Discover Internet of Things

Research

Detection of road traffic anomalies based on computational data science

Jamal Raiyn[1]

Received: 8 December 2021 / Accepted: 11 February 2022
Published online: 20 October 2022

Abstract

The development of 5G has enabled the autonomous vehicles (AVs) to have full control over all functions. The AV acts autonomously and collects travel data based on various smart devices and sensors, with the goal of enabling it to operate under its own power. However, the collected data is affected by several sources that degrade the forecasting accuracy. To manage large amounts of traffic data in different formats, a computational data science approach (CDS) is proposed. The computational data science scheme introduced to detect anomalies in traffic data that negatively affect traffic efficiency. The combination of data science and advanced artificial intelligence techniques, such as deep leaning provides higher degree of data anomalies detection which leads to reduce traffic congestion and vehicular queuing. The main contribution of the CDS approach is summarized in detection of the factors that caused data anomalies early to avoid long-term traffic congestions. Moