in WDM metro-aggregation networks. *IEEE/OSA Journal of Optical Communications and Networking*, *12*(12), 403–413.

Zeb, S., Rathore, M. A., Mahmood, A., Hassan, S. A., Kim, J., & Gidlund, M. (2021). Edge intelligence in softwarized 6G: Deep learning-enabled network traffic predictions. arXiv preprint arXiv:2108.00332.

Zeng, Q., Sun, Q., Chen, G., & Duan, H. (2021). Attention based multi-component spatiotemporal cross-domain neural network model for wireless cellular network traffic prediction. *EURASIP Journal on Advances in Signal Processing*, *2021*(1), 1–25.

Zeng, Q., Sun, Q., Chen, G., Duan, H., Li, C., & Song, G. (2020). Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data. *IEEE Access*, *8*, 172387–172397.

Zhan, S., Yu, L., Wang, Z., Du, Y., Yu, Y., Cao, Q., et al. (2021). Cell traffic prediction based on convolutional neural network for software-defined ultra-dense visible light communication networks. *Security and Communication Networks*, *2021*.

Zhang, C., Dang, S., Shihada, B., & Alouini, M.-S. (2021). Dual attention-based federated learning for wireless traffic prediction. In *IEEE INFOCOM 2021-IEEE conference on computer communications* (pp. 1–10). IEEE.

Zhang, D., Liu, L., Xie, C., Yang, B., & Liu, Q. (2020). Citywide cellular traffic prediction based on a hybrid spatiotemporal network. *Algorithms*, *13*(1), 20.

Zhang, Q., Mozaffari, M., Saad, W., Bennis, M., & Debbah, M. (2018). Machine learning for predictive on-demand deployment of UAVs for wireless communications. In *2018 IEEE global communications conference (GLOBECOM)* (pp. 1–6). IEEE.

Zhang, C., & Patras, P. (2018). Long-term mobile traffic forecasting using deep spatio-temporal neural networks. In *Proceedings of the eighteenth ACM international symposium on mobile ad hoc networking and computing* (pp. 231–240).

Zhang, Q., Saad, W., Bennis, M., Lu, X., Debbah, M., & Zuo, W. (2020). Predictive deployment of UAV base stations in wireless networks: Machine learning meets contract theory. *IEEE Transactions on Wireless Communication*, *20*(1), 637–652.

Zhang, C., Zhang, H., Qiao, J., Yuan, D., & Zhang, M. (2019). Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE Journal on Selected Areas in Communications*, *37*(6), 1389–1401.

Zhang, C., Zhang, H., Yuan, D., & Zhang, M. (2018). Citywide cellular traffic prediction based on densely connected convolutional neural networks. *IEEE Communications Letters*, *22*(8), 1656–1659.

Zhang, J., Zuo, X., Xu, M., Han, J., & Zhang, B. (2021). Base station network traffic prediction approach based on LMA-DeepAR. In *2021 IEEE 6th international conference on computer and communication systems (ICCCS)* (pp. 473–479). IEEE.

Zhao, S., Jiang, X., Jacobson, G., Jana, R., Hsu, W.-L., Rustamov, R., et al. (2020). Cellular network traffic prediction incorporating handover: A graph convolutional approach. In *2020 17th annual IEEE international conference on sensing, communication, and networking (SECON)* (pp. 1–9). IEEE.

Zhao, D., Qin, H., Song, B., Han, B., Du, X., & Guizani, M. (2020). A graph convolutional network-based deep reinforcement learning approach for resource allocation in a cognitive radio network. *Sensors*, *20*(18), 5216.

Zhou, J., Zhao, W., & Chen, S. (2020). Dynamic network slice scaling assisted by prediction in 5G network. *IEEE Access*, *8*, 133700–133712.

Zhu, Y., & Wang, S. (2021). Joint traffic prediction and base station sleeping for energy saving in cellular networks. In *ICC 2021-IEEE international conference on communications* (pp. 1–6). IEEE.