International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Aug-2023)

http://dx.doi.org/10.12785/ijcds/140147

Traffic Flow Prediction Using Big Data and Geographic Information Systems: A Survey of Data Sources, Frameworks, Challenges, and Opportunities

Sayed A. Sayed[1], Yasser Abdelhamid[2] and Hesham A. Hefny[3]

[1,2,3]*Computer Science Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt*

Received 18 Dec. 2022, Revised 07 Jul. 2023, Accepted 22 Jul. 2023, Published 01 Aug. 2023

Abstract: Big Data has been utilized extensively in numerous fields, including transportation. Using several sources of data, traffic conditions can be accurately anticipated and forecasted to enhance the overall operation's efficacy. This study introduces and outlines the different data types that can be utilized in traffic prediction. In addition, the most popular Big Data processing frameworks will be characterized. It also reviews the applications of Big Data in traffic flow prediction. Furthermore, a conceptual framework to adopt Big Data and Geographic Information Systems in traffic prediction is proposed in this study. To further encourage Big Data utilization in traffic prediction operations, future research obstacles, and opportunities are also provided in this study.

Keywords: Big Data; Traffic Prediction; Conceptual Framework; Processing Frameworks; Transportation

1. INTRODUCTION

Intelligent Transportation (IT) improves our travel and makes it more comfortable and effective. Data on traffic flows may now be collected thanks to advances in mobile Internet and positioning technology; this information could be used to improve traffic forecasts and other aspects of intelligent transportation. It is reasonable to use collected data to improve traffic flow prediction and provide consumers with easy transportation services. In other words, the objective is to intelligently improve transportation using recorded traffic flow data [1]. Traffic flow prediction can direct traveling behavior, thereby reducing road network congestion, enhancing travel efficiency, and promoting traffic safety. A key component of data-driven intelligent transportation systems is the accurate and efficient prediction of traffic flow. The traffic flows on a transportation network are mutually interconnected and interactive. It is challenging to characterize traffic flow dynamics using analytical methods. Common methods for estimating the traffic flow on a road retrieve only the time series properties and exclude the close linkages and dynamic evolution between link flows [2]. Big data (BD) has been utilized extensively in numerous fields. The transportation industry is one of them. Using several data sources, traffic flows can be accurately anticipated and forecasted to enhance the performance of the overall operation. Transportation is an example of the ideal application domain where BD has been excessively implemented. Regarding Intelligent Transportation Systems (ITS), the growing infrastructures of the Internet of Things (IoT), Cyber-Physical Systems (CPS), and smart cities facilitated the acquisition of massive amounts of data via surveillance cameras, mobile devices, and static sensors. The size of data has increased from Trillion bytes to Petabytes[3]. This study will concentrate on predicting traffic flow, which is considered a serious challenge among various applications. In some instances, traffic information (volume, speed, travel time, etc.) cannot be retrieved without manipulation of the raw data, even though it is possible to obtain a variety of data about traffic conditions. This procedure is known as the problem of traffic estimation, and it aims to retrieve traffic data accurately from any raw data. Despite assessing traffic information and providing us with historical data only, anticipating the future based on historical data and taking necessary action are among the most essential applications of BD. Statistical models, Machine Learning (ML) models, and Deep Learning (DL) models are just some of the methods that have been created for traffic estimate and prediction challenges, with the latter gaining favor due to their improved performance. Because DL models benefit from a large training dataset, BD aids in their development [4]. Recently, numerous researchers have focused their attention on the ITS's many technological components. Among these tools is the Geographic Information System (GIS). These investigations led to the development of increasingly in-demand GIS for Transportation (GIS-T) technologies. When it comes to storing and organizing transportation data, GIS mapping is an indispensable, time-saving resource [5], [6]. In this paper, a systematic review of the potential BD role

E-mail address: sayed.ahmed@narss.sci.eg, yasserabdelhamid@gmail.com, hehefny@cu.edu.eg

in traffic prediction was introduced to answer the following questions. The first question indicates the most popular BD sources that can be utilized in traffic flow prediction. The second question trending non-spatial and spatial BD Framework that are commonly used in such cases. The Third question indicates how GIS, BD sources, and BD processing frameworks can be integrated to enhance traffic prediction performance and accuracy. Therefore, A conceptual framework to adopt BD and GIS in traffic prediction was presented. What follows is a brief overview of the most relevant contributions from this investigation. First, Various data sources employed for predicting traffic flow are introduced and summarized. Second, the most popular BD processing frameworks will be characterized and grouped based on the types of data they handle. Third, a review of the applications of big data in traffic flow prediction will be provided. Fourth, a conceptual framework to adopt BD in traffic prediction is presented. Finally, the obstacles and opportunities of utilizing BD for traffic assessment and prediction are highlighted. Here's how the rest of the paper is organized: a theoretical background about BD, traffic prediction, and GIS is presented in Section 2. Section 3, discusses the different BD sources that can be employed in traffic prediction. Section 4, discusses the off-the-shelf BD processing frameworks that can be employed in traffic prediction tasks. Section 5, outlines the survey methodology and presents a literature review of the various forecasting techniques for traffic flow prediction using BD and GIS technologies. In Section 6, a conceptual framework for employing BD in traffic flow prediction tasks is proposed. Section 7, discusses the obstacles as well as future opportunities for utilizing BD in traffic prediction. Finally, the paper is concluded in Section 8.

2. BACKGROUND

The adoption of several efficient traffic data acquisition systems and the development of BD technologies for the storing and processing of massive amounts of data have permitted the deployment of many methods for traffic volume prediction [7]. Due to the complexity of urban development concerns, BD technologies, and GIS have become indispensable for managing urban infrastructure and devising municipal development policies.

A. *Big Data*

The term ”Big Data” was introduced by Michael Cox and David Ellsworth in [8], who explained it as follows: ”Visualization offers an extra conundrum for computing systems: data sets are often fairly large, testing the capacity of main memory, local disc, and even remote storage”. The term ”Big Data” was used in the middle of 2011 to describe a specific type of data that is extremely large and varied, making it very challenging to manage and handle with conventional tools and techniques [9]. An alternate definition of ”Big Data” is the technology paradigm that enables academics to efficiently analyze the huge quantities of data made available by contemporary practices [10], [11]. The characteristics of BD include volume, velocity, variety, veracity, and value. The volume of data collected can reach tens or even hundreds of petabytes. Velocity refers to the rapid rate of data reception and accumulation, as well as the increasing need with which data must be acted upon. The variety here refers to the availability of a wide range of data formats, including semi-structured and unstructured forms such as audio, video, and text. These types of data frequently necessitate further processing. Veracity is the precision and quality of data. In the surrounding world, the data can be disorganized, particularly in the context of BD as data sizes and sources expand. The final element is value [3]. Data value is the most significant dimension, as it measures the utility of data [12]. The BD paradigm includes the resources, infrastructure, and safeguards that makeup BD management [13]. BD paradigm allows for the analysis of vast quantities of data. It is comprised of four components: storage, processing, representation, and methods [14]. They strive to uncover hidden tendencies and patterns within enormous quantities of data from multiple sources. BD storage offers management strategies and tools for the storage of structured and unstructured data. Cloud-based processing is made possible by the infrastructure made available by BD processing. BD Representation offers tools for building real-time data dashboards and visualizations. Finally, BD methods are concerned with data acquisition to discover hidden trends and patterns [15], [16], [17], [18]. In BD, processing data in real-time is a critical issue. The necessity to analyze and respond to real-time streaming data, such as traffic statistics, via continuous queries to enable on-the-fly analysis while in-stream is the backbone of big data streaming analytics. The process of BD stream analysis begins with the ingestion of data in the form of an infinite tuple, continues with the analysis of that data, and concludes with the production of useful outputs, which are often presented in the form of an output stream [19].

B. *Traffic Prediction*

Currently, transportation networks have evolved swiftly, and increasingly sophisticated modes of transportation have emerged. Almost all cities in the world are plagued by severe traffic congestion. The daily influx of heavy traffic paralyzes the urban transportation system, which has a negative influence on people's ability to travel [20]. Hence, transportation systems need to be intelligibly managed. Fortunately, ITS provides these features. ITS integrates and uses information, communications, computers, and other technology in the transportation sector. It seeks to integrate people, roads, and vehicles via sophisticated data communication technology [21]. Predicting the number of vehicles moving through an area of traffic requires investigating historical data and making an educated guess as to how many will be moving through the area at some point in the future [22]. Numerous elements, such as weather conditions and the possibility of accidents, can contribute to traffic congestion in a huge and complex transportation system. Therefore, precise prediction becomes extremely difficult. Therefore, comprehensive consideration must be given to the forecasting and analysis of traffic flow. Historically, the

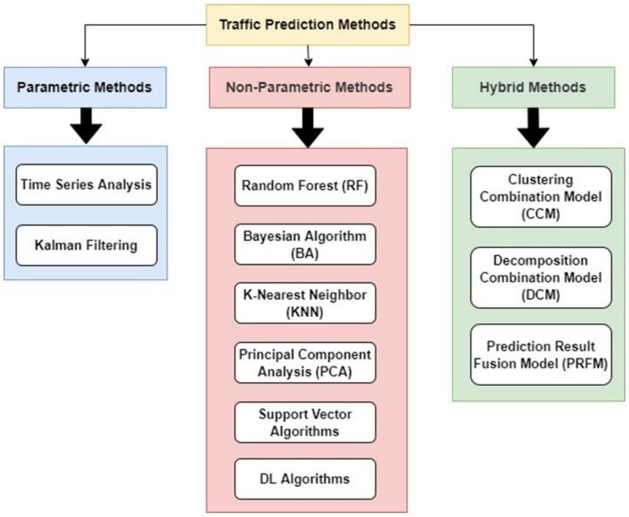

Traffic Prediction Methods
Parametric Methods
Non-Parametric Methods
Hybrid Methods
Time Series Analysis
Kalman Filtering
Random Forest (RF)
Bayesian Algorithm (BA)
K-Nearest Neighbor (KNN)
Principal Component Analysis (PCA)
Support Vector Algorithms
DL Algorithms
Clustering Combination Model (CCM)
Decomposition Combination Model (DCM)
Prediction Result Fusion Model (PRFM)

Figure 1. Traffic Prediction Methods

traffic flow change has a periodicity that can be every day, week, quarter, or year. Consequently, the traffic flow will have roughly periodic variations because of external variables. Since traffic patterns tend to repeat themselves, we can reliably predict future traffic volumes [21]. Parametric methods, nonparametric methods, and hybrid methods are the three main types of traffic flow forecast approaches [23], as shown in figure 1. Parametric methods such as Time Series Analysis (TSA) and Kalman Filtering (KF). TSA aims to find temporal patterns in a set of observed readings x at a given time t and use these results to make predictions. Another model that can help with regression difficulties and lower variance to get to the best possible outcomes in mobile stochastic challenges is KF. Non-Parametric methods rely on ML and DL techniques in which the prediction model is pre-trained with a set of training data and then tested and validated using new data. Support Vector Algorithms (SVA), Random Forest (RF), K-Nearest Neighbor (KNN), Bayesian Algorithm (BA), Principal Component Analysis (PCA), and DL Algorithms are all non-Parametric techniques. The hybrid forecasting approach is the integration of two or more forecasting techniques to produce joint forecasting. Models for clustering, decomposition, and fusing predictions are all utilized in this approach [24].

