applied sciences

MDPI

Article

A New Vehicle Dataset in the City of Los Angeles for V2X and Machine Learning Applications

Ibtihal Ahmed Alablani [1,2,*] and Mohammed Amer Arafah [1]

[1] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; arafah@ksu.edu.sa
[2] Department of Computer Technology, Technical College, Technical and Vocational Training Corporation, Riyadh 11472, Saudi Arabia
* Correspondence: 438203904@student.ksu.edu.sa

Abstract: The fifth-generation (5G) network is the current emerging technology that meets the increasing need for higher throughputs and greater system capacities. It is expected that 5G technology will enable many new applications and services. Vehicle-to-everything (V2X) communication is an example of an application that is supported by 5G technology and beyond. A V2X communication system allows a vehicle to be connected to an entity, such as a pedestrian, another vehicle, infrastructure, and a network, to provide a robust transportation solution. It uses many models and strategies that are usually based on machine learning (ML) techniques, which require the use of a vehicle dataset. In this paper, a real vehicle dataset is proposed that was generated in the city of Los Angeles (LA). It is called the Vehicle dataset in the city of LA (VehDS-LA). It has 74,170 samples that are located on 15 LA streets and each sample has 4 features. The LA dataset has been opened to allow researchers in V2X and ML fields to use it for academic purposes. The main uses of the VehDS-LA dataset are studies related to 5G networks, vehicle automation, or ML-Based vehicle mobility applications. The proposed dataset overcomes limitations experienced by previous related works.

Keywords: 5G; Google Maps; IoV; ITS; Los Angeles; machine learning; V2X; vehicle dataset

Citation: Alablani, I.A.; Arafah, M.A. A New Vehicle Dataset in the City of Los Angeles for V2X and Machine Learning Applications. *Appl. Sci.* **2022**, *12*, 3751. https://doi.org/10.3390/app12083751

Academic Editors: Omar Sami Oubbati and Abderrezak Rachedi

Received: 22 February 2022
Accepted: 4 April 2022
Published: 8 April 2022

1. Introduction

The fifth generation is the current generation of cellular networks and aims to make significant improvements in service quality to enhance reliability, throughput, delay, and connectivity [1]. Some examples of 5G emerging applications are smart houses, intelligent transportation, health monitoring, and the Internet of Things (IoT) [2]. The IoT is an emerging revolution that associates physical things to the Internet [3]. The Internet of Vehicles (IoV) is a subset of the IoT in which vehicles are connected to the internet and can send and receive data [4,5]. Vehicle-to-everything technology is an evolution towards the IoV era and the Intelligent Transportation System (ITS). V2X aims to enhance road safety, the reliability of communications, and traffic efficiency [6,7]. There are four kinds of V2X communications, as shown in Figure 1: vehicle-to-vehicle (V2V), vehicle-to-pedestrian (V2P), vehicle-to-infrastructure (V2I), and vehicle-to-network (V2N). An ITS provides end users with comfort and safety by employing many models and strategies, the majority of which are based on machine learning techniques [8].

Machine learning (ML) is a branch of artificial intelligence (AI) that allows computers to learn from data without having to be explicitly programmed [9,10]. ML techniques have recently received a lot of attention and the future prospects for this technology are extremely bright [11]. There are three types of learning techniques, i.e., supervised, unsupervised, and reinforcement methods. Supervised learning uses labeled data to perform a specific learning task, while unsupervised learning uses unlabeled data [12]. Reinforcement learning is a kind of learning that uses reward signals to make the computer learn; the learner is not taught which actions to take, but it must try to see which ones give the most rewards [13].

Building an effective ML model needs data features that are closely associated with each other and with the prediction target [14].

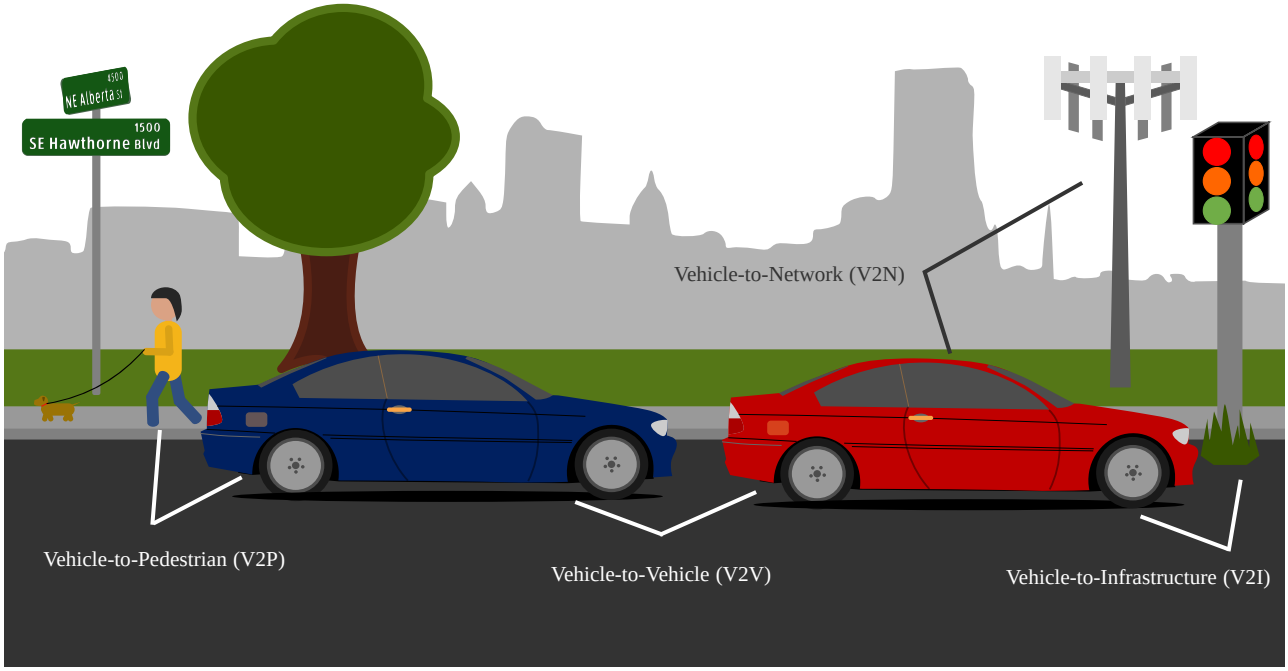

Figure 1. Types of vehicle-to-everything communications.

A smart city is an urban area that utilizes advanced technologies to make life easier for its citizens [15,16]. Smart cities focus on improving the quality of services provided to individuals through the management of public resources, convenience, maintenance, and sustainability [17]. They can overcome issues related to the fields of health, education, environment, governance, economic, and transportation [18,19]. By 2025, it is expected that there will be 88 smart cities around the world. Based on the global smart cities index, the top ten smart cities in terms of smart infrastructure, economy, and governance are London, New York, Paris, Berlin, Tokyo, Los Angeles, Singapore, Seoul, Chicago, and Hong Kong [20]. Three of these top cities are located in the United States of America. New York is one of the largest cities in the world and it has many attractions for tourists and a diversity of cultures, as 40% of its residents come from other countries [21]. Los Angeles lies in Southern California and it is the United States' second-largest city in terms of population [22,23]. Chicago is located in northeastern Illinois and it is the third largest city in the United States in terms of population [24,25].

In the field of transportation, a very limited number of real vehicle databases is available for scientists and engineers to perform academic research related to V2X and machine learning. The existing databases require effort, time, and equipment to collect data samples. In addition, the resulting data lack location accuracy and up-to-date versions.

The main contribution of this paper is proposing a real vehicle dataset, called VehDS-LA that was generated accurately using Google Maps in the city of Los Angeles, California. The database has 74,170 samples that are located on 15 LA streets and each sample has 4 vehicle features. This paper introduces a general mechanism in generating vehicle datasets for smart cities based on Google My Maps. The main uses of the proposed dataset, which was collected in the smart city of LA, are studies related to 5G networks, automation, and driverless vehicles, together with ML-based vehicle mobility applications.

The rest of this paper is arranged as follows. Section 2 discusses related works on generating real vehicle datasets. Section 3 illustrates the proposed LA vehicle dataset in terms of how it was created, its contents, and its representation of it on the LA map. Section 5 concludes the paper and highlights suggested future directions.

2. Literature Review

In this section, works on generating vehicle datasets to be used in many fields are discussed and their limitations are given.

2.1. Related Work

In [26], Jensen et al., who are researchers at the Aalborg University department of Development and Planning, recorded a vehicle dataset during an intelligent speed adaptation project called INFATI. The dataset was generated in February and March 2001 in Aalborg, Denmark. It is non-commercial and is available free of charge for researchers. Each vehicle was equipped with a Global Positioning System (GPS) receiver in addition to a small computer. When vehicles were moved, their GPS location was sampled every second. When vehicles were parked, no sampling was generated. The process of collecting vehicle information took more than a month. The generated datasets were saved in Universal Transverse Mercator (UTM) format. Figure 2 shows the vehicle samples on the INFATI dataset. In [27], the authors found that the resolution of the INFATI dataset was low and inconsistent.

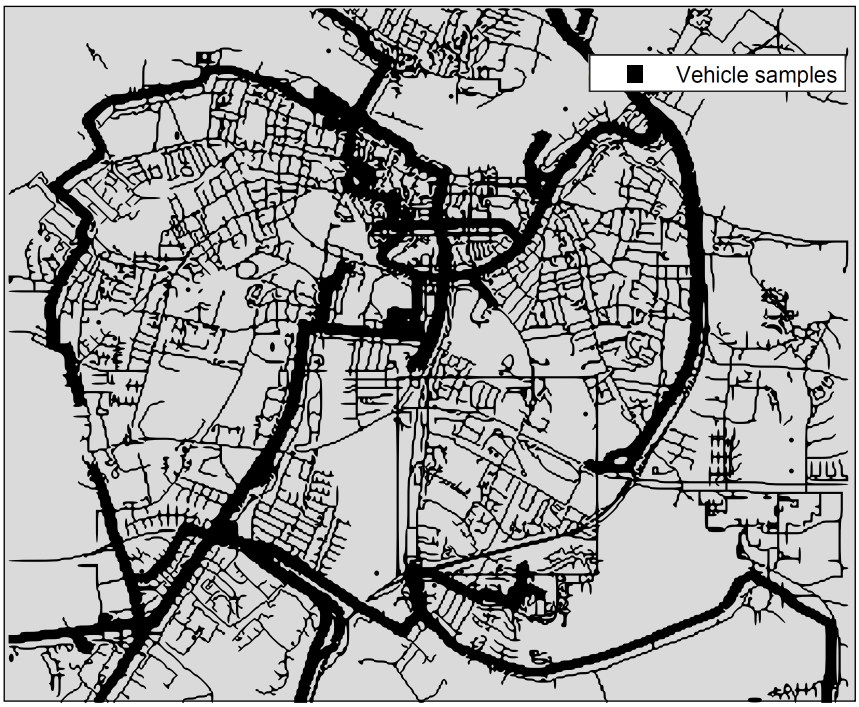

Figure 2. Illustration of vehicle samples of the INFATI dataset.

In [28], Cho and Kim introduced a vehicle dataset which is based on real data that were recorded on 13 February 2017 in the city of Los Angeles. It was created for research purposes to investigate the movement of vehicles in a real-world environment. The database includes 128,199 samples, distributed over 64 comma-separated values (CSV) files. Figure 3 depicts a snapshot from one of these cvs files and Figure 4 shows the locations of the vehicle samples on the LA map. Five kinds of sensors have been used: GPS, orientation, acceleration, gyroscope, and magnetic field sensors. A platform called MediaQ was utilized to achieve vehicle sample collection, organization and sharing of the recorded dataset. The MediaQ platform includes a server and an application for smartphones. It can be used to record videos in MP4 format. Figure 5 shows how a smartphone was mounted during the data recording process using the MediaQ application. The driving time to collect the data took about 22.4 h and the driving distance was 1177.4 km [23].

	date	time	speed	kspeed	accSpeed	azimuth	pitch	roll	light	gps_lat	gps_lon
1	2017-02-10	09:52:01	0.000000	0.000000	−0.322648	227.762207...	−20.505966...	12.5901489...	0.62646192...	34.0231058...	−118.29183...
2	2017-02-10	09:52:01	0.000000	0.000000	−0.521802	228.961410...	−19.907852...	11.3122558...	0.62646192...	34.0231058...	−118.29183...
3	2017-02-10	09:52:01	0.000000	0.000000	−0.558541	230.147094...	−18.576416...	5.97937011...	0.62646192...	34.0231058...	−118.29183...
4	2017-02-10	09:52:01	0.000000	0.000000	0.150292	231.142959...	−19.102294...	5.45849609...	0.62646192...	34.0231058...	−118.29183...
5	2017-02-10	09:52:01	0.000000	0.000000	0.105919	231.785247...	−18.877243...	5.53344726...	0.62646192...	34.0231058...	−118.29183...
6	2017-02-10	09:52:01	0.000000	0.000000	0.385535	232.405746...	−20.152023...	5.22436523...	0.62646192...	34.0231058...	−118.29183...
7	2017-02-10	09:52:01	0.000000	0.000000	−0.347830	232.913711...	−19.476623...	5.44628906...	0.62646192...	34.0231058...	−118.29183...
8	2017-02-10	09:52:01	0.000000	0.000000	−0.293108	233.665542...	−20.153015...	4.78106689...	0.62646192...	34.0231058...	−118.29183...
9	2017-02-10	09:52:02	8.194253	4.292042	7.807077	234.303314...	−19.854064...	4.18310546...	0.62646192...	34.0231062...	−118.29186...
10	2017-02-10	09:52:02	8.194253	4.292042	8.973851	234.653244...	−20.529373...	4.03692626...	0.62646192...	34.0231062...	−118.29186...
11	2017-02-10	09:52:02	8.194253	4.292042	8.793981	234.883789...	−20.529769...	3.73620605...	0.62646192...	34.0231062...	−118.29186...
12	2017-02-10	09:52:02	8.194253	4.292042	9.357249	234.965835...	−20.379348...	3.80395507...	0.62646192...	34.0231062...	−118.29186...
13	2017-02-10	09:52:02	8.194253	4.292042	9.622275	234.915924...	−20.892562...	4.26489257...	0.62646192...	34.0231062...	−118.29186...
14	2017-02-10	09:52:03	8.564157	4.574732	10.137592	234.719665...	−20.891510...	4.78466796...	0.62646192...	34.0231065...	−118.29190...

Figure 3. The vehicle dataset introduced by Cho and Kim in LA.

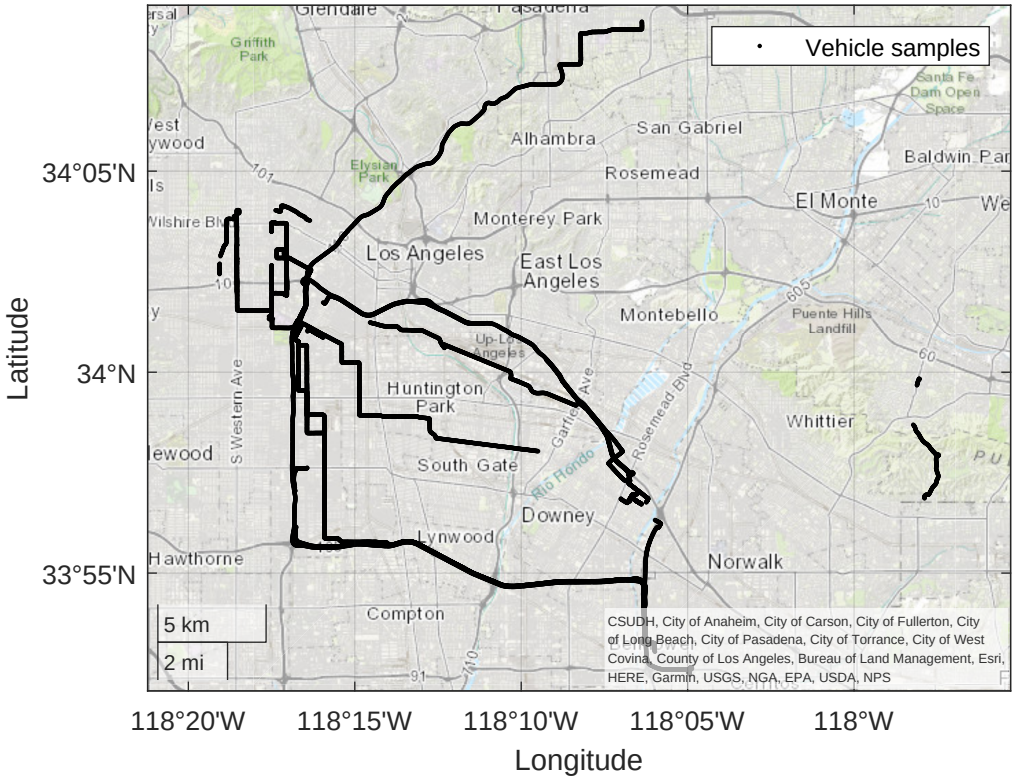

Figure 4. Illustration on the LA map of vehicle samples collected by Cho and Kim.

Figure 5. Smartphone mounted on a vehicle dashboard to generate the vehicle dataset.

In [27], Alzyout et al. proposed a real vehicle dataset in Jordan in 2019. An Android application called Ultra GPS Logger (UGL) was used to collect the samples, using a Samsung Galaxy S Duos 2 S7582 smartphone, as shown in Figure 6. The vehicle sample generation process took about eight months (from January to August). Once per second, vehicle information was collected, recording GPS position, speed, direction, and distance between successive positions. The dataset covered a distance of around 6600 kilometers.

Figure 6. Using the Ultra GPS Logger application on an Android smartphone.

2.2. Limitations of the Related Works

The limitations of Cho and Kim's dataset, which was collected in LA, are the following:

- Most of the vehicle samples are located on freeways, such as Harbor, Passadena, and Santa Ana, as shown in Figure 4. The distribution of vehicle samples should not focus on a particular type of street.
- The geographical distance between two consecutive samples is large around 20 m, as shown in Figure 7. A large space between samples is undesirable when applying machine learning techniques.
- The driving time for collecting the LA vehicle dataset was long (around 22 h).

- The recording process of the dataset required considerable effort, equipment, and tools (i.e, five types of sensors, MediaQ platform, smartphone, and a vehicle smartphone holder).
- The database includes samples that are not moving (i.e., vehicles with a speed of 0 km/h).

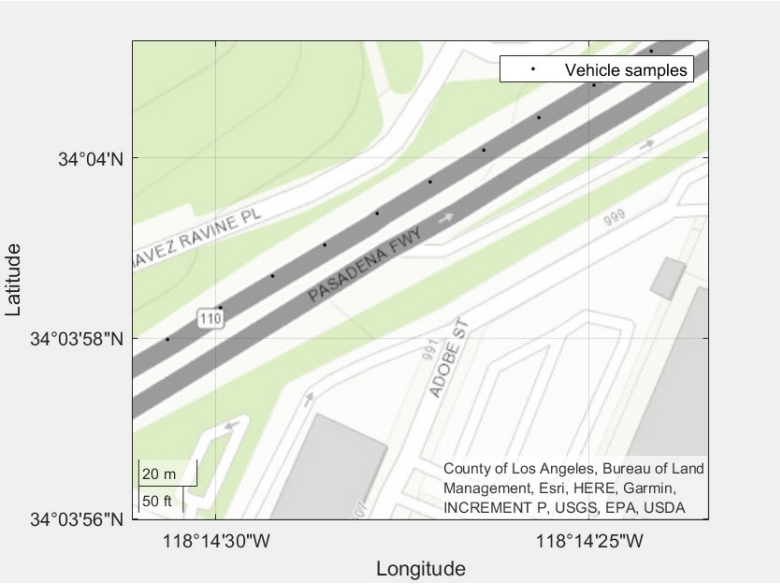

Figure 7. The distance between two consecutive geographical points.

In general, based on the previous works on recording vehicle datasets represented in this section, we find the following limitations:

- The long time and huge effort required to record vehicle dataset samples.
- The need for equipment in the vehicle during the collecting process, such as GPS receivers, computers, and smartphones.
- The accuracy of the resulting samples is not guaranteed and it may deviate from the road on which the vehicles moved.
- Difficulty in updating and adding new samples to the resulting dataset, whereas, after some years, changes may occur to the streets on which the data were collected.

Consequently, there is an urgent need to provide a general and simple mechanism to generate a vehicle dataset that considers different types of roads. In addition, the geographic distance between samples should be small, so that the dataset can be used to train a good machine learning model. In fact, Google Maps is a powerful mapping service that can be utilized to develop a new mechanism in generating vehicle datasets.

3. The Proposed Vehicle Dataset

3.1. Dataset Generation Method

In this paper, a real vehicle dataset in the city of Los Angeles is proposed. The VehDS-LA was generated by utilizing Google Maps and the MATLAB R2021b simulator. The database production process is divided into two main phases, as shown in Figure 8.

- Phase 1: Creating Driving Routes: This phase was implemented through Google Maps. It includes three steps:
  - Step 1: Creating a new map of the city of Los Angeles.
  - Step 2: Adding driving routes for all the selected streets (15 streets in this study).
  - Step 3: Exporting a Keyhole Markup Language (KMZ) file for each driving route. An example of the contents of a KMZ file is shown in Figure 9.

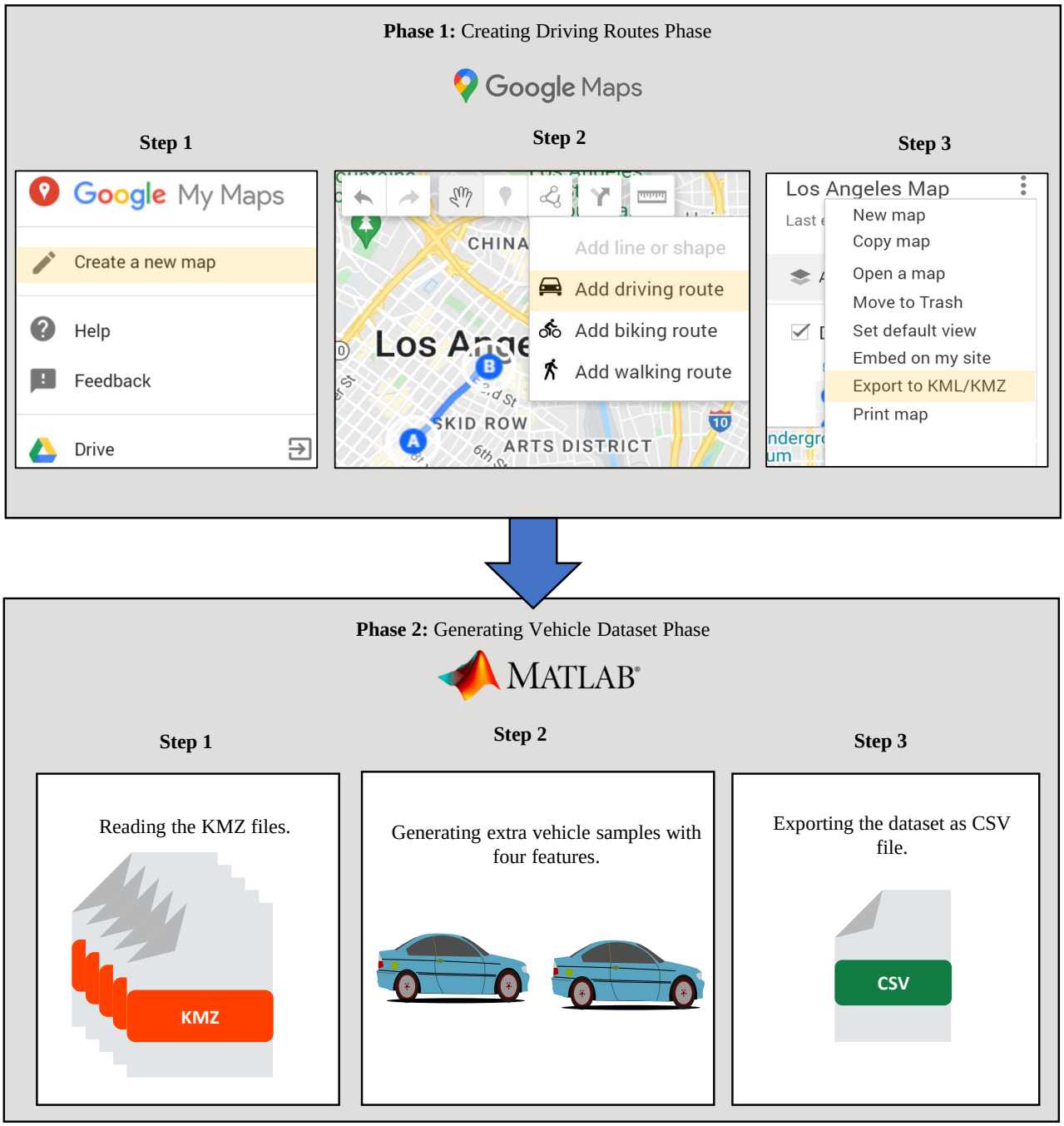

Figure 8. The phases of generating the proposed vehicle DS.

Geometry	Name	Lon	Lat	BoundingBox	Color
'Line'	'San Pedro / 16th'	*1x120 double*	*1x120 double*	[−118.2559,34.0300; −118.2397,34.0525]	[0.0706,0.4039,1]
'Point'	'San Pedro / 16th, Los Angeles, CA 90015, USA'	-118.2559	34.0300	[−118.2559,34.0300; −118.2559,34.0300]	[0.6758,0.8438,0.8984]
'Point'	'Temple St & Judge John Aiso St, Los Angeles, CA 90012, USA'	-118.2397	34.0525	[−118.2397,34.0525; −118.2397,34.0525]	[0.6758,0.8438,0.8984]

Figure 9. An example of the contents of a KMZ file.

- Phase 2: Generating the Vehicle Dataset: This phase was performed using the MATLAB simulator. This phase has three steps:
  - Step 1: Reading the KMZ files and converting them into structure objects.
  - Step 2: Generating extra vehicle samples so that the distance between two samples is small (0.25 m in this study). For each vehicle sample, four features were assigned: (1) latitude coordinate, (2) longitude coordinate, (3) vehicle speed, and (4) vehicle azimuth. The speeds were generated randomly in the range from 10 to 40 km per hour (km/h).
  - Step 3: Exporting the proposed VehDS-LA as a comma-separated values (CSV) file.

3.2. LA Vehicle Dataset Characteristics

The generated LA vehicle dataset has 74,170 samples that are located on 15 LA streets. Figure 10 shows the locations of the proposed vehicle samples on the LA map. Each sample has four vehicle features: latitude coordinate, longitude coordinate, vehicle speed, and vehicle azimuth. The azimuth refers to the angle between the vehicle direction and north. Figure 11 displays an overview of the proposed VehDS-LA. A description of the vehicle dataset fields is given in Table 1. Figure 12 gives a snapshot of the proposed dataset.

Table 1. Description of the proposed LA vehicle dataset fields.

Field Name	Description	Values
'STREET_NAME'	Name of LA street where vehicle is located.	'San Pedro St', 'S Hill St', 'N Hill St', 'Flower St', 'S Hope St', 'E Olympic Bivd', 'E 3rd St', 'W 3rd St', 'E 6th St', 'W 6th St', 'E 9th St', 'W 9th St', 'James M Wood Blvd', 'S Los Angeles St', 'N Los Angeles St'
'LAT'	Latitude coordinate of vehicle.	[34.03 to 34.056]
'LON'	Longitude coordinate of vehicle.	[−118.27 to −118.24]
'AZIMUTH'	Angle between vehicle direction and north in degrees.	[0 to 342.74]
'KSPEED'	Speed of vehicle in km/h.	[10 to 40]

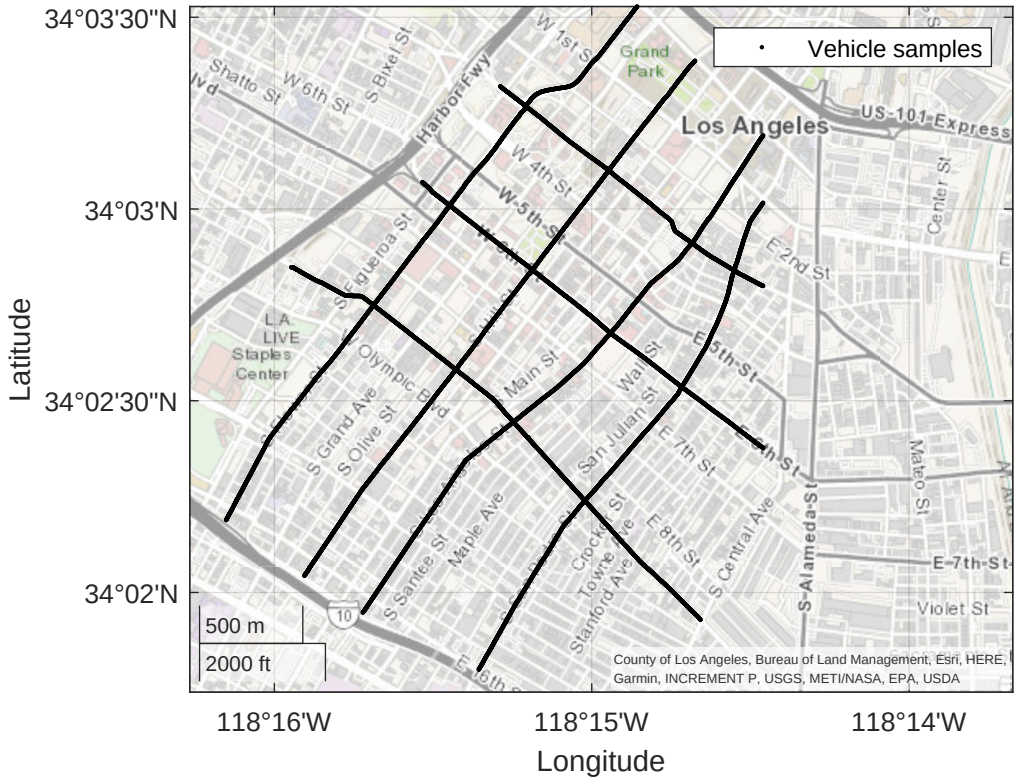

Figure 10. Illustration of the proposed LA vehicle samples.

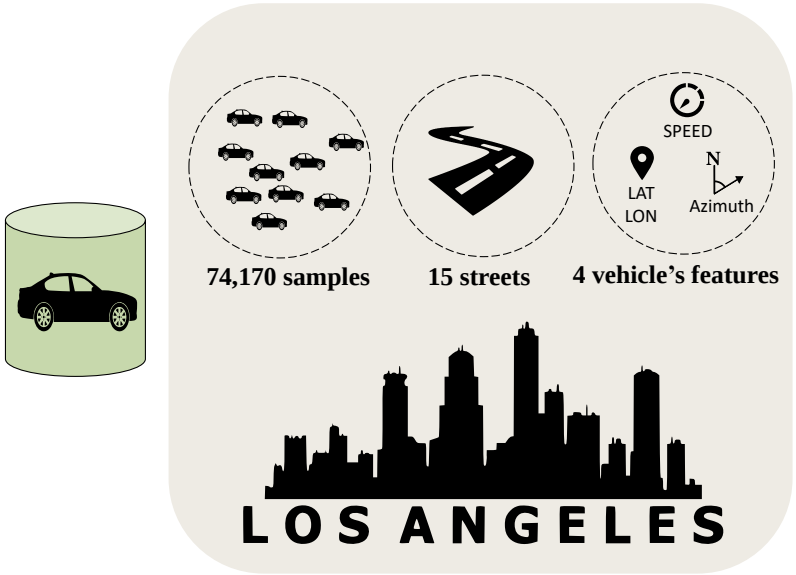

Figure 11. An overview of the proposed LA vehicle dataset.

	STREET_NAME	LAT	LON	AZIMUTH	KSPEED
1	S Los Angeles St	34.0534	−118.2408	35.0856	30
2	S Los Angeles St	34.0529	−118.2412	32.6282	15
3	S Los Angeles St	34.0525	−118.2416	34.6226	15
4	S Los Angeles St	34.0524	−118.2417	32.1978	20
5	S Los Angeles St	34.0521	−118.2419	33.3093	30
6	S Los Angeles St	34.0518	−118.2421	32.3576	20
7	S Los Angeles St	34.0515	−118.2424	35.6807	40
8	S Los Angeles St	34.0513	−118.2425	33.2618	35
9	S Los Angeles St	34.0511	−118.2427	33.5376	40
10	S Los Angeles St	34.0511	−118.2427	31.3506	10
11	S Los Angeles St	34.0507	−118.2430	35.3816	15
12	S Los Angeles St	34.0506	−118.2430	31.4043	35
13	S Los Angeles St	34.0505	−118.2432	31.4043	30
14	S Los Angeles St	34.0503	−118.2433	33.5379	10
15	S Los Angeles St	34.0500	−118.2435	32.5112	25

Figure 12. The proposed LA vehicle dataset.

3.3. The Advantages of the Proposed Dataset

The following list presents the advantages of the proposed VehDS-LA dataset compared to related dataset generation works:

- Generating the database does not require a long time, as in the related works, where it took days and months.
- The accuracy of the positions of vehicle samples which were produced based on Google Maps and the MATLAB simulator. It was verified that the samples are located on the LA streets without any deviation.
- There is no need to install special equipment and devices in the vehicle, such as a GPS receiver, small computer, or smartphone.
- The number of dataset samples is large and each sample has four features, which are the most important features of a vehicle for traffic simulation purposes.
- The method of generating the proposed VehDS-LA dataset introduces a general mechanism that can be followed in generating new databases in any region of the world on the basis of Google Maps.

In fact, the VehDS-LA dataset is based on the current state of the selected streets of Los Angeles city. After a few years, the database may need to be updated according to future street-related information.

3.4. The Uses of the VehDS-LA Dataset

The proposed VehDS-LA is appropriate for use with applications related to 5G technology, machine learning techniques and transportation systems. The main uses of the VehDS-LA are:

- **5G network studies:** A heterogeneous ultra-dense network is a 5G-enabling technology that consists of a high density of small cells in addition to the legacy Long-Term Evolution (LTE) macro cells. HUDN aims to meet the requirements of increased capacity, low latency, and distributed traffic load with low installation cost [23,29]. The major issues associated with 5G HUDNs are cell selection, interference mitigation, and resource allocation [30]. Cell selection refers to the process of choosing the serving base station to which a vehicle will connect. The conventional approach of selecting cells is

based on the received signal strength indicator (RSSI) value. In fact, this approach is inefficient in 5G HUDNs due to the existence of a large number of cells with different distribution and sizes [31]. Figure 13 shows the cell selection issue in an HUDN, where a red vehicle should select a serving cell, and RSSI values are not enough.

HUDNs suffer from two types of interference: co-tier and cross-tier interference. Co-tier interference occurs between homogeneous cells, while cross-tier interference happens between heterogeneous cells [32], as shown in Figure 14. The proposed VehDS-LA dataset can be used in studies related to 5G HUDNs.

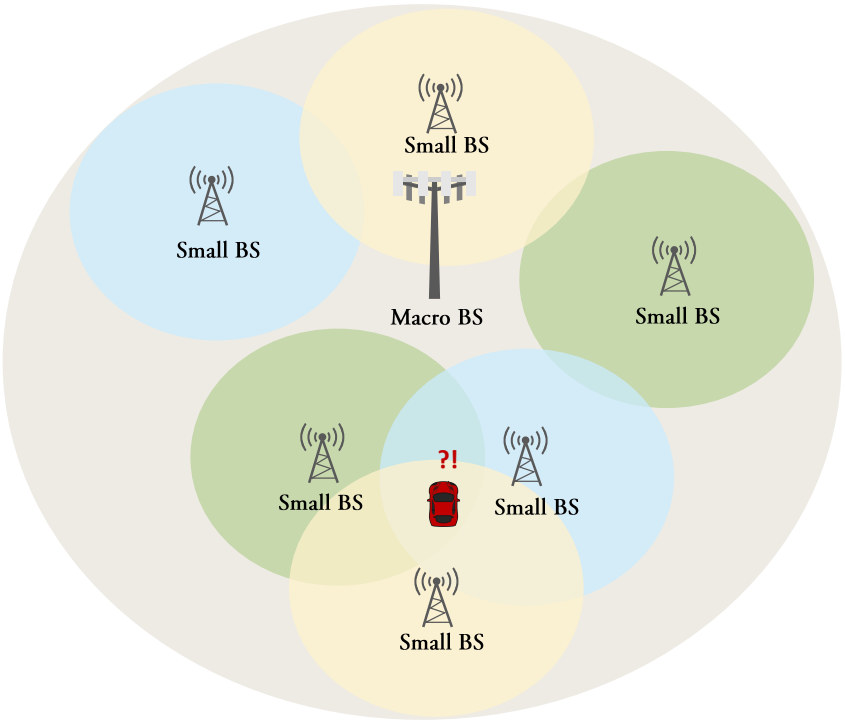

Figure 13. Cell selection issue in HUDNs.

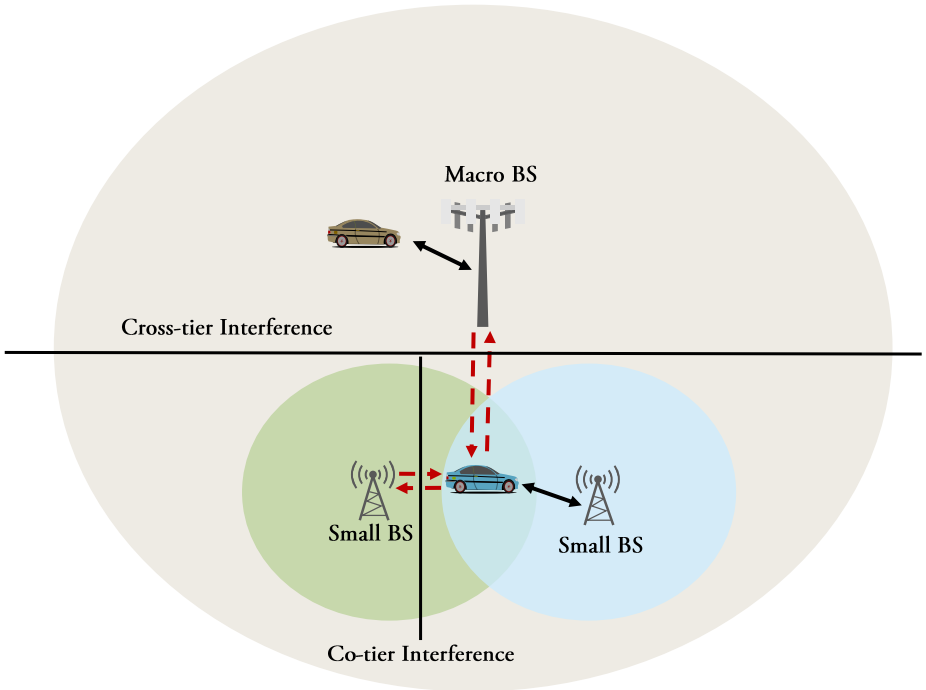

Figure 14. Interference issue in HUDNs.

- **Automation and driverless vehicles studies:** Nowadays, vehicle automation is becoming a solution that is used to provide road safety and to prevent accidents [33]. The Society of Automotive Engineers (SAE) defines six levels of vehicle automation, as illustrated in Figure 15. The first three levels, i.e., levels 0 to 2, require driver attention. On the other hand, levels 3 to 4 give part of the responsibility for driving and monitoring roads to the vehicle itself, while level 5 provides full automation of vehicles [34]. Thus, the proposed dataset includes the essential vehicle features, i.e., geographical latitude and longitude coordinates, azimuths, and speeds of vehicle samples, which can be used in research related to vehicle automation.

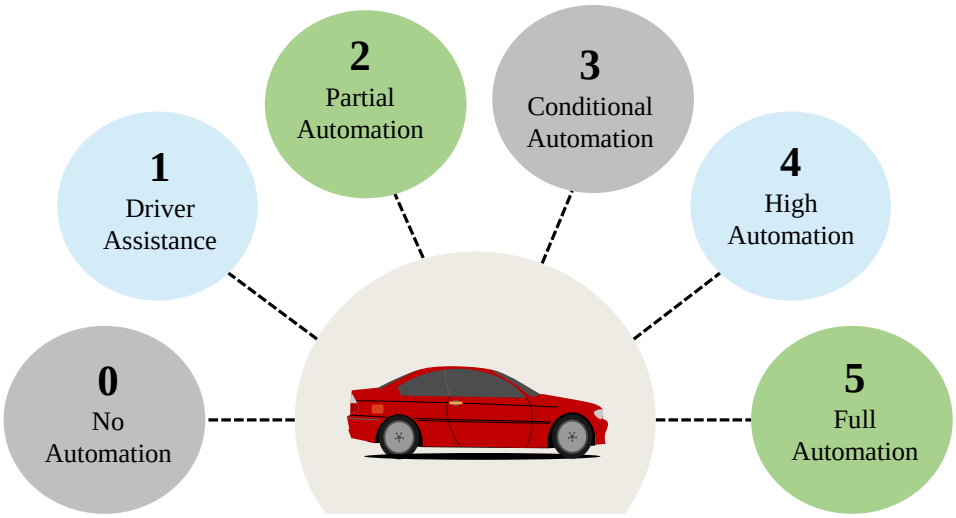

Figure 15. The levels of vehicle automation.

- **ML-based vehicle mobility studies:** ML techniques provide remarkable opportunities in several fields, including transportation [35]. A good machine learning model needs a large number of samples to train the ML model [36]. Recent works that focus on vehicle movement issues, including [2,37], relied on solving research problems using machine learning algorithms, such as artificial neural networks (ANN) and support vector machine (SVM), Naive Bayes (NB), and Tree-based techniques. Figure 16 represents the process of building a machine learning model that is based on supervised learning to solve a vehicle mobility issue. The building process passes through many phases: data cleaning, data labeling, data dividing, ML model training, and ML model testing [2].
  - Data cleaning: In this phase, data that will not be used to solve the research problem are removed [38].
  - Data Labeling: This refers to the process of tagging vehicle samples so that the ML model can learn from it [39].
  - Data Dividing: This refers to splitting the dataset into two parts: training and testing sets. The dataset is usually divided into 80:20 or 70:30 ratios [40].
  - ML Model Training: The training set is used train the ML model.
  - ML Model Testing: The test set is used to evaluate the performance of the trained ML model.

  Research that is based on solving vehicle mobility problems using ML algorithms can utilize the proposed database. It provides a sufficient number of vehicle samples, i.e., 74,170 samples, that can be used for ML model training and testing. Moreover, the accuracy of the locations of vehicle samples was verified without any deviation.

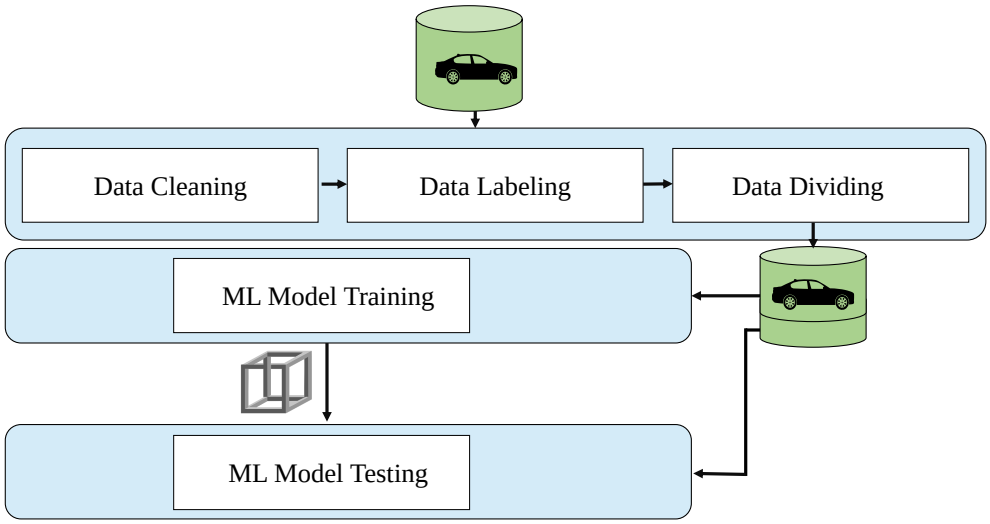

Figure 16. Building a machine learning model.

- **Intelligent transportation system studies in the LA smart city:** Smart city and intelligent transportation system are recently developed concepts [41]. The term ITS is defined as a comprehensive system that consists of vehicles and transportation infrastructure and it performs communication, controlling, and information processing in smart cities to facilitate their environmental sustainability [42,43]. The proposed VehDS-LA can be used for studies related to ITS in the downtown of the city of Los Angeles, as shown in Figure 17. Our VehDS-LA includes information of vehicle samples in terms of their real-world geographical locations, as well as the vehicles' movement-related information in terms of directions and speeds based on the infrastructure of LA streets. Therefore, studies related to vehicle-to-vehicle, vehicle-to-pedestrian, and vehicle-to-network communications in LA city can utilize the vehicles information stored in our proposed dataset.

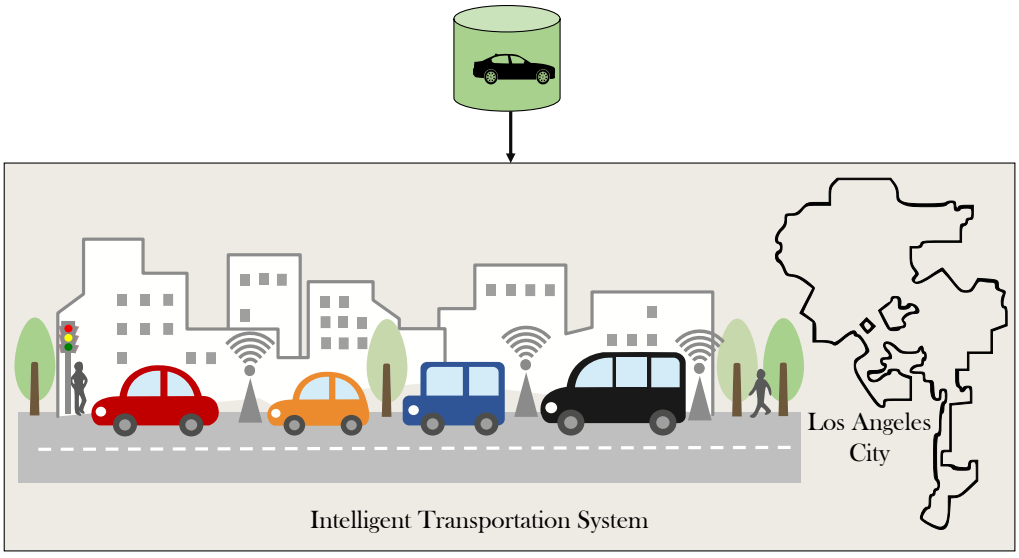

Figure 17. Using the proposed vehicle dataset in ITS studies in LA city.

- **SDN-based vehicular networks studies:** Software-defined networking is one of the most recent network architectures that aims to facilitate the network management task and to enhance the utilization of network resources in an efficient way [44,45]. The architecture of SDN is made up of three components, which are data plane, control plane, and application plane [46]. The data plane comprises network devices that are responsible for forwarding data [47]. The control plane is made up of a set of

SDN controller(s) to control and manage operations of the whole network [48]. The application plane consists of end user applications that interact with SDN controller(s) to perform specific tasks [49,50]. Southbound interface is used to perform the communication between the data and control planes based on a standardized protocol [51]. Northbound interface is utilized to establish the communication between the control and application planes [48]. Figure 18 shows the architecture of SDN-based vehicular networks, where vehicle samples of our proposed VehDS-LA can be utilized to construct a vehicular network. The studies that are focused on SDN-based vehicular networks can benefit from our proposed dataset in performing vehicle mobility management and supervision tasks by SDN, where realistic vehicle location coordinates and movement-related information exist.

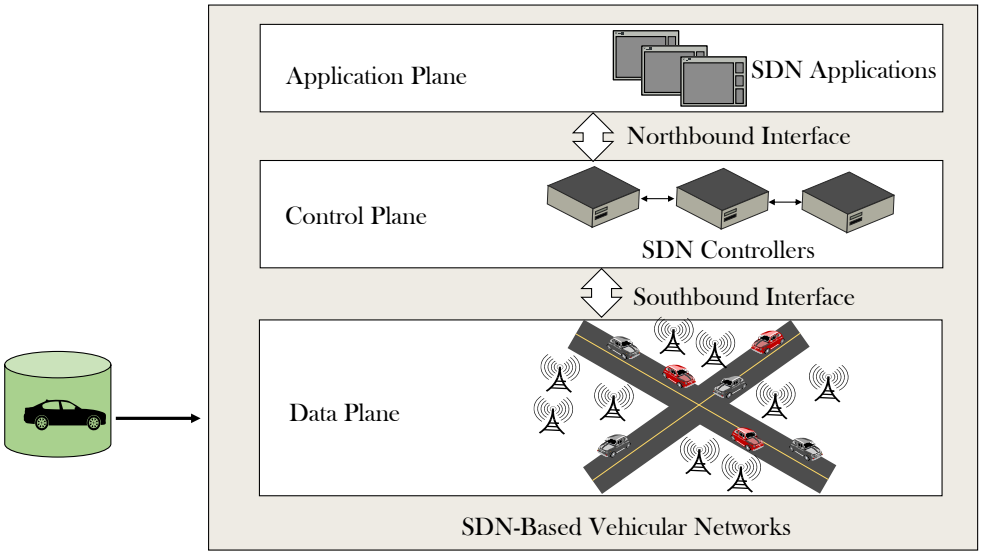

Figure 18. Using the proposed VehDS-LA in SDN-based vehicular network studies.

3.5. Ethical Issues

The proposed VehDS-LA dataset is available for research purposes on the GitHub website [52]. When the proposed dataset is used for academic or research purposes, there are no proprietary or copyright restrictions. However, this paper should be cited in the references list, indicating the title of the article, names of authors, publication year, journal information, volume number (issue number), and page range.

4. Using the Proposed VehDS-LA to Perform Cell Selection in 5G Networks

In this section, the proposed VehDS-LA was used to perform the cell selection process in 5G networks. The distribution of 5G small base stations (BSs) depends on a dataset that was published by data.LAcity.org (accessed on 22 February 2022) [53]. The dataset contains information about 5G small BSs in the city of Los Angeles, which are attached to street lights. To model the network and to accomplish the cell selection process, MATLAB 2021b simulator was used because it provides a powerful platform. The simulation parameters, which are used in this work, are shown in Table 2. Path loss is modeled based on a model called urban microcell-line-of-sight (UMi-LOS) (street canyon), which is described in the 3rd-Generation Partnership Project (3GPP) technical report 38.901 version 16.1 [54].

Table 2. Simulation parameters.

Simulation Parameters	Values
Transmit power (dBm)	30
Path loss model (dB)	3GPP UMi Model
Carrier frequency (GHz)	28
Number of 5G small BSs	198
Small BS height (meters)	10
Small cell radius (meters)	600
RSSI threshold (dBm)	−90
Handover delay (ms)	50 [55]
Simulation time (sec)	500

Handover (HO), which is the process of transferring the connection from one BS to another [56], is performed based on the strongest value of the received signal strength indicator (RSSI). Figure 19 displays the average number of horizontal handovers, which occur between small BSs, under various vehicle speeds. Sojourn time of vehicles inside a serving small cell is shown in Figure 20. The results demonstrate that there is an inverse relationship between the sojourn time and the number of horizontal handovers. As the vehicle speed increases, the sojourn time decreases and the number of horizontal HOs will increase.

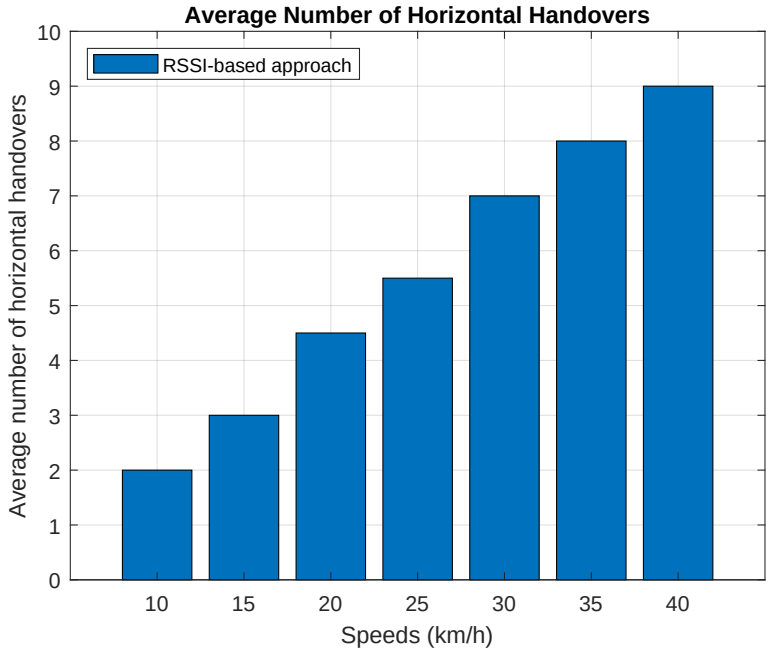

Figure 19. Average number of horizontal handovers under various speeds.

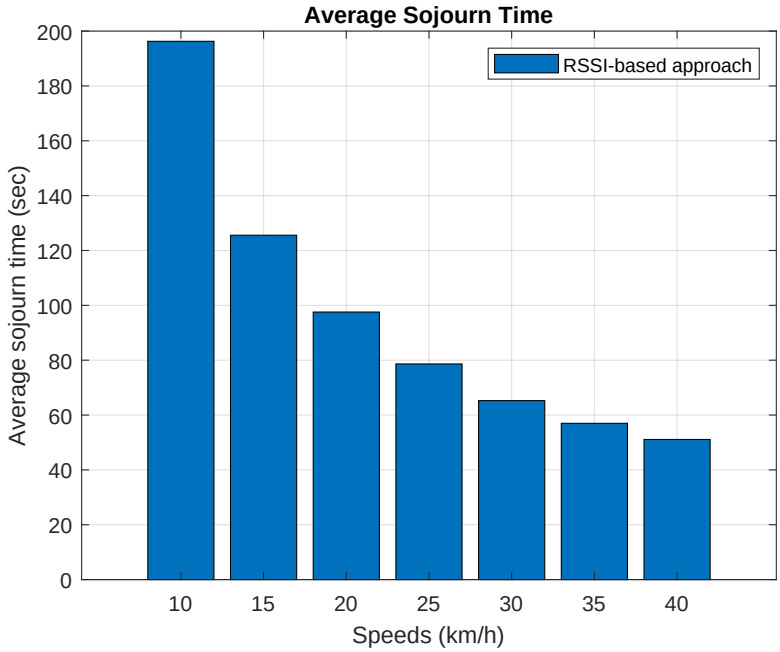

Figure 20. Average sojourn time under various speeds.