C. Geographic Information Systems and Traffic

Due to the variety of urban development concerns, GIS and information technology have become indispensable not just for managing urban infrastructure but also for developing municipal development policies. A GIS is a system that is built to collect, store, manipulate, analyze, manage, and show all kinds of data based on their location [25]. In the past fifty years, GIS software was mostly used to automate the manufacturing of maps without offering many analytical capabilities. Currently, GIS combines geospatial and non-geospatial data with decision support and expert systems. Using GIS integrated with traffic information, trip flows, and traffic conditions including accidents or congestion can be monitored in real-time. In addition, by evaluating historical data, the system must be able of recognizing traffic flow patterns and flag trouble locations where overcrowding occurs at specific periods and their possible causes. Therefore, experts in the fields of transportation are rapidly incorporating conventional ways of research in this field with spatial analysis, giving a variety of applications for both conventional and innovative modes of transportation [26]. GIS allows users to combine a range of geospatial data associated with transport and traffic, such as type of roads, road width, types of sidewalks, number of lanes, lanes and intersection capacities, velocity limits, incidents, vehicles, etc. GIS is beneficial to the subject of traffic not only due to the spatial and distributed characteristics of transport data, but also due to the requirement for many forms of network-level, statistical, and spatial analysis. In addition, GIS-based systems allow the integration of socioeconomic data with spatial information on the road network for a range of planning purposes [27]. Due to its various benefits, GIS is an ideal solution for modeling and resolving traffic problems, particularly because of its capacity to handle enormous volumes of geographic data. Consequently, GIS systems have progressed beyond the early use of data management and mapping to include modeling and analysis, thereby facilitating geographical decision-making. Several attempts have been made to incorporate GIS with the development of ITS. As a result of this integration, GIS-T has emerged. The primary benefit of GIS-T is the potential incorporation of all spatial and tabular data types. Because of its focus on the transportation system's relationship to its surroundings, GIS is well-suited for tasks including route planning, risk analysis, and decision-making [28], [24]. Additionally, GIS-T can be combined with complex mathematical models and simulation techniques meant to examine alternative planning and management practices. Nowadays, GIS-T is now used in several different applications such as traffic modeling, route planning, accident analysis, and impact assessment.

3. BIG DATA SOURCES

In this section, the major BD sources will be described first, followed by a summary of the various data sources that can be employed for traffic prediction.

A. Major BD Sources

- Satellite Imagery: Satellite imagery can be defined as images of the Earth's surface captured by imaging satellites. Usually, these data are captured by active or passive sensors [29]. Using passive sensors, the gathered images determine the amount of the sun's reflected sunlight. On the other hand, images are often obtained using active sensors. In conditions of cloud cover, precipitation, and darkness, these sensors are unable to detect the reflectance of Earth's surface objects. Active sensors like the Synthetic Aperture Radar (SAR) are used effectively to increase their observational capacity and so help scientists get over these kinds of limitations. Examples of satellite imagery datasets are (1) USGS Earth Explorer [30] which is a world-class source of data that offers the latest satellite view like Landsat, Spy Satellite, and Hyperspectral, and (2) Sentinel Open Access Hub repository [31] from which Sentinel images can be downloaded. With Sentinel-2, open and free satellite photography enters a new and exciting era. It's not only that every corner of the earth is covered. But another reason is that Sentinel-2 offers recent, high-resolution satellite images.

- Wireless Sensor Network and IoT: The term ”Wireless Sensor Network” (WSN) refers to a collection of sensors that may be placed almost anywhere to monitor environmental conditions and report back information like temperature, humidity, wind speed, and more. IoT describes the integration of Internet and web technologies into the physical world through the widespread use of geographically dispersed devices that contain recognition, sensing, and/or actuation capabilities. The IoT is a vision for the future that connects disparate digital and physical objects through relatively low-cost and accessible Information Communication and Technology (ICT), paving the way for a whole new set of software tools and services [32]. Examples of WSN and IoT datasets are (1) WSN-DS [33] which is a dataset for intrusion detection systems in wireless sensor networks, and (2) WSN-Indfeat-Dataset [34] which is a set of data about network features taken from WSN measurements taken in an industrial setting from June 5 to June 6, 2014. Network monitoring includes metrics from the Physical, Media Access Control (MAC), and Network (NWK) layers, as well as temperature and humidity readings from ambient sensors on each sensor node and the voltage threshold.

- Social Media and Crowdsourcing: Public contributions are well-known in both crowdsourcing and social media data. Contributors to a crowdsourced project are likely aware that they are providing data. However, social media data are supplied passively, meaning that the contributors are usually oblivious to the fact that they are providing data [35]. Platforms for active crowdsourcing have been developed so that individuals can eagerly contribute the data that is required. Members of the affected public or Non-Governmental Organizations (NGOs) construct and run most of these platforms [36], [37]. Despite widespread popularity, active crowd-sourcing platforms continue to struggle with issues including data trustworthiness and inclusion in decision-making [38]. Real-time BD analytics depending on social media networks [39] offer substantial opportunities for intelligent traffic forecast and monitoring. Visual analytics based on data from social media can facilitate Spatiotemporal analysis and produce spatial decisions for the supporting environment. The social network relies not only on text communications but also on user-posted videos and photographs. Using image/video-based analysis and visual analytics, social media posts are mined for vital information [14]. Examples of crowdsourced datasets are (1) 2015 New Year's Resolutions [40] which is a Twitter sentiment analysis of users' 2015 New Year's resolutions. Contains demographic and geographical data of users and resolution categorizations, and (2) Academy Awards Demographics [41] which is a dataset concerning the race, religion, age, and other demographic details of all Oscar winners since 1928 in the following categories Best Actress, Best Supporting Actor, Best Supporting Actress, and Best Director. Examples of social media datasets are (1) Social Influence on Shopping [42] which is a survey of 2,676 millennials that investigates the impact of social platforms on online shopping. This data was collected on the social survey mobile platform Whatsgoodly which has 300,000 millennial and Gen Z members and collected 150,000,000 survey responses from this demographic to date, and (2) Airline Twitter Sentiment [43] which is a sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as ”late flight” or ”rude service”).

- Big Traffic Data: This section discusses the various sources of data that can be utilized for traffic forecasting. There are various approaches to classifying big traffic data. Previously, the big traffic data were categorized into unsupervised and supervised types, with supervised data such as loop detectors and Global Positioning System (GPS) traces representing direct sources of traffic information, and unsupervised data such as call detail records as well as cell phone location data representing indirect sources from which traffic information can be inferred. In this review, big traffic data will be further categorized based on data sources.

- Travel Surveys: Typically collected by municipal authorities or researchers, travel surveys are in-depth questionnaires regarding mobility habits. When diverse travel modes are evaluated and location privacy is respected, it is sometimes the only choice to quantify and comprehend the evolving daily travel habits of individuals effectively. The advantages of travel surveys are high-resolution and comprehensive data, which directly include trip information and reduce the necessity of traffic prediction. While real-time applications, such as traffic prediction, can benefit greatly from the information provided by travel surveys, these surveys have several limitations, including a relatively small sample size, a restricted temporal and geographic scale, self-reporting errors, and a high cost to gather. Examples of publicly available travel survey datasets are (1) ”My Daily Travel Survey” [44] that published by the Chicago Metropolitan Agency for Planning between August 2018 and April 2019, Households in north-eastern Illinois were polled on their travel habits, including how often they drove to and from work, school, stores, and social events, and (2) The National Renewable Energy Laboratory (NREL) which conducted the ”California Household Travel Survey (CHTS)” between 2010 and 2012 [45]. More than 42,500 households participated in the survey, making it the largest of its kind in the United States. Data was collected through a variety of channels, including computer-assisted telephone interviewing, online and mail surveys, wearable (7574 participants) and in-vehicle (2910 vehicles) GPS devices, and on-board diagnostic sensors that gathered information directly from a vehicle’s engine.

- Demographic Information and Official Statistics: Demographic information and official statistics can also include traffic information as well as extra data essential for addressing traffic issues, such as the place of a person’s residence and employment. Typically, governments collect this information regularly. Demographic information and official statistics have a big volume and extensive coverage, which can typically be the entire nation. The drawbacks are: 1) Flows are collected at the levels of coarse municipal or county, 2) There is a lack of traffic information available, and 3) It is costly and time-consuming to acquire such information through travel surveys. With a few exceptions, official statistics and demographic data are currently not commonly employed for traffic estimation and forecast activities [46]. Examples of publicly available datasets are (1) The US Census Bureau’s Commuting Flows [47] in which the primary workplace location from the respondents are collected, and (2) The Internal Revenue Service’s (IRS) Migration Data [48] in which both inflows and outflows of migration patterns are available for filing years 1991 through 2018 with the IRS.

- Traffic Sensors Data: For traffic data collection, static sensors such as radar sensors, acoustics sensors, infrared sensors, and inductive loop detectors, can be utilized. Loop sensors represent the most known technology of many road sensors and offer numerous well-known open datasets for traffic prediction. The advantages of traffic sensor data include their capacity to collect enormous volumes of data automatically and constantly, typically in minutes or even seconds. On the other hand, the disadvantages of these data are their constrained spatial coverage and the issue of missing data produced by sensors’ insufficiency. In addition, the traffic flow speed collected via traffic sensors may be a measurement point that is unsuitable for calculating the average travel time along the road link. Examples of publicly available travel survey datasets are (1) Caltrans Performance Measurement System (PeMS) [49], in which traffic data is collected in real-time from over 39,000 individual detectors. These sensors span the freeway system across all major metropolitan areas of the State of California. It covers the period from 2001 to 2019, and (2) The ”Traffic Flow Madrid [50] published by the Open Data Portal of the Madrid City Council which publishes the data in different files formats such as CSV, XLS, XML, TXT, SHP, KML, KMZ, and other different formats. This dataset covers the period from 2013 till now.

- Call Detail Records: Information about a phone call, including its location, is stored in a Call Detail Record (CDR) and is generated by the cell phone company. CDRs are generated by a mobile carrier to retain details of calls traveling via a device and typically include several properties of the call, such as source, destination, timestamp, duration, and tower. CDRs also contain comparable information regarding human mobility patterns. CDRs offer a crude locating method. Not only is data preparation required for eliminating distortion, but also for determining the appropriate spots. CDR data, or data like it, has many benefits, including its widespread availability (through mobile phones), massive volume, and varied characteristics (social, mobile, time, demographics). The disadvantages, however, include weak locating capabilities, noise, and the need for pre-processing. Considering that location information is only available from cell towers, it is only partially detected when making calls and is only known on a tower-by-tower basis. On the other hand, the ping-pong effect between towers might muddy the data [3]. Examples of publicly available CDRs datasets are (1) The OpenCellID [51], which aims to build a community-driven and -inspired open cellular dataset. Customers all over the world use this cellular data for a wide variety of commercial and personal applications. OpenCellID supports it by offering easy access to the data via an Application Programming Interface (API) and data dumps, which can be used for everything from identifying devices to comprehending network coverage patterns. This dataset is regularly updated, and (2) The Telecommunications - SMS, Call, Internet - MI [52] dataset, which provides information regarding the city of Milan’s telecommunications activity. The dataset is the outcome of a computation performed on the CDRs produced by the cellular network of Telecom Italia in Milan. CDRs record user activities for billing and network administration.

- GPS Trajectory Data: Floating vehicles that are integrated with onboard communication devices and positioning systems commonly acquire GPS trajectory data. In addition, smartphones can be used to obtain GPS trajectory data, such as through crowdsourcing. GPS trajectory data are simple to acquire, but they require pre-processing processes, such as position recognition to assemble points in a meaningful position and trajectory segments to divide a trajectory into sub-trajectories. Using map matching, the GPS coordinates are also mapped onto the road network. Most public GPS trajectory data are acquired from taxis due to privacy concerns. In many situations, just the predicted traffic status determined by the GPS traces is made available to the public. The benefits of utilizing GPS trace data involve high temporal and spatial precision in addition to the ability to track the complete traces. Taxis may also be viewed as mobile sensors that reflect the entire road traffic situation. Like CDRs, the disadvantages include the need for pre-processing and the distortion in the data. In addition, GPS traces data are frequently acquired at a high rate, such as per second, necessitating storage and processing capabilities that can only be provided by big data technologies. Examples of publicly available GPS traces datasets are (1) GeoLife GPS Trajectories [53], which were collected by 182 users over the world of more than three years as part of the Microsoft Research Asia GeoLife project. It covers the period from April 2007 to August 2012. In this set of data, a GPS track is represented by a series of time-stamped points, each of which has information about its latitude, longitude, and height. This set of data has 17,621 trajectories with a total length of more than 48,000 hours and about 1.2 million kilometres. (2) Mobile Data Challenge (MDC) [54], which consists of large quantities of continuous data about the behaviour of individuals and social networks, recorded via mobile phones from 2009 to 2011 in the Lausanne/Geneva area. About 200 persons participated in the data-collecting campaign.

- Big Location-Based Services Data: With the emergence of Mobile Internet Technology (MIT), data from Location Based Services (LBS) are generated via the GPS feature of smartphones. From the location-based service providers, such as Google Maps, Baidu Maps, check-ins, geotagged tweets, and micro-blogs, various location-based BD is obtained. Typically, these data are gathered through a crowdsourcing method in which a traffic information extraction process is required. The benefits of employing LBS BD include their semantic information and widespread availability, such as restaurants, shopping centers, etc. Nevertheless, drawbacks are also present. Whenever geotagged messages or check-ins are made, only a portion of human motion may be identified. Thus, the data go sparse and may be sparser than CDRs. Self-selection bias also exists, which would result in erroneous traffic data. Examples of publicly available LBS datasets are (1) Baidu-Traffic dataset [55] which includes data about traffic speed acquired by navigation apps covering the period from 1 April 2017 to 31 May 2017, in Beijing, China, and (2) Foursquare Dataset [56] which has two subsets: A) New York City's (NYC) Restaurant Rich Dataset (Check-ins, Tips, Tags), which has data from Foursquare about check-ins, tips, and tags at restaurants in NYC from October 24, 2011 to February 20, 2012. It has 3112 users, 3298 places where people have checked in, 27149 check-ins, and 10377 tips. B) NYC and Tokyo Check-in Dataset, which has about 10 months of check-ins in NYC and Tokyo (from 12 April 2012 to 16 February 2013). It has 227,428 New York City check-ins and 573,703 Tokyo check-ins. Each check-in is linked to its time stamp, GPS coordinates, and meaning (represented by fine-grained venue-categories).

- Public Transport Activities Data: Several types of public transportation, especially those with automatic wage systems, can acquire activity data to assess and predict traffic. For example, Origin-Destination (OD) estimates [57] often use data from smart cards. A two-step method was used for transportation analysis [58] that uses bus routes and smart card data to reconstruct individual journeys and then cluster these trips to identify transit nodes. Researchers in China were able to ascertain the goal of dockless shared-bike users by analyzing a public bike dataset using a gravity model and Bayes rules. The incorporation of activity form ratio and facility of the Point of Interest (POI) was shown to be helpful for the inference [59]. The advantages of utilizing public transportation transaction data include their widespread availability and strong relationship to traffic state. If the temporal and geographic ranges of the data are expansive, the storage and processing requirements can be prohibitive. Examples of publicly available Public Transport Activities datasets are (1) TaxiNYC [60] published by The NYC Taxi and Limousine Commission (TLC). It covers the period from 2009 till now. It is available in different formats such as CSV, SHP, and JPG, and (2) UberNYC [61], which contains data on over 4.5 million Uber pickups in NYC from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. Trip-level data on 10 other For-Hire Vehicles (FHV) companies, as well as aggregated data for 329 FHV companies, is also included.

- Airborne and Monitoring Digital Cameras: Closed-Circuit Television (CCTV) cameras are commonly employed to supervise traffic states and assist law enforcement in large cities. By acquiring monitoring film, traffic data can be evaluated and utilized for forecasting purposes. In [62], the traffic situation of CCTV systems is studied, and the difficulty is the tremendous processing load of digital multimedia data. The authors in [63] examined transportation video data management and developed a high-performance computing architecture using dispersed files and distributed computing systems. As an alternative to fixed traffic monitoring cameras, Remote Sensing (RS) instruments like Unmanned Aerial Vehicles (UAVs) and aerial digital cameras can be used to keep tabs on the flow of vehicles. Average trip time and density are calculated in [64] using UAV aerial footage captured by an optical 3K camera system, and the traffic flow parameter is determined in [65]. The video data has many advantages such as their constant monitoring, broad coverage of road traffic, and many applications for traffic estimation, forecast, accident investigation, and vehicle tracking. However, video data presents some drawbacks, such as the need for extensive storage and processing resources. Compression problems, blurring, and hardware flaws are just a few of the additional engineering challenges presented by the video stream. Examples of publicly available Airborne and Monitoring Digital Cameras datasets are (1) The MIT Traffic Dataset [66] which is a set of data that can be used to analyze crowds and activity. It has a 90-minute video sequence about traffic. It is recorded by a stationary camera. The scene is 720 by 480 pixels. It's made up of 20 clips, and (2) Video Surveillance Data (VSD) [67] which contains 982 video frames with more than 60,000 objects presented in various conditions.

- Electronic Toll Collection Data: Electronic Toll Collection (ETC) systems became commonly used on toll highways and toll bridges, among other applications. The information gathered from ETC systems is a comprehensive and cost-effective tool for calculating expressway traffic. To extract traffic flow from data collected by ETC systems in Shandong, China with a high degree of examination precision, a simulation-dependent dynamic traffic allocation algorithm was proposed in [68]. One of the main benefits of ETC systems is the data they collect, which is both comprehensive and inexpensive to collect. Although, the negatives are obvious. Only in areas with functional ETC systems can toll data be collected. The Amap data for Knowledge Discovery and Data Mining (KDD) CUP 2017 [69] is the only public source of toll ticket data that the authors know of that is used to predict traffic flow. To summarize the above discussion, this section discussed the major BD sources as well as other different sources of data that can be used as input for various BD Processing Frameworks that will be discussed in detail in the next section.

4. PROCESSING FRAMEWORKS OF BIG DATA

As conventional BD tools are not optimal for applying BD in the transportation field, there are already prior attempts of utilizing the most recent BD tools to traffic big data. Some exploratory efforts have been made for enhancing the support of spatiotemporal BD using commercially available software. For example, Hadoop is augmented with the capabilities of spatiotemporal indexing and trajectory analytics in [70] and [71]respectively. MobilityDB, an open-source mobility database depending on PostgreSQL and PostGIS, is proposed in [72] for moving object spatial trajectories. In existing systems, online and offline trajectory analyses are frequently separated. To tackle this deficiency, Dragoon, an efficient Spark-based hybrid framework for online and offline trajectory data analyses, is proposed in [73]. Compared to the state-of-the-art massive Ultraman System for trajectory data management presented in [74], Dragoon reduces storage overhead by up to twofold while retaining offline trajectory query processing. Dragoon also delivers a minimum 40% gain in extensibility over Spark and Flink Streaming and an average performance increase of 100% for online traces data processing. Using ground station data, 2G/3G/4G cellular data, road network, and user information, real-time traffic management, and mobility monitoring system have been proposed in [75] to leverage BD tools such as Spark, HBase, and Hive. In this section, the off-the-shelf BD processing frameworks that are already employed in traffic prediction jobs or might be leveraged further for these purposes are summarized. These frameworks are categorized into two categories: Non-Spatial and Spatial BD processing frameworks, as depicted in figure 2. To provide further insight and understanding, at the end of this section, table I compares the non-spatial BD processing frameworks based on some characteristics such as cluster architecture, processing mode, data processing model, data flow, etc. Table II also provides a summary of the key distinctions and similarities amongst spatial BD processing frameworks based on similarities such as geographic partitioning, geospatial indexing, DataFrame API, in-memory processing, etc.

A. *Non-Spatial BD Processing Frameworks*

In this section, the non-spatial BD processing frameworks will be further categorized into three sub-categories namely: Batch BD Frameworks, Stream BD Frameworks, and Hybrid BD Frameworks.

1) *Batch BD Frameworks*

To be processed in batch mode, BD had to accumulate for a considerable amount of time, often hours or days. Unless the information was imported into memory before processing, it would be stored in a file system or database [29]. This section will explain several common architectures for BD batch processing and provide examples.

- Apache Hadoop; Hadoop is a fantastic BD solution because it is developed in Java and can process both structured and unstructured data from a variety of sources [76]. Apache Hadoop began in April 2006 as an offshoot of the Apache Software Foundation's Lucene project. In Hadoop, data is processed in

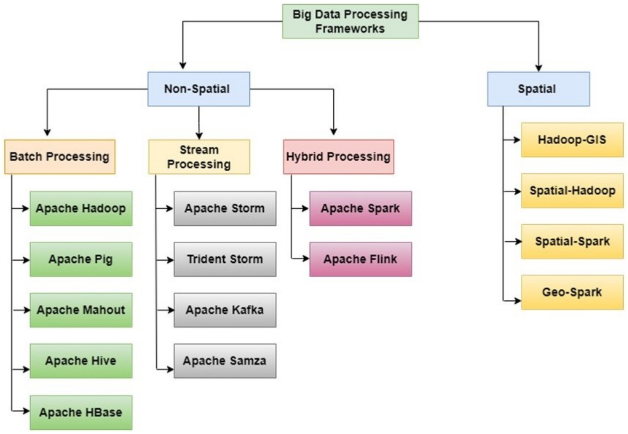


Figure 2. BD Processing Frameworks

parallel using a computer cluster, thereby speeding up big computations and concealing I/O latency by increasing parallelism. Also, it can utilize distributed file system to affordably and reliably duplicate pieces of data into cluster's nodes, which makes data locally accessible on the computing machine. In addition, Hadoop offers application developers an abstraction of the map and reduces operations [77].

- Apache Pig: Apache Pig was created in September 2008 by Yahoo's researchers to run MapReduce operations on massive datasets at a high level of abstraction. It offers the high-level scripting language Pig Latin for coding for data analysis. The Hadoop Distributed File System (HDFS) relies on user-created scripts written in the Pig Latin Language to do data processing. Each script would be converted by the Pig Engine into a map-and-reduce processing job. It's important to note that developers wouldn't have access to these steps otherwise, as the process is totally internal. The output of Pig will be saved to HDFS upon completion. Using a multi-query method, Apache pig reduces development time in comparison to MapReduce. Using Pig Latin, the identical task might be accomplished with significantly less code than using Java. User-Defined Functions (UDFs) written in multiple languages such as Python, Java, JavaScript, Groovy, or Ruby can be used to extend Pig Latin, which can then be called straight from the language [78].

- Apache Mahout: Apache Mahout, which was released for the first time in April 2010 as an open-source framework that offers a service for the adoption of scalable and distributed ML algorithms that emphasize statistics and linear algebra. It's coded using Java and Scala. Mahout functions in conjunction with Hadoop, allowing customers to apply ML algorithms using parallel processing on Hadoop. The fundamental Mahout algorithms involve clustering, classification, recommendation mining, and frequent item-set mining. Even though Mahout is still being developed, several algorithms have been developed [79].

- Apache Hive: Facebook developed the data center-friendly information extraction and analysis tool Apache Hive. It is based on Apache Hadoop and offers users a SQL-like environment for interacting with data stored in Hadoop-based databases and storage systems. To perform SQL applications and queries on distributed data, a Java API is no longer necessary thanks to Hive's SQL-like querying language, HiveQL or HQL. Hive promotes the usage of SQL-based querying languages with Hadoop because these are used by most data warehouse applications [78].

- Apache HBase: Apache HBase is a free and open-source Java-based non-relational database that was originally developed by Google and is based on Bigtable. The original version was released in March of 2008. It's a Bigtable add-on that operates on top of HDFS or Alluxio in the Hadoop ecosystem. It can safely and securely store a massive volume of sparse data. HBase also offers column-level compression, Bloom filters, and in-memory operations. Hadoop MapReduce tasks can use HBase tables as input using Java APIs and similar APIs, and HBase tables can be output via MapReduce operations. HBase's enormous adoption can be attributed, in no little part, to its interoperability with HDFS and Hadoop. HBase is not intended to be used in place of a relational database, but thanks to the Java Database Connectivity (JDBC) driver and SQL layer developed as part of the Apache Phoenix project, it can be used alongside relational databases for business analytics and intelligent applications [80].

2) Stream BD Frameworks

When data streams in micro-batches or in real-time, Stream BD frameworks can manage it [29]. Apache Storm, Trident Storm, Apache Samza, and Apache Kafka are all examples of stream BD frameworks [77].

- Apache Storm: Apache Storm was created by Nathan Marz and the BackType team in September 2011 and released as a distributed BD framework for stream processing. Then Twitter made it open-source, and it became the gold standard for distributed real-time processing systems [81], [82], [83]. Directed Acyclic Graphs (DAGs) form the basis of a typical Apache storm structure, with edges representing data exchange and vertices representing computational resources. The overall architecture is like a MapReduce job, with the key difference being that data is processed in real-time as opposed to in batches. In addition, Storm topologies keep running until they are stopped, or a mistake occurs, whereas a MapReduce task DAG must end [84].

- Trident Storm: The Apache Storm topology makes use of the many high-level operators provided by the Trident APIs [81]. To accomplish more, Trident APIs break down jobs into smaller groups. Performance and latency can be fine-tuned by modifying the batch size. However, DAGs cannot be used to run iterative algorithms because of their topologies [85].

- Apache Kafka: LinkedIn released the Scala and Java open-source project Apache Kafka [77] in January 2011. It's an open-source Peer-To-Peer (P2P) messaging system with a publish-subscribe architecture. Its goal is to establish a reliable, high-throughput, and low-latency infrastructure for the coordination and dissemination of massive amounts of time-sensitive information. It is also an optimized Transmission Control Protocol (TCP) based binary protocol for efficiently grouping messages to reduce network roundtrip overhead. Kafka is developed based on the ZooKeeper synchronization service.

- Apache Samza: Apache Samza is introduced by LinkedIn and built based on two other frameworks: Kafka and Hadoop [77], [86]. Apache Samza was developed to tackle some issues in stream processing like scalability, resource allocation, etc. [86].

3) Hybrid BD Frameworks

Some applications need frameworks for both BD processing modes: batch and stream processing. In such circumstances, hybrid processing frameworks are required. Apache Spark and Apache Flink are among the most prominent examples.

- Apache Spark: Hadoop is the foundation upon which Apache Spark, a hybrid processing platform, is constructed. It enhances batch processing workloads with in-memory processing, which improves BD processing overall [77]. Data loading for processing into memory and result storage were the only two use cases where it was allowed to build linkages between data stores. Spark, unlike Apache MapReduce, keeps the results it obtains in memory rather than persistent storage. The core data model of Apache Spark is called Resilient Distributed Datasets (RDDs), and it lets programmers gather and reuse intermediate data. Due to their high fault tolerance and ability to manage partitions with minimal data loss, RDDs are increasingly used in modern data management [87]. Many high-level libraries, such as MLlib for ML [88], GraphX for stream processing [89], and Spark SQL for structured data processing [90], complement the core components of the Apache Spark framework [87]. Spark's multi-cluster backbone is implemented in Scala. Spark runs various data visualization and analysis techniques and supports high-level APIs including Scala, Java, Python, and R [14].

- Apache Flink: The Apache Flink [91] is a free and open-source framework for running iterative algorithms, batch processing, and data-stream pipelines in real-time. The primary advantage of a distributed system is that it can process massive data sets with a low cost of delay and high fault tolerance. Processing in batches of data that are separated by some threshold requires the use of the Dataset API [91].

4) Spatial BD Processing Frameworks

In this section, four spatial BD processing frameworks will be discussed, namely: Hadoop-GIS, Spatial-Hadoop, Spatial-Spark, and Geo-Spark.

- Hadoop-GIS: Hadoop-GIS is a MapReduce-based framework for handling geographic queries, vector data segmentation, and massive data sets [92]. In addition to geographical mining and statistical methods, geospatial queries can be descriptive, spatial, and distance based. Hadoop-GIS makes use of Sample-Analyze-Tear-Optimize (SATO) geographic partitioning and local geospatial indexing to speed up query times [93]. Convex polygons, line strings, multiple points, and multiple polygons are not supported. Hadoop-GIS favors 2D data and two spatial query techniques beyond geometric objects: box range and spatial joins [14].

- Spatial-Hadoop: To get over the constraints of Hadoop-GIS, a comprehensive MapReduce framework called Spatial-Hadoop was developed. It has two cutting-edge features that aid in geospatial data storage, indexing, and operations: Reader for Spatial Records and Splitter for Spatial Files [94]. The geometry types that can be processed by Spatial-Hadoop are not limited to just points and lines but also include polygons and other multi-point geometries. Spatial partitioning algorithms including R-Tree, K-Dimensional Tree (KD-Tree), Quad-Tree, uniform grids, and Hilbert curves have all been used to organize geographic indexes [14].

- Spatial-Spark: Spatial Spark is a platform for cluster-based Geospatial data processing. It was built upon Spark RDD to provide a wide variety of spatial operations, such as spatial filtering, spatial join, R-Tree index, range query, and R-Tree splitting to improve query performance [95]. Spatial-Spark is an in-memory BD framework designed to enable two spatial join operations, namely partitioned spatial join and broadcast spatial join [95].

- Geo-Spark: A Spark-based in-memory cluster computing platform, Geo-Spark [96] is capable of processing large-scale geographic data at a much faster rate than Spatial-Hadoop. To accommodate geographical data types, indexes, and massive geometric computations, Geo-Spark expands on the ideas of RDDs and SparkSQL. Developers working with Apache

Spark can take advantage of operational quickly (like Java and Scala), declarative (like SQL), and spatial RDD APIs to build powerful apps for geographic analysis. It also makes it easier to use techniques like R-tree, KD-Tree, Quad-tree, and KNN searches for dividing geographical data. To strike a good middle ground between processing speed and memory/processor load, Geo-Spark is optimized to determine the best possible connect method [97].

5. RESEARCH METHODOLOGY

The publications reviewed in this work were published in high-quality Elsevier, IEEE, IOP, and Springer conferences and journals. The following search phrases were used to locate these articles: Big Data Processing Frameworks, Big data in traffic flow prediction, and GIS in traffic flow prediction. The reviewed articles are directly related to the use of Big Data techniques and GIS in traffic flow prediction. This work considered both empirical and scholarly reviews on the mentioned topics.

A. Related Work

This survey investigates various forecasting techniques for traffic flow prediction using BD techniques and GIS. It provides a detailed discussion of the approaches and algorithms which are utilized for predictions, performance measurements, and tools used for these procedures. The main objective of this survey is to investigate how GIS, BD sources, and BD processing frameworks can be integrated to enhance traffic prediction performance and accuracy, hence presenting an easy-to-adopt unified conceptual framework to be a guide for researchers on how to adopt BD and GIS in traffic prediction. In [98], the authors proposed a traffic flow prediction system in real-time with a simple method and vivid visualization. Many highways were analyzed simultaneously using the suggested system, and the results of the predictions are shown on a heat map. In the first step, the authors constructed a Prophet model to anticipate traffic volumes in subsequent time steps. Sensors placed throughout the roadways collect training data, which may include the average speed of vehicles, the time of day, the placement of sensors, and so on. All the system's data come from a trustworthy source that has been made public using IEEE DataPort. The obtained results showed that predictions of future speeds are less accurate than those made at earlier times, and this trend continues as the time horizon expands. Also, the Prophet model's performance is not very great, although it works well for trend time series prediction with long-term knowledge. Meanwhile, the traffic flow is quite sensitive to recent data. In [20], the authors proposed a model that combines IoT and BD to apply and analyze its social advantages in intelligent traffic flow forecasting, and it also analyses the model's three-tier network architecture, consisting of a perception layer, a network layer, and an application layer. Also, it examined and evaluated the most effective means of bringing together cloud and edge computing. By combining KF and BP techniques, the method suggested in this paper can increase prediction accuracy while decreasing the error value for the entire traffic flow. In this study, the authors highlighted the challenges associated with fusing multivariate data for predictive purposes and argue that more advanced optimization methods are required. The aim of the work presented in [99] was to shed light on the effects of mobility choices on urban air quality and to address the issue of traffic monitoring and analysis across location and time. In this work, the authors presented a revolutionary traffic data analytics platform, complete with capabilities for real-time traffic monitoring, effective traffic analytics over time, and the ability to comprehend how adjustments to the urban vehicle fleet can lessen the influence of traffic on air pollution. By providing a centralized, user-friendly interface, the authors of the proposed Trafair Traffic Dashboard (TTD) aimed to facilitate research into both current and past traffic statistics, as well as provide a visual representation of the environmental toll that traffic congestion has on a city. The dashboard was supported by several data analysis techniques, such as visualization of sensor observations, anomaly detection, and traffic simulation analysis, enabling the investigation of behavioral similarities between sensors or neighborhoods, the visual detection of unusual events, and the simulation of traffic flow on a new hypothetical vehicle fleet scenario defined by the user. With two examples based on actual traffic data, the suggested dashboard was shown to be useful in easing the decision-making process. In [100], The authors aimed to propose an intelligent real-time traffic model to address the traffic congestion problem. The proposed model assists the urban population in their everyday lives by assessing the probability of road accidents and accurate traffic information prediction. It also helps in reducing overall carbon dioxide emissions in the environment and assists the urban population in their everyday lives by increasing overall transportation quality. This study offered a real-time traffic model based on the analysis of numerous sensor data. The proposed model incorporated data from road sensors as well as a variety of other sources. Sensor data is consumed by streaming analytics platforms that use big data technologies, which are then processed using a range of deep learning and machine learning techniques. The study provided in this paper would fill a gap in the data analytics sector by delivering a more accurate and trustworthy model that uses internet of things sensor data and other data sources. This method can also assist organizations such as transit agencies and public safety departments in making strategic decisions by incorporating it into their platforms. The model has a big flaw in that it makes predictions for the period following January 2020 that are not particularly accurate due to the Covid-19 pandemic has impacted the traffic scenario, resulting in erratic data for the period after February 2020. Using Big Data Analysis (BDA) on the copious amounts of information produced by the smart city IoT, the authors of [101] hoped to push the smart city in the direction of better management and more secure data storage. This study aimed to apply DL algorithms and BDA to smart city data, and it proposed a distributed parallelism technique for

TABLE I. A comparison between popular Non-Spatial Big Data Frameworks

Framework	Hadoop	Storm	Trident Storm	Samza	Spark	Flink
Architecture	YARN	Nimbus	Nimbus	YARN and Kafka	YARN and Mesos	YARN and Kafka
Mode	Batch	Stream	Stream	Stream	Hybrid	Hybrid
Model	MapReduce	At-Least-Once	Exactly-Once	At-Least-Once	Exactly-Once	
Flow	MapReduce	Cyclic Graph	DAG	Kafka-Kafka job-Kafka	RDDs processed individually.	Stream-System-sinks
Latency	Low	MS	MS/Small Batches	MS	High	Low
Scalability	Yes	Yes	Parallel Processing	Yes	On Demand	Parallel Tasks
ML	Yes	Compatible with SAMOA API	Trident-ML	Compatible with SAMOA API	Spark MLlib	FlinkML
Languages	Java	Java, Ruby, and Perl	Java, Ruby, and Perl	Java	Scala, Java, Python, and R	Scala and Java