If the sojourn time of a vehicle within a small cell is less than the handover delay, HO failure happens. Unnecessary handover occurs when the sum of HO latencies to move into and out of a 5G small cell is longer than the sojourn time in the small cell [31]. Figures 21 and 22 show the averages of the number of HO failures and unnecessary HOs, respectively.

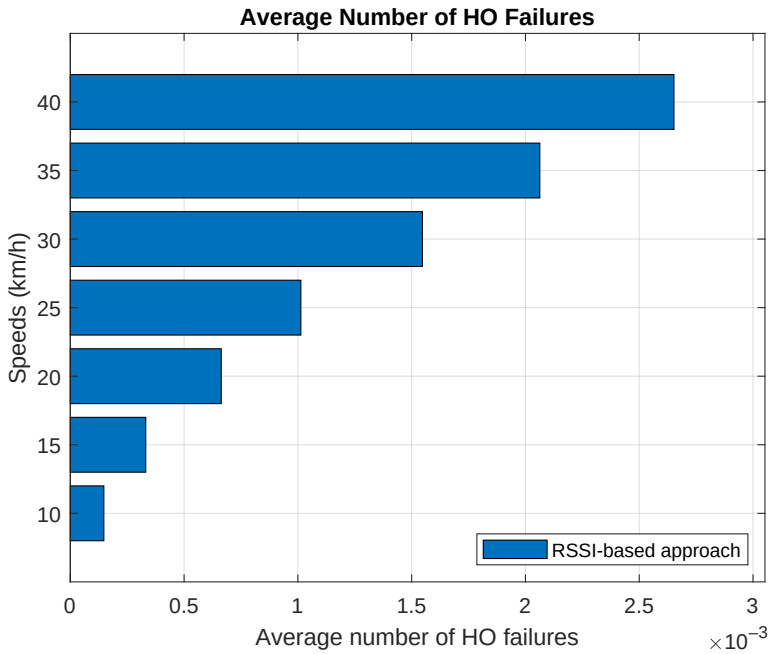

Figure 21. Average number of HO failures under various speeds.

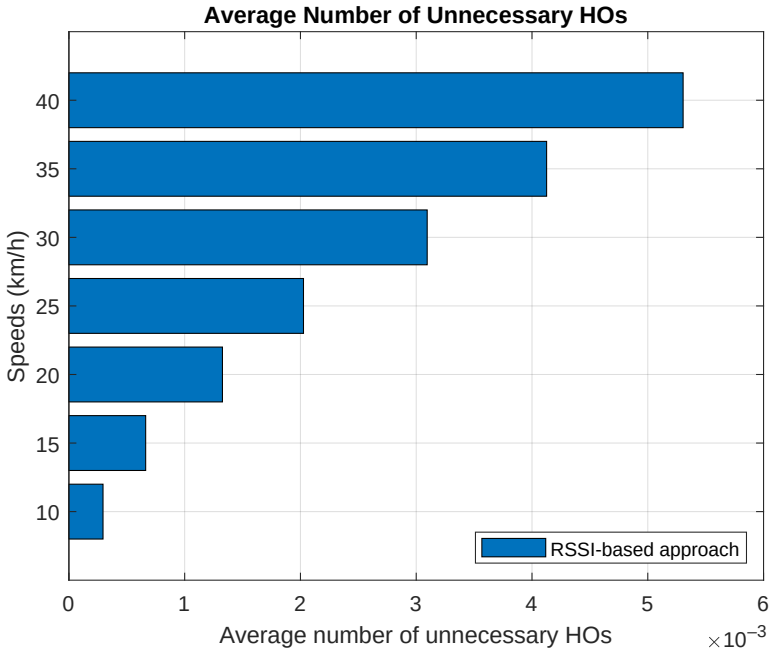

Figure 22. Average number of unnecessary HOs under various speeds.

5. Conclusions and Future Work

In this paper, we have proposed a real vehicle dataset, called VehDS-LA, that is designed for researchers and scientists in the field of V2X and machine learning. It is available on the GitHub website and it is characterized by its ability to take advantage of the power of Google Maps and MATLAB to produce a database with high location accuracy of vehicle samples. The vehicle samples are located on 15 streets in the city of Los Angeles. Each sample has four features; namely, latitude and longitude coordinates, speed, and azimuth. The total number of samples in the dataset is 74,170. The proposed dataset overcomes the limitations of related vehicle datasets in terms of generation time, vehicle location accuracy, effort savings, and the absence of requirements for special equipment and devices. The proposed dataset can be used as the basis for a new line of future research related to 5G networks, ML-based vehicle mobility applications, automation and driverless vehicles, ITS in the LA smart city, and SDN-based vehicular networks.

Author Contributions: I.A.A. collected the data, generated the dataset, analyzed the results, and wrote the paper. M.A.A. supervised the research and critically revised the paper. All authors have read and approved the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group No (RG-1440-122).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

A list of the abbreviations that are mentioned in this paper is given in following table.

Abbreviation	Meaning
3GPP	3rd-Generation Partnership Project
5G	Fifth Generation
AI	Artificial Intelligence
ANN	Artificial Neural Networks

BSs	Base Stations
CSV	Comma-Separated Values
GPS	Global Positioning System
HO	Handover
IoT	Internet of Things
IoV	Internet of Vehicles
ITS	Intelligent Transportation System
KMZ	Keyhole Markup Language
LA	Los Angeles
ML	Machine learning
NB	Naive Bayes
SAE	Society of Automotive Engineers
SDN	Software-Defined Networking
SVM	Support Vector Machine
UGL	Ultra GPS Logger
UTM	Universal Transverse Mercator
V2I	Vehicle-to-Infrastructure
V2N	Vehicle-to-Network
V2P	Vehicle-to-Pedestrian
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VehDS-LA	Vehicle Dataset in the city of LA

References

1. Hassan, N.; Yau, K.L.A.; Wu, C. Edge Computing in 5G: A Review. *IEEE Access* **2019**, *7*, 127276–127289. [CrossRef]
2. Alablani, I.A.; Arafah, M.A. Enhancing 5G small cell selection: A neural network and IoV-based approach. *Sensors* **2021**, *21*, 6361. [CrossRef] [PubMed]
3. Alablani, I.; Alenazi, M. Performance Evaluation of Sensor Deployment Strategies in WSNs Towards IoT. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 3–7 November 2019; pp. 1–8. [CrossRef]
4. Rehman, G.U.; Ghani, A.; Zubair, M.; Naqvi, S.H.A.; Singh, D.; Muhammad, S. IPS: Incentive and Punishment Scheme for Omitting Selfishness in the Internet of Vehicles (Iov). *IEEE Access* **2019**, *7*, 109026–109037. [CrossRef]
5. Fabian, P.; Rachedi, A.; Guéguen, C. Selection of relays based on the classification of mobility-type and localized network metrics in the Internet of Vehicles. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e4246. [CrossRef]
6. Chen, S.; Hu, J.; Shi, Y.; Peng, Y.; Fang, J.; Zhao, R.; Zhao, L. Vehicle-to-Everything (V2X) Services Supported by LTE-Based Systems and 5G. *IEEE Commun. Stand. Mag.* **2017**, *1*, 70–76. [CrossRef]
7. Raza, N.; Jabbar, S.; Han, J.; Han, K. Social vehicle-to-everything (V2X) communication model for intelligent transportation systems based on 5G scenario. In Proceedings of the 2nd International Conference on Future Networks and Distributed Systems, Amman, Jordan, 26–27 June 2018; pp. 1–8.
8. Sirohi, D.; Kumar, N.; Rana, P.S. Convolutional neural networks for 5G-enabled intelligent transportation system: A systematic review. *Comput. Commun.* **2020**, *153*, 459–498. [CrossRef]
9. Liu, H.; Lang, B. Machine learning and deep learning methods for intrusion detection systems: A survey. *Appl. Sci.* **2019**, *9*, 4396. [CrossRef]
10. Alablani, I.A.; Arafah, M.A. An SDN/ML-Based Adaptive Cell Selection Approach for HetNets: A Real-World Case Study in London, UK. *IEEE Access* **2021**, *9*, 166932–166950. [CrossRef]
11. Ziolkowski, P.; Niedostatkiewicz, M. Machine learning techniques in concrete mix design. *Materials* **2019**, *12*, 1256. [CrossRef]
12. Alzahrani, T.; Al-Bander, B.; Al-Nuaimy, W. Deep Learning Models for Automatic Makeup Detection. *AI* **2021**, *2*, 497–511. [CrossRef]
13. Modi, B.; Jethva, H. Reinforcement Learning with Neural Networks: A Survey. In *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 467–475.
14. Li, J.; Horiguchi, Y.; Sawaragi, T. Counterfactual inference to predict causal knowledge graph for relational transfer learning by assimilating expert knowledge–Relational feature transfer learning algorithm. *Adv. Eng. Inform.* **2022**, *51*, 101516. [CrossRef]
15. Alablani, I.; Alenazi, M. EDTD-SC: An IoT sensor deployment strategy for smart cities. *Sensors* **2020**, *20*, 7191. [CrossRef] [PubMed]
16. Shaaban, K.; Adalbi, M.A. Smart City Transportation System in Developing Countries: The Case of Lusail City, Qatar. In Proceedings of the International Conference on Applied Human Factors and Ergonomics, New York, NY, USA, 25–29 July 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 445–452.

17. Gohar, A.; Nencioni, G. The Role of 5G Technologies in a Smart City: The Case for Intelligent Transportation System. *Sustainability* **2021**, *13*, 5188. [CrossRef]
18. Silva, B.N.; Khan, M.; Han, K. Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. *Sustain. Cities Soc.* **2018**, *38*, 697–713. [CrossRef]
19. Ortega-Fernández, A.; Martín-Rojas, R.; García-Morales, V.J. Artificial intelligence in the urban environment: Smart cities as models for developing innovation and sustainability. *Sustainability* **2020**, *12*, 7860. [CrossRef]
20. Sharma, P.; Rajput, S. Perspectives of smart cities: Introduction and overview. In *Sustainable Smart Cities in India*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 1–13.
21. Ye, X.; Duan, L.; Peng, Q. Spatiotemporal Prediction of Theft Risk with Deep Inception-Residual Networks. *Smart Cities* **2021**, *4*, 204–216. [CrossRef]
22. Pincetl, S.; Graham, R.; Murphy, S.; Sivaraman, D. Analysis of high-resolution utility data for understanding energy use in urban systems: The case of Los Angeles, California. *J. Ind. Ecol.* **2016**, *20*, 166–178. [CrossRef]
23. Alablani, I.A.; Arafah, M.A. Applying a Dwell Time-Based 5G V2X Cell Selection Strategy in the City of Los Angeles, California. *IEEE Access* **2021**, *9*, 153909–153925. [CrossRef]
24. Lin, T.; Rodríguez, L.F.; Davis, S.; Khanna, M.; Shastri, Y.; Grift, T.; Long, S.; Ting, K. Biomass feedstock preprocessing and long-distance transportation logistics. *Gcb Bioenergy* **2016**, *8*, 160–170. [CrossRef]
25. Talen, E.; Anselin, L. City cents: Tracking the spatial imprint of urban public expenditures. *Cities* **2021**, *108*, 102962. [CrossRef]
26. Jensen, C.; Lahrmann, H.; Pakalnis, S.; Runge, J. The Infati Data. *arXiv* **2004**, arXiv:cs/0410001. Available online: https://arxiv.org/abs/cs/0410001 (accessed on 3 January 2022).
27. Alzyout, M.S.; Alsmirat, M.A. Performance of design options of automated ARIMA model construction for dynamic vehicle GPS location prediction. *Simul. Model. Pract. Theory* **2020**, *104*, 102148. [CrossRef]
28. Cho, W.; Kim, S.H. Multimedia Sensor Dataset for the Analysis of Vehicle Movement. In Proceedings of the 8th ACM on Multimedia Systems Conference, MMSys'17, Taipei, Taiwan, 20–23 June 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 175–180. [CrossRef]
29. Zhang, Y.; Xiong, L.; Yu, J. Deep Learning Based User Association in Heterogeneous Wireless Networks. *IEEE Access* **2020**, *8*, 197439–197447. [CrossRef]
30. Alam, M.J.; El-Saleh, A.A.; Tan, C.K.; Ku, I.; Lee, Y.L.; Chuah, T.C. Improved Joint Cell Association and Interference Mitigation for LTE-A Heterogeneous Networks. In Proceedings of the 2018 IEEE 4th International Symposium on Telecommunication Technologies (ISTT), Bangi, Malaysia, 26–28 November 2018; pp. 1–4. [CrossRef]
31. Alablani, I.A.; Arafah, M.A. An Adaptive Cell Selection Scheme for 5G Heterogeneous Ultra-Dense Networks. *IEEE Access* **2021**, *9*, 64224–64240. [CrossRef]
32. Ezhilarasi, S.; Bhuvaneswari, P. Modified RRP scheme for interference management in OFDMA based heterogeneous networks. *Wirel. Netw.* **2021**, *27*, 5105–5124. [CrossRef]
33. van Hoek, R.; Ploeg, J.; Nijmeijer, H. Cooperative Driving of Automated Vehicles Using B-Splines for Trajectory Planning. *IEEE Trans. Intell. Veh.* **2021**, *6*, 594–604. [CrossRef]
34. Schneble, C.O.; Shaw, D.M. Driver's views on driverless vehicles: Public perspectives on defining and using autonomous cars. *Transp. Res. Interdiscip. Perspect.* **2021**, *11*, 100446. [CrossRef]
35. Akyuz, E.; Cicek, K.; Celik, M. A Comparative research of machine learning impact to future of maritime transportation. *Procedia Comput. Sci.* **2019**, *158*, 275–280. [CrossRef]
36. Wang, J.; Zhu, H.; Liu, J.; Li, H.; Han, Y.; Zhou, R.; Zhang, Y. The application of computer vision to visual prosthesis. *Artif. Organs* **2021**, *45*, 1141–1154. [CrossRef]
37. Siddique, A.; Afanasyev, I. Deep Learning-based Trajectory Estimation of Vehicles in Crowded and Crossroad Scenarios. In Proceedings of the 2021 28th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 27–29 January 2021; pp. 413–423. [CrossRef]
38. Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J. Mater.* **2017**, *3*, 159–177. [CrossRef]
39. Escobar, C.A.; Chakraborty, D.; McGovern, M.; Macias, D.; Morales-Menendez, R. Quality 4.0—Green, Black and Master Black Belt Curricula. *Procedia Manuf.* **2021**, *53*, 748–759. [CrossRef]
40. Cao, R.; Sun, L. Design and Practice of Machine Learning Course Based on CDIO and Student Behavior Data. In Proceedings of the 2020 15th International Conference on Computer Science Education (ICCSE), Delft, The Netherlands, 18–22 August 2020; pp. 553–556. [CrossRef]
41. Raut, C.M.; Devane, S.R. Intelligent transportation system for smartcity using VANET. In Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP), Tamilnadu, India, 6–9 April 2017; pp. 1602–1605. [CrossRef]
42. Riaz, M.T.; Aaqib, S.M.; Ahmad, S.; Amin, S.; Ali, H.; Husnain, S.; Riaz, S. The Intelligent Transportation Systems with Advanced Technology of Sensor and Network. In Proceedings of the 2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, Pakistan, 26–27 October 2021; pp. 1–6. [CrossRef]
43. Guerrero-Ibáñez, J.; Zeadally, S.; Contreras-Castillo, J. Sensor technologies for intelligent transportation systems. *Sensors* **2018**, *18*, 1212. [CrossRef] [PubMed]
44. Alsaeedi, M.; Mohamad, M.M.; Al-Roubaiey, A.A. Toward adaptive and scalable OpenFlow-SDN flow control: A survey. *IEEE Access* **2019**, *7*, 107346–107379. [CrossRef]

45. Mekki, T.; Jabri, I.; Rachedi, A.; Chaari, L. Software-defined networking in vehicular networks: A survey. *Trans. Emerg. Telecommun. Technol.* **2021**, e4265. [CrossRef]
46. Gupta, V.; Kaur, K.; Kaur, S. Network programmability using software defined networking. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 1170–1173.
47. Hussein, A.; Elhajj, I.H.; Chehab, A.; Kayssi, A. SDN security plane: An architecture for resilient security services. In Proceedings of the 2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW), Berlin, Germany, 4–8 April 2016; pp. 54–59.
48. Natanzi, S.B.H.; Majma, M.R. Secure northbound interface for SDN applications with NTRU public key infrastructure. In Proceedings of the 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, Iran, 22 December 2017; pp. 0452–0458. [CrossRef]
49. Nkosi, M.; Lysko, A.; Ravhuanzwo, L.; Nandeni, T.; Engelberencht, A. Classification of SDN distributed controller approaches: A brief overview. In Proceedings of the 2016 International Conference on Advances in Computing and Communication Engineering (ICACCE), Durban, South Africa, 28–29 November 2016; pp. 342–344. [CrossRef]
50. Susilo, B.; Sari, R.F. Intrusion Detection in Software Defined Network Using Deep Learning Approach. In Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 27–30 January 2021; pp. 0807–0812. [CrossRef]
51. de Almeida Amazonas, J.R.; Santos-Boada, G.; Ricciardi, S.; Solé-Pareta, J. Technical challenges and deployment perspectives of SDN based elastic optical networks. In Proceedings of the 2016 IEEE 18th International Conference on Transparent Optical Networks (ICTON), Trento, Italy, 10–14 July 2016; pp. 1–5.
52. Github.com. *Vehicle-Dataset-in-Los-Angeles*. Available online: https://github.com/Ibtihal-Alablani/Vehicle-Dataset-in-Los-Angeles (accessed on 7 January 2022).
53. data.lacity.org. *Small Cell Locations*. Available online: https://data.lacity.org/City-Infrastructure-Service-Requests/Small-Cell-Locations/3nrm-mq6k (accessed on 22 February 2022).
54. V16.1.0, G.T.; 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Channel Model for Frequencies from 0.5 to 100 GHz (Release 16). 3GPP. 2019. Available online: https://www.3gpp.org/ftp//Specs/archive/38_series/38.901/38901-g10.zip (accessed on 7 January 2022).
55. Alhammadi, A.; Roslee, M.; Alias, M.Y.; Shayea, I.; Alquhali, A. Velocity-aware handover self-optimization management for next generation networks. *Appl. Sci.* **2020**, *10*, 1354. [CrossRef]
56. Ahmad, R.; Sundararajan, E.A.; Othman, N.E.; Ismail, M. Efficient handover in LTE-A by using mobility pattern history and user trajectory prediction. *Arab. J. Sci. Eng.* **2018**, *43*, 2995–3009. [CrossRef]