TABLE II. A comparison between popular Spatial Big Data frameworks

Feature	HadoopGIS	SpatialHadoop	SpatialSpark	GeoSpark
In-memory processing	No	No	Yes	Yes
DataFrame APIs	No	No	No	Yes
Geo-Indexing	R-Tree	R-Tree /QuadTree	R-Tree	R-Tree /QuadTree
Geo-Partitioning	SATO	Multiple	Multiple	Multiple
Query Optimizer	No	No	No	Yes
KNN query	Yes	Yes	No	Yes
Distance Query	Yes	Yes	Yes	Yes
Distance join	Yes	Yes	Yes	Yes
Filter (Contains)	Yes	Yes	Yes	Yes
Filter (Intersects)	Yes	Yes	Yes	Yes
Filter (ContainedBy)	Yes	Yes	Yes	No
Filter (WithinDistance)	Yes	Yes	Yes	No

Convolutional Neural Networks (CNNs). Meanwhile, the concept of Digital Twins (DTs) was presented alongside multi-hop transmission technology to build a DL-based IoT-BDA system for a smart city, with the ability to simulate and analyze the system's performance beforehand. It is shown that the proposed model's prediction accuracy is as high as 97.8%, based on an analysis of the data. In [7], two case studies were used to analyze and evaluate supervised ML techniques (Decision Tree, KNN, M5P, RF, Random Committee, and Random Tree algorithms) as approaches to BD analytics for traffic volume forecasting. Both sets of experiments used traffic data collected by selected automatic traffic counters installed along roads in the Republic of Serbia from 2011 to 2018 for training and testing prediction models. The M5P algorithm performs well in one scenario, whereas the KNN approach performs best in another. The study in [102] used a DL model to create a data fusion-based traffic jam control system for smart cities. In smart cities, traffic flow predictions are made at the regional level using a hybrid model built upon the CNN and Long Short-Term Memory (LSTM) architectures. In this context, CNN is employed for temporal data classification and LSTM for spatial data classification. Root Mean Square Error (RMSE), the time required, and accuracy were all measured during the tests that made use of the CityPulse Traffic and CityPulse Pollution datasets. When compared to other baseline models, the suggested model's low RMSE value of 49 and greatest accuracy of 92.3% illustrated its viability in the region-based traffic flow prediction challenges in smart cities. In [103], There were two goals for this paper. First, it provided a comprehensive literature assessment on big data analytics for ITSs. Second, it relied upon prior research to suggest a basic but comprehensive framework for designing an architecture to handle BD analytics in

ITSs. The results of this work will be used to examine transportation data for a major Colombian metropolis. For traffic flow prediction, the authors of [104] proposed a CNN-LTSM model trained using incremental learning (IL-TFNet). The forecasting performance and efficiency of the model have been optimized using a lightweight CNN-based model architecture. This architecture is built to handle spatiotemporal and external environment variables simultaneously. They used the K-means clustering approach with an uncertainty feature to derive previously unknown data about traffic accidents. To meet the needs of high real-time performance and low computational overhead in short-term traffic prediction, an incremental learning method is utilized during model training instead of the conventional batch learning algorithm. In addition, the authors presented a method for improving the prediction model's precision in exceptional cases by combining incremental and active learning. Multi-Layer Hybrid Network (MLHN) was created by the authors of [105] to analyze and anticipate network traffic; MLHN is made up of three separate networks to process various inputs for individualized feature extraction. For the first network, the input is the sequence of network traffic within a specified time interval. To comprehend and leverage the deep features, the third input takes the cross-domain parameter, and the second network takes input as references and data corresponding to traffic dates and times. To further improve parameter learning, which typically results in less inaccuracy, a new, improved, and efficient parameter-tuning algorithm is presented. The suggested MLHN considered call detail records from Milan, Italy, by evaluating 100*1000 grids of size 235*235 square meters from the Telecom Italia BD challenge dataset [106]. The objective of [107] was to provide an officer with a visual representation of the prevailing traffic trend and a time-dependent prediction of traffic flow. This information includes traffic volumes and timestamps. Time-series data analysis is performed using a deep learning method called an LSTM Recurrent Network (RN). Python would be used for both the data preprocessing and model training. Tableau Prep Builder is used to clean and arrange the data before it is published to Tableau Server using Python. Tableau would be used to create an online interactive dashboard for internal use. Output predictions for the next 15 minutes can be made with an accuracy of up to 83% using 3 hours of input data. A model to forecast when and where traffic jams may form was suggested in [108]. This research used Big Data, such as POIs and check-in, to analyze the features of residents' spatial activities based on urban planning data and inhabitants' distribution data of Beijing's Second Ring Road. A more accurate reflection of the city's traffic congestion at specific times and roadways can be achieved by the simulation of urban traffic operating circumstances, which is based on the actions of the city's population. Authors rasterized all data through a GIS system, and each grid data individually calculates its connection with the surrounding grid, allowing for a more convenient quantitative examination of citizens' activities in cities. According to the data collected throughout the experiments, the model that considers the spatial activities of locals is both efficient and accurate in making predictions. The findings were consistent with reality and have significant practical importance. Using both little data from conventional transport surveys and BD from ICT, the authors of [109] investigated how to build Transport System Models (TSMs). The ex-post effects of planned actions and policies cannot be assessed using BD, but it may be used to examine historical mobility patterns and transport infrastructure and services. The research proposed a strategy for optimizing the benefits of TSMs' in-built forecasting capabilities, on the one hand, while minimizing the costs of traditional surveys, on the other, by making use of big data, on the other. Free-Circulating Data (FCD) was added to incomplete databases to fill in gaps, such as the census. The authors provided an actual, extra-urban environment where data fusion was accomplished using a GIS tool to demonstrate the viability of their proposed approach. In [110], a new and all-encompassing method was proposed for predicting traffic on a global scale faster and more accurately, and in real-time. It brought together four cutting-edge technologies that are different but work well together: BD, DL, in-memory computing, and Graphical Processing Unit (GPU). The deep networks were trained with information from the California Department of Transportation (Caltrans). Aside from using a small amount of data, the proposed method has a low level of accuracy when it comes to making predictions. In [111], the authors aimed to use BD to look at how traffic flows in a city. In particular, the authors came up with a DL model for predicting traffic flow by using data visualization techniques for data mining. Specifically, the first set of data was collected and made by Vehicles Detection Systems (VDS) in Daejeon, South Korea. Then, an LSTM-RN has been set up to predict traffic flow. They set the size of the window to be 24 hours, with 2 hours of the prediction. The results of the experiments were promising, so it's worth continuing to work on them. With the ability to capture the temporal correlation and periodicity of traffic flow data and the disturbance of weather factors, a combined framework of Stacked Auto-Encoder (SAE) and Radial Basis Function (RBF) neural network is proposed in [112] to predict traffic flow. Before modeling for traffic prediction, many works for data processing were involved in their experiment. The authors first employed one-hot coding for the original expression of the non-numerical weather type parameter. Then, an embedding component is used to acquire the interpretable expression. The Pearson Correlation Coefficient (PCC) is used to identify the flow-related parameters when dealing with many weather parameters, and PCA is then used to process the selected parameters into a new parameter with increased correlation. Additionally, based on past traffic data, Historical Average (HA) is used to build a time expression for use in the prediction process. Specifically, SAE was used to learn the temporal correlation in traffic flow, RBF was used to learn the regular progression under weather disturbance, and another RBF was used to realize the decision-level data fusion of the former models. This integrated framework

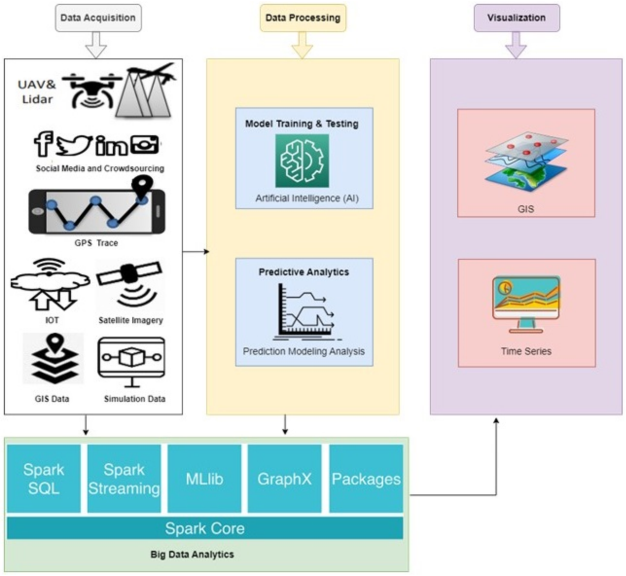


Figure 3. The Proposed Conceptual Framework for Traffic Prediction

can efficiently learn the temporal correlation and periodicity of traffic flow and weather disturbance to strengthen the prediction model's accuracy and robustness. The authors tested their proposed framework, and the results were 10.37 %, 10.06 %, 9.5, 151.15, and 12.29 respectively for Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Error (MAE), Mean Square Error (MSE), and RMSE.

6. A CONCEPTUAL FRAMEWORK FOR TRAFFIC PREDICTION

In this section, a conceptual framework for employing BD analytics and GIS in traffic prediction is presented. This framework is still under construction and has not been implemented yet. The motivations behind it are: 1) To help us better understand real-world systems in traffic prediction, facilitate communication, and integrate knowledge across disciplines and 2) to present an easy-to-adopt unified framework that can be employed by other researchers who are interested in traffic prediction problems. To the best of our knowledge, it is the first time that a conceptual framework for traffic prediction is to be proposed in a review study. As depicted in figure 3, the conceptual framework compromises four steps: Data Acquisition, Data Processing, BD Analytics, and Visualization. The next subsections discuss these steps in detail.

- Data Acquisition: The first and most necessary stage of BD applications in traffic prediction is the collecting of traffic flow data from different sources, such as traffic sensor data, GPS traces, LBS data, satellite imagery, social media, Crowdsourcing, and IoT. The fundamental modules implemented at this step include data filtration, data harmonization, and unnecessary data elimination. Additionally, for each dataset, metadata are produced to describe how the data are rendered and processed.

- Data Processing: It is necessary to process the acquired data by extracting the essential information for future investigation. In this phase, data are cleaned, interpreted, integrated, mined, analyzed, and warehoused to obtain important information. To enhance data comprehension, many analyses, including visual, prescriptive, diagnostic, and predictive, could be applied. In general, visual analysis has the potential to gain insights into the discovered links within enormous datasets and to equip researchers with more intuitive visual perception and effective decision-making support. GIS is the primary tool for achieving visualization in the proposed framework. In this context, GIS technology needs to be first equipped with the appropriate BD. No longer is data acquisition limited to classical facilities and procedures, including satellite RS, and stations, in addition to field measures. The acquired data can be converted into a format compatible with GIS tools. To successfully integrate data from a variety of sources, decision-makers used spatial tools like GoogleEngine, which renders digital interactive maps in near real-time using a significant amount of free satellite data and provides a stable working environment. Prescriptive analytics evaluate analysts' ability to determine ideal actions and conclusions based on responses to a variety of ”what may happen?” inquiries. For instance, analysts may have multiple options for assigning maintenance actions to a particular asset. Prescriptive analytics synthesize BD, varied sciences' fundamentals, business rules, and IoT disciplines to determine the merits of forecasts and then make the most optimal selections. Forecasts are surpassed by prescriptive analytics. The ”why will it happen?” questions need to be supported by the ”what will happen?” and ”when will it happen?” questions. An intelligent, professional dashboard integrated with time-series analysis provides analysts with the essential tools for swiftly summarizing an overview that aligns with company goals. In this regard, diagnostic analytics encourage analysts to conduct the main cause analysis to identify the primary causes of events. The diagnostic analytics address the inquiry ”why did it occur?”. Many Artificial Intelligence (AI) techniques such as data mining and correlation analysis can provide profound insight into determining the targeted challenges and issues. In addition, DL represents a complete shift in the orientation of supervised ML, such as Natural Language Processing (NLP) and pattern recognition. Predictive analytics aims to anticipate the future by answering ”what will happen?” queries. Numerous statistical and ML techniques try to correlate historical and current data to predict the future.

- Big Data Analytics: In this stage, the data acquired from different sensors in addition to the models trained and tested could be incorporated to improve management, monitoring, and anticipation of traffic flow. BD analytics frameworks can be utilized to speed up predictive insights in real-time. One of the most powerful BD analytics frameworks that can be utilized in this stage is Apache Spark, which was discussed earlier in this paper. Apache Spark includes several upper-level libraries for ML, stream processing, and structured data processing which make traffic prediction more intelligent and effective.

- Visualization: In the final stage, the obtained data are then transformed into actionable perceptions. Visualization tools such as GIS and TSA graphs are being used in this stage to display the prediction results efficiently. This stage is so important as it enables decision-makers and experts in the transportation field to take the appropriate decision to mitigate a specific traffic situation.

7. CHALLENGES AND OPPORTUNITIES

This section first discusses the results of the analysis of the state of the art and presents the strengths and the weaknesses of existing works, as shown in Table III. It then discusses the open issues and challenges that face big data in traffic prediction. Despite the many successful application cases for BD in traffic prediction tasks, there still exist some challenges. Data density varies significantly between modes of transportation, and the challenges of data scarcity, excessive missing data, data distortion, and deficiency persist. Previous research has not adequately explored data quality, privacy, and policies. This section outlines some future directions for addressing these difficulties. For instance, crowdsourced data suffer from poor data quality, complexity in noise removal, and privacy issues. To address these obstacles, some attempts have been made, such as the use of sparse BA for traffic conditions prediction with undersampled data [113]. The geographical and temporal ranges of available data for traffic estimating and prediction activities are varied. In the transportation domain, combining data from multiple sources across multiple scales remains a challenge. Another obstacle is the scarcity of ”true” BD in the transport industry, particularly open ones. Although numerous data sources were investigated in this research, some of them contain data volumes that are difficult to classify as ”large”. This difficulty is exacerbated by the costly and time-consuming nature of obtaining certain types of data. The other reason prohibiting the acquisition of fine-grained data is the fear of location privacy leakage.

1) *Multi-Sources Data Incorporation*

As multiple modes of transportation interact with one another, it may be difficult to rely on a single data source to manage traffic flow effectively. Alternatively, it is possible to enhance traffic evaluation and predictions by combining data from multiple sources. This finding motivates the use of DL for urban BD fusion, which in turn promotes the integration of data from many sources [114]. There are already some relevant studies that discussed the relevant sources of data. For example, cell phone data and loop sensors are used together to estimate the speed of traffic on a freeway [115]. In [116], GPS trace data and data from a few stationary traffic sensors are used to estimate the flow of traffic on the whole road network. Data from geomagnetic detectors, data from floating cars, and data from reading license plates are all put together [117] to get the average link travel time. Using data from the General Transit Feed Specification (GTFS) on bus schedules, real-time data on bus positions, and cell phone data from geographical mapping software, it is possible to predict bus delays with a MAPE of about 6% [118].

2) *Hybrid Learning and Computing Techniques*

Because there are so many sources to get data, numerous computing methods have been used to collect and process various kinds of data, including mobile computing, cloud computing, fog computing, edge computing, etc. It is still challenging to connect and use different computing techniques with existing BD infrastructures due to differences in their communication and computing capabilities. Numerous ongoing studies were investigating this topic. For instance, in [119] a low-cost traffic monitoring system was developed using IoT technologies and fog computing to process the gathered data and extract traffic details from car GPS traces. Supervised ML formulations are the standard when addressing traffic prediction concerns. Despite this, alternative forms of learning, such as Transfer Learning (TL) and Generative Adversarial Learning (GAL), can be used to address these issues. To tackle the issue of data scarcity in various locations, TL has great potential. For combining human movement data with urban POI data, the authors of [120] suggested employing an embedding technique for POI. CNN and an LSTM record both spatiotemporal and spatial information, and mobility data is transferred between cities. This improves the accuracy of predictions for the target city with limited information. In addition, a traffic Generative Adversarial Network (GAN) technique was presented in [121] to tackle the unexplored problem of off-deployment traffic prediction. TrafficGAN can account for changes in travel demand and basic road network parameters, allowing for a more accurate description of traffic patterns.

3) *Multi-Sources Data Incorporation*

In recent years, several initiatives have been made to develop and employ distributed systems and frameworks. In addition to greater use of the existing frameworks, various emerging technologies may be applied to traffic prediction as well as other pertinent challenges. In [122], for instance, blockchain is utilized as a viable option for BD exchange trust and privacy security. In [123], federated learning is also used to forecast traffic flow while maintaining anonymity.

TABLE III. A comparison between related works

Study	Approaches	Findings
[98]	Prophet model	Low Accuracy, Poor Performance.
[20]	BD, IoT, KF, and BP	High Accuracy, Low error rate
[99]	Anomaly detection,traffic simulation.	Useful for decision-making.
[100]	BD, DL, and ML.	Low Accuracy.
[101]	DL, BDA, CNN.	Accuracy: 97.8%.
[7]	DT, KNN, RF, Random Committee, RT	High Accuracy in M5P and KNN.
[102]	CNN	RMSE: 49, Accuracy: 92.3%
[103]	BDA	Will be used to examine transportation data for Colombian metropolis.
[104]	CNN, K-means clustering, and LTSM.	High real-time performance, Low computational overhead.
[105]	MLHN	Efficient parameter-tuning algorithm.
[107]	LSTM RN.	Accuracy: 83%
[108]	BD and GIS.	Efficient and accurate predictions.
[109]	BD and GIS.	Optimized forecasting capabilities, Minimized costs of traditional surveys.
[110]	BD, DL, in-memory computing, and GPUs.	Small amount of data used, Low prediction accuracy.
[111]	LSTM RN.	Promising prediction accuracy.
[112]	SAE, RBF, PCC, PCA, and HA.	MAPE: 10.37%. SMAPE: 10.06 % MAE: 9.5, MSE: 151.15. RMSE:12.29.

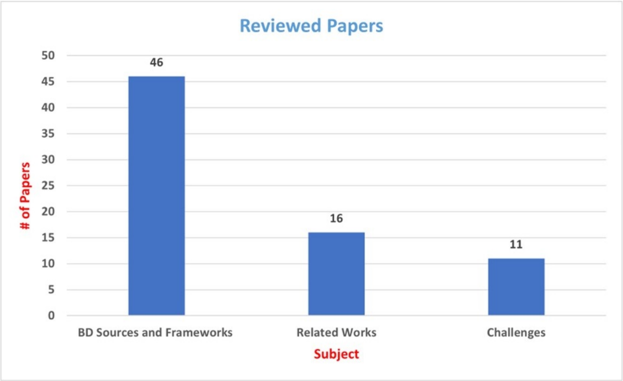


Figure 4. Categorization of reviewed papers

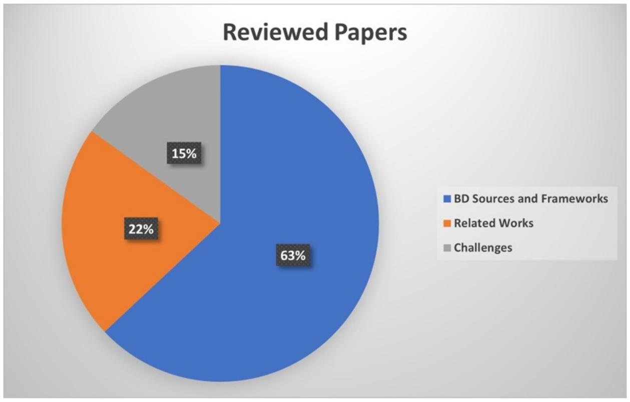


Figure 5. Percentage of each category of papers

8. CONCLUSION

This study conducted a systematic review to inspect the recent cutting-edge research of BD and GIS in traffic prediction. It concentrated on the processes of traffic prediction and how BD and GIS can be appropriate and powerful choices. It presented an investigation of the different sources of BD that can be utilized in traffic prediction tasks. In this regard, this paper reviews 73 peer-reviewed articles on BD and GIS in traffic prediction, indicating BD's prominent role in tackling the challenges of traffic prediction tasks. As depicted in figure 4 and figure 5, 46 paper discussed BD sources and Frameworks, which forms about 63% of the reviewed papers, 16 paper discussed related works about how BD and GIS can be utilized in traffic prediction tasks, which forms about 22% of the reviewed papers, and finally, 11 paper discussed the challenges and future opportunities for employing BD analytics in traffic prediction, which forms about 15% of the reviewed papers. Finally, the present study proposed a conceptual framework for employing BD analytics and GIS in traffic prediction.

From the above discussion, ample conclusions were drawn:

- An abundance of new tools for data collection, processing, storage, and usage has emerged in tandem with the rise of BD, which can be beneficial for work in traffic prediction.
- Despite there being many sources of data, the richness of the data varies a lot, and the volumes of data aren't big enough because they are limited in space and time.
- BD storage and processing capabilities can facilitate the integration of more powerful techniques such as DL, which can make traffic prediction more accurate and intelligent.

References

[1] H. Yuan and G. Li, "A survey of traffic prediction: from spatio-temporal data to intelligent transportation," *Data Science and*

Engineering, vol. 6, pp. 63–85, 2021.

[2] Y. Zhang, Y. Zhou, H. Lu, and H. Fujita, "Traffic network flow prediction using parallel training for deep convolutional neural networks on spark cloud," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7369–7380, 2020.

[3] W. Jiang and J. Luo, "Big data for traffic estimation and prediction: a survey of data and tools," *Applied System Innovation*, vol. 5, no. 1, p. 23, 2022.

[4] K.-H. N. Bui, J. Cho, and H. Yi, "Spatial-temporal graph neural network for traffic forecasting: An overview and open research issues," *Applied Intelligence*, vol. 52, no. 3, pp. 2763–2774, 2022.

[5] G. Lan, L. Yaojun, and C. Xinfa, "Study and development of webgis-t for transportation," *Journal of Changsha Communications University*, vol. 17, no. 4, pp. 18–22, 2001.

[6] N. T. Gill and A. Al-Akhras, "Transportation plan information management system," in *Transportation, Land Use, and Air Quality: Making the Connection*. ASCE, pp. 606–613.

[7] S. Janković, A. Uzelac, S. Zdravković, D. Mladenović, S. Mladenović, and I. Andrijanić, "Traffic volumes prediction using big data analytics methods." *International Journal for Traffic & Transport Engineering*, vol. 11, no. 2, 2021.

[8] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization'97 (Cat. No. 97CB36155)*. IEEE, 1997, pp. 235–244.

[9] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," *Computer science review*, vol. 17, pp. 70–81, 2015.

[10] R. L. Villars, C. W. Olofson, and M. Eastwood, "Big data: What it is and why you should care," *White paper, IDC*, vol. 14, pp. 1–14, 2011.

[11] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information systems*, vol. 47, pp. 98–115, 2015.

[12] M. Arslan, A.-M. Roxin, C. Cruz, and D. Ginhac, "A review on applications of big data for disaster management," in *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2017, pp. 370–375.

[13] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information sciences*, vol. 275, pp. 314–347, 2014.

[14] A. Sayed, A. S. Mahmoud, E. Farg, A. M. Mohamed, M. S. Moustafa, M. A. AbdelRahman, H. M. AbdelSalam, and S. M. Arafat, "A conceptual framework for using big data in egyptian agriculture," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, 2022.

[15] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[16] P. Zikopoulos and C. Eaton, *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[17] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 international conference on collaboration technologies and systems (CTS)*. IEEE, 2013, pp. 42–47.

[18] W. Yang, X. Liu, L. Zhang, and L. T. Yang, "Big data real-time processing based on storm," in *2013 12th IEEE international conference on trust, security and privacy in computing and communications*. IEEE, 2013, pp. 1784–1787.

[19] T. Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review," *Journal of Big Data*, vol. 6, no. 1, p. 47, 2019.

[20] B. Liu, T. Zhang, and W. Hu, "Intelligent traffic flow prediction and analysis based on internet of things and big data," *Computational intelligence and neuroscience*, vol. 2022, 2022.

[21] S.-h. An, B.-H. Lee, and D.-R. Shin, "A survey of intelligent transportation systems," in *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*. IEEE, 2011, pp. 332–337.

[22] Y. Jia, J. Wu, and M. Xu, "Traffic flow prediction with rainfall impact using a deep learning method," *Journal of advanced transportation*, vol. 2017, 2017.

[23] R. Ravish and S. R. Swamy, "Intelligent traffic management: A review of challenges, solutions, and future perspectives," *Transport and Telecommunication Journal*, vol. 22, no. 2, pp. 163–182, 2021.

[24] D. Wilkie, J. Sewall, and M. C. Lin, "Transforming gis data into functional road models for large-scale traffic simulation," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 6, pp. 890–901, 2011.

[25] A. Dabhade, K. Kale, and Y. Gedam, "Network analysis for finding shortest path in hospital information system," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 7, pp. 618–623, 2015.

[26] M. Loidl, G. Wallentin, R. Cyganski, A. Graser, J. Scholz, and E. Haslauer, "Gis and transport modeling—strengthening the spatial perspective," *ISPRS International Journal of Geo-Information*, vol. 5, no. 6, p. 84, 2016.

[27] P. Gupta, N. Jain, P. Sikdar, and K. Kumar, "Geographical information system in transportation planning," in *Map Asia Conference*. Citeseer, 2003.

[28] B. Huang and X. Pan, "Gis coupled with traffic simulation and optimization for incident response," *Computers, environment and urban systems*, vol. 31, no. 2, pp. 116–132, 2007.

[29] S. P. Cumbane and G. Gidófalvi, "Review of big data and processing frameworks for disaster response applications," *ISPRS International Journal of Geo-Information*, vol. 8, no. 9, p. 387, 2019.

[30] U. E. Explorer. Science for a changing world. [Online]. Available: (https://earthexplorer.usgs.gov/)

[31] C. O. A. Hub. Copernicus open access hub. [Online]. Available: (https://scihub.copernicus.eu/dhus/#/home)

[32] M. Ben-Daya, E. Hassini, and Z. Bahroun, "Internet of things and supply chain management: a literature review," *International*

Journal of Production Research, vol. 57, no. 15-16, pp. 4719–4742, 2019.

[33] M. A.-A. Iman Almomani, Bassam Al-Kasasbeh. Wsn-ds: A dataset for intrusion detection systems in wireless sensor networks. [Online]. Available: (https://www.kaggle.com/datasets/bassamkasasbeh1/wsnds)

[34] M. Azkune. Wsn indfeat dataset. [Online]. Available: (https://github.com/apanouso/wsn-indfeat-dataset)

[35] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal*, vol. 78, pp. 319–338, 2013.

[36] E. Loukis and Y. Charalabidis, "Active and passive crowdsourcing in government," *Policy practice and digital science: Integrating complex systems, social simulation and public administration in policy research*, pp. 261–289, 2015.

[37] H. Qin, R. M. Rice, S. Fuhrmann, M. T. Rice, K. M. Curtin, and E. Ong, "Geocrowdsourcing and accessibility for dynamic environments," *GeoJournal*, vol. 81, pp. 699–716, 2016.

[38] M. Büscher, M. Liegl, and V. Thomas, "Collective intelligence in crises," *Social collective intelligence: Combining the powers of humans and machines to build a smarter society*, pp. 243–265, 2014.

[39] T. Balan, C. Dumitru, G. Dudnik, E. Alessi, S. Lesecq, M. Correvon, F. Passaniti, and A. Licciardello, "Smart multi-sensor platform for analytics and social decision support in agriculture," *Sensors*, vol. 20, no. 15, p. 4127, 2020.

[40] crowdflower. 2015 new year's resolutions. [Online]. Available: (https://data.world/crowdflower/2015-new-years-resolutions)

[41] CrowdFlower. Academy awards demographics. [Online]. Available: (https://data.world/crowdflower/academy-awards-demographics)

[42] A. Halper. Social influence on shopping. [Online]. Available: (https://data.world/ahalps/social-influence-on-shopping)

[43] CrowdFlower. Airline twitter sentiment. [Online]. Available: (https://data.world/crowdflower/airline-twitter-sentiment)

[44] T. C. M. A. for Planning (CMAP). My daily travel survey. [Online]. Available: (https://www.cmap.illinois.gov/data/transportation/travel-survey#My_Daily_Travel_Survey)

[45] T. S. D. Center. 2010–2012 california household travel survey. [Online]. Available: (https://www.nrel.gov/transportation/secure-transportation-data/tsdc-california-travel-survey.html)

[46] M. Katranji, S. Kraiem, L. Moalic, G. Sanmarty, G. Khodabandelou, A. Caminada, and F. Hadj Selem, "Deep multi-task learning for individuals origin–destination matrices estimation from census data," *Data Mining and Knowledge Discovery*, vol. 34, pp. 201–230, 2020.

[47] T. A. C. S. (ACS). Us census bureau's commuting flows. [Online]. Available: (https://www.census.gov/topics/employment/commuting/guidance/flows.html)

[48] S. T. S. M. Data. U.s. population migration data. [Online]. Available: (https://www.irs.gov/statistics/soi-tax-stats-migration-data)

[49] Caltrans. Caltrans performance measurement system (pems). [Online]. Available: (http://pems.dot.ca.gov/)

[50] M. C. Council. Open data portal of the madrid city council. [Online]. Available: (http://datos.madrid.es)

[51] OpenCelliD. The world's largest open database of cell towers. [Online]. Available: (https://www.opencellid.org)

[52] T. Italia. Telecommunications - sms, call, internet - mi. [Online]. Available: (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV)

[53] M. R. Asia. Geolife gps trajectories. [Online]. Available: (https://www.microsoft.com/en-us/download/details.aspx?id=52367)

[54] Idiap and NRC-Lausanne. Lausanne data collection campaign (ldcc). [Online]. Available: (https://www.idiap.ch/en/dataset/mdc)

[55] B. Liao, J. Zhang, C. Wu, D. McIlwraith, T. Chen, S. Yang, Y. Guo, and F. Wu. Deep sequence learning with auxiliary information for traffic prediction. [Online]. Available: (https://github.com/JingqingZ/BaiduTraffic)

[56] D. YANG. Foursquare dataset. [Online]. Available: (https://sites.google.com/site/yangdingqi/home/foursquare-dataset)

[57] E. Hussain, A. Bhaskar, and E. Chung, "Transit od matrix estimation using smartcard data: Recent developments and future research challenges," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103044, 2021.

[58] T. Zhang, Y. Li, H. Yang, C. Cui, J. Li, and Q. Qiao, "Identifying primary public transit corridors using multi-source big transit data," *International Journal of Geographical Information Science*, vol. 34, no. 6, pp. 1137–1161, 2020.

[59] S. Li, C. Zhuang, Z. Tan, F. Gao, Z. Lai, and Z. Wu, "Inferring the trip purposes and uncovering spatio-temporal activity patterns from dockless shared bike dataset in shenzhen, china," *Journal of Transport Geography*, vol. 91, p. 102974, 2021.

[60] N. Y. C. D. of Information Technology and T. (DOITT). Tlc trip record data. [Online]. Available: (https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page)

[61] D. M. Andrew Flowers, Reuben Fischer-Baum. Uber tlc foil response. [Online]. Available: (https://github.com/fivethirtyeight/uber-tlc-foil-response)

[62] K.-H. N. Bui, H. Yi, and J. Cho, "A multi-class multi-movement vehicle counting framework for traffic analysis in complex areas using cctv systems," *Energies*, vol. 13, no. 8, p. 2036, 2020.

[63] Q. Hao and L. Qin, "The design of intelligent transportation video processing system in big data environment," *IEEE Access*, vol. 8, pp. 13 769–13 780, 2020.

[64] J. Leitloff, D. Rosenbaum, F. Kurz, O. Meynberg, and P. Reinartz, "An operational system for estimating road traffic information from aerial images," *Remote Sensing*, vol. 6, no. 11, pp. 11 315–11 341, 2014.

[65] I. Brkić, M. Miler, M. Ševrović, and D. Medak, "An analytical framework for accurate traffic flow parameter calculation from uav aerial videos," *Remote Sensing*, vol. 12, no. 22, p. 3844, 2020.

[66] M. Laboratory. Mit traffic dataset. [Online]. Available: (https://mmlab.ie.cuhk.edu.hk/datasets/mit_traffic/index.html)

[67] alnfedorov. Traffic flow estimation data. [Online]. Available: (https://github.com/alnfedorov/traffic-analysis)

[68] E. Yao, X. Wang, Y. Yang, L. Pan, and Y. Song, "Traffic flow estimation based on toll ticket data considering multitype vehicle impact," *Journal of Transportation Engineering, Part A: Systems*, vol. 147, no. 2, p. 04020158, 2021.

[69] Tianchi. Kdd cup 2017 highway tollgates traffic flow prediction dataset. [Online]. Available: (https://tianchi.aliyun.com/dataset/dataDetail?dataId=60

[70] Z. Li, Q. Huang, G. J. Carbone, and F. Hu, "A high performance query analytical framework for supporting data-intensive climate studies," *Computers, Environment and Urban Systems*, vol. 62, pp. 210–221, 2017.

[71] M. Bakli, M. Sakr, and T. H. A. Soliman, "Hadooptrajectory: a hadoop spatiotemporal data processing extension," *Journal of geographical systems*, vol. 21, pp. 211–235, 2019.

[72] E. Zimányi, M. Sakr, and A. Lesuisse, "Mobilitydb: A mobility database based on postgresql and postgis," *ACM Transactions on Database Systems (TODS)*, vol. 45, no. 4, pp. 1–42, 2020.

[73] Z. Fang, L. Chen, Y. Gao, L. Pan, and C. S. Jensen, "Dragoon: a hybrid and efficient big trajectory management system for offline and online analytics," *The VLDB Journal*, vol. 30, pp. 287–310, 2021.

[74] X. Ding, L. Chen, Y. Gao, C. S. Jensen, and H. Bao, "Ultraman: A unified platform for big trajectory data management and analytics," *Proceedings of the VLDB Endowment*, vol. 11, no. 7, pp. 787–799, 2018.

[75] Z. Yao, Y. Zhong, Q. Liao, J. Wu, H. Liu, and F. Yang, "Understanding human activity and urban mobility patterns from massive cellphone data: Platform design and applications," *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 3, pp. 206–219, 2020.

[76] J. Dittrich and J.-A. Quiané-Ruiz, "Efficient big data processing in hadoop mapreduce," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2014–2015, 2012.

[77] V. Gurusamy, S. Kannan, and K. Nandhini, "The real time big data processing framework: Advantages and limitations," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 12, pp. 305–312, 2017.

[78] U. R. Pol, "Big data analysis: comparison of hadoop mapreduce, pig and hive," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 5, no. 6, pp. 9687–93, 2016.

[79] R. Anil, G. Capan, I. Drost-Fromm, T. Dunning, E. Friedman, T. Grant, S. Quinn, P. Ranjan, S. Schelter, and Ö. Yilmazel, "Apache mahout: machine learning on distributed dataflow systems," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 4999–5004, 2020.

[80] S. M. Razavi, M. Kahani, and S. Paydar, "Big data fuzzy c-means algorithm based on bee colony optimization using an apache hbase," *Journal of Big Data*, vol. 8, pp. 1–22, 2021.

[81] S. Kamburugamuve, G. Fox, D. Leake, and J. Qiu, "Survey of distributed stream processing for large stream sources," *Grids Ucs Indiana Edu*, vol. 2, pp. 1–16, 2013.

[82] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham *et al.*, "Storm@ twitter," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014, pp. 147–156.

[83] M. H. Iqbal, T. R. Soomro *et al.*, "Big data analysis: Apache storm perspective," *International journal of computer trends and technology*, vol. 19, no. 1, pp. 9–14, 2015.

[84] M. Jankowski, P. Pathirana, and S. Allen, *Storm Applied: Strategies for real-time event processing*. Simon and Schuster, 2015.

[85] W. Wingerath, F. Gessert, S. Friedrich, and N. Ritter, "Real-time stream processing for big data," *it-Information Technology*, vol. 58, no. 4, pp. 186–194, 2016.

[86] S. A. Noghabi, K. Paramasivam, Y. Pan, N. Ramesh, J. Bringhurst, I. Gupta, and R. H. Campbell, "Samza: stateful scalable stream processing at linkedin," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1634–1645, 2017.

[87] D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "A comparison on scalability for batch big data processing on apache spark and apache flink," *Big Data Analytics*, vol. 2, no. 1, pp. 1–11, 2017.

[88] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen *et al.*, "Mllib: Machine learning in apache spark," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.

[89] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica, "Graphx: Graph processing in a distributed dataflow framework," in *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, 2014, pp. 599–613.

[90] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi *et al.*, "Spark sql: Relational data processing in spark," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015, pp. 1383–1394.

[91] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache flink: Stream and batch processing in a single engine," *The Bulletin of the Technical Committee on Data Engineering*, vol. 38, no. 4, 2015.

[92] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz, "Hadoop gis: a high performance spatial data warehousing system over mapreduce," *Proceedings of the VLDB Endowment*, vol. 6, no. 11, pp. 1009–1020, 2013.

[93] M. Langhammer, J. Thober, M. Lange, K. Frank, and V. Grimm, "Agricultural landscape generators for simulation models: A review of existing solutions and an outline of future directions," *Ecological Modelling*, vol. 393, pp. 135–151, 2019.

[94] A. Eldawy and M. F. Mokbel, "Spatialhadoop: A mapreduce framework for spatial data," in *2015 IEEE 31st international conference on Data Engineering*. IEEE, 2015, pp. 1352–1363.

[95] S. You, J. Zhang, and L. Gruenwald, "Large-scale spatial join query

processing in cloud," in *2015 31st IEEE international conference on data engineering workshops*. IEEE, 2015, pp. 34–41.

[96] R. K. Lenka, R. K. Barik, N. Gupta, S. M. Ali, A. Rath, and H. Dubey, "Comparative analysis of spatialhadoop and geospark for geospatial big data analytics," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2016, pp. 484–488.

[97] J. Yu, Z. Zhang, and M. Sarwat, "Spatial data management in apache spark: the geospark perspective and beyond," *GeoInformatica*, vol. 23, pp. 37–78, 2019.

[98] D.-K. Tran, D.-Q. Hoang, V.-T. Le, M.-D. N. Thai, and T.-H. Do, "Real-time traffic flow prediction using big data analytics," in *Intelligence of Things: Technologies and Applications: The First International Conference on Intelligence of Things (ICIot 2022), Hanoi, Vietnam, August 17–19, 2022, Proceedings*. Springer, 2022, pp. 398–405.

[99] C. Bachechi, L. Po, and F. Rollo, "Big data analytics and visualization in traffic monitoring," *Big Data Research*, vol. 27, p. 100292, 2022.

[100] P. Chawla, R. Hasurkar, C. R. Bogadi, N. S. Korlapati, R. Rajendran, S. Ravichandran, S. C. Tolem, and J. Z. Gao, "Real-time traffic congestion prediction using big data and machine learning techniques," *World Journal of Engineering*, no. ahead-of-print, 2022.

[101] X. Li, H. Liu, W. Wang, Y. Zheng, H. Lv, and Z. Lv, "Big data analysis of the internet of things in the digital twins of smart city based on deep learning," *Future Generation Computer Systems*, vol. 128, pp. 167–177, 2022.

[102] S. Khan, S. Nazir, I. García-Magariño, and A. Hussain, "Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion," *Computers & Electrical Engineering*, vol. 89, p. 106906, 2021.

[103] J. R. Montoya-Torres, S. Moreno, W. J. Guerrero, and G. Mejía, "Big data analytics and intelligent transportation systems," *IFAC-PapersOnLine*, vol. 54, no. 2, pp. 216–220, 2021.

[104] Y. Shao, Y. Zhao, F. Yu, H. Zhu, and J. Fang, "The traffic flow prediction method using the incremental learning-based cnn-ltsm model: the solution of mobile application," *Mobile Information Systems*, vol. 2021, pp. 1–16, 2021.

[105] S. HS and C. BM, "An efficient multi-layer hybrid neural network and optimized parameter enhancing approach for traffic prediction in big data domain." *Special Education*, vol. 1, no. 43, 2022.

[106] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of milan and the province of trentino," *Scientific data*, vol. 2, no. 1, pp. 1–15, 2015.

[107] L. P. Loon, E. Refaie, and A. A. M. Faudzi, "Data analytics for traffic flow prediction in custom using long short term memory (lstm) networks," in *Journal of Physics: Conference Series*, vol. 2107, no. 1. IOP Publishing, 2021, p. 012006.

[108] Z. Lv, H. Fu, W. Tang, and X. Chen, "Traffic jam prediction based on analysis of residents spatial activities," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*. IEEE, 2020, pp. 37–40.

[109] A. I. Croce, G. Musolino, C. Rindone, and A. Vitetta, "Transport system models and big data: Zoning and graph building with traditional surveys, fcd and gis," *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 187, 2019.

[110] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. M. Altowaijri, "Smarter traffic prediction using big data, in-memory computing, deep learning and gpus," *Sensors*, vol. 19, no. 9, p. 2206, 2019.

[111] K.-H. N. Bui, H. Yi, H. Jung, and J. Seo, "Big data analytics-based urban traffic prediction using deep learning in its," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2019, pp. 270–273.

[112] Y. Hou, Z. Deng, and H. Cui, "Short-term traffic flow prediction with weather conditions: based on deep learning algorithms and data fusion," *Complexity*, vol. 2021, pp. 1–14, 2021.

[113] C. N. Babu, P. Sure, and C. M. Bhuma, "Sparse bayesian learning assisted approaches for road network traffic state estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1733–1741, 2020.

[114] J. Liu, T. Li, P. Xie, S. Du, F. Teng, and X. Yang, "Urban big data fusion based on deep learning: An overview," *Information Fusion*, vol. 53, pp. 123–133, 2020.

[115] J. Zhang, S. He, W. Wang, and F. Zhan, "Accuracy analysis of freeway traffic speed estimation based on the integration of cellular probe system and loop detectors," *Journal of Intelligent Transportation Systems*, vol. 19, no. 4, pp. 411–426, 2015.

[116] O. Gkountouna, D. Pfoser, and A. Züfle, "Traffic flow estimation using probe vehicle data," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2020, pp. 579–588.

[117] Y. Guo and L. Yang, "Reliable estimation of urban link travel time using multi-sensor data fusion," *Information*, vol. 11, no. 5, p. 267, 2020.

[118] M. Shoman, A. Aboah, and Y. Adu-Gyamfi, "Deep learning framework for predicting bus delays on multiple routes using heterogenous datasets," *Journal of Big Data Analytics in Transportation*, vol. 2, pp. 275–290, 2020.

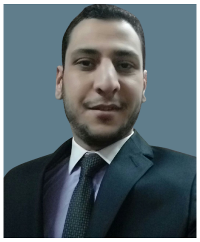
[119] S. Vergis, V. Komianos, G. Tsoumanis, A. Tsipis, and K. Oikonomou, "A low-cost vehicular traffic monitoring system using fog computing," *Smart Cities*, vol. 3, no. 1, pp. 138–156, 2020.

[120] R. Jiang, X. Song, Z. Fan, T. Xia, Z. Wang, Q. Chen, Z. Cai, and R. Shibasaki, "Transfer urban human mobility via poi embedding over multiple cities," *ACM Transactions on Data Science*, vol. 2, no. 1, pp. 1–26, 2021.

[121] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo, "Off-deployment traffic estimation—a traffic generative adversarial networks approach," *IEEE transactions on big data*, vol. 8, no. 4, pp. 1084–1095, 2020.

[122] V. Hassija, V. Gupta, S. Garg, and V. Chamola, "Traffic jam probability estimation based on blockchain and deep neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 3919–3928, 2020.

[123] Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, “Privacy-preserving traffic flow prediction: A federated learning approach,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, 2020.



Sayed A. Sayed Systems and Databases Analyst and Designer at the National Authority for Remote Sensing and Space Sciences (NARSS), Cairo, Egypt. He received a B.S. degree in Computer Sciences from the Faculty of Computers and Information Sciences, Asyut University, Asyut, Egypt, in 2007, and an M.Sc. degree in computer sciences from the Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University, Cairo, Egypt, in 2019. Currently, he pursues a Ph.D. degree in Computer Sciences. His primary research interests are GIS and its applications, Cloud Computing, Mobile Computing, Big Data, Statistical Analysis, Software Engineering, and Data Sciences.

Yasser Abdelhamid was born in Giza, Egypt in 1962. Abdelhamid earned his higher diploma degree in computer science in 1986 from Institute of Statistical Studies and Research (ISSR), Cairo University, Cairo, Egypt. He earned his MSc. degree and Ph.D. degree in computer science from ISSR, Cairo University, Cairo, Egypt in 1992, 1998 respectively. He worked as a computer programmer, system analyst, and IT manager during the period from 1986 to 1994. He worked as a research assistant at the Central Lab for Agricultural Expert Systems (CLAES) during the period from 1994 to 1998, and as a researcher at (CLAES) during the period from 1998 to 1999. During the period from 1999 to 2002 he worked as an Assistant professor at (ISSR), Cairo University, Egypt, then as an Assistant professor at the Community College, Computer Science department King Abdualaziz University, Tabuk, Saudi Arabia during the period from 2002 to 2006, then as an Associate professor at the Community College, Department of computer science, University of Tabuk, Tabuk, Saudi Arabia. Dr. Abdelhamid has many publications in the domain of artificial intelligence and its applications in the domain of agriculture and education. He was the chairman of computer science department in Community College, University of Tabuk during the period from 2006 to 2018. Now he is delegated to the Egyptian E-Learning University Cairo Egypt.

Hesham A. Hefny received the B.Sc., M.Sc. and Ph.D. all in Electronics and Communication Engineering from Cairo University in 1987, 1991 and 1998 respectively. He is currently a professor of Computer Science at the Faculty of Graduate Studies of Statistical Research (FGSSR), Cairo University. Prof. Hefny has authored more than 200 papers in international conferences, journals and book chapters. His major research interests include: computational intelligence (Neural networks – Fuzzy systems – Genetic algorithms – Swarm intelligence), Data mining, Uncertain Decision Making. He is a member in the following professional societies: IEEE Computer, IEEE Computational Intelligence, and IEEE System, Man and Cybernetics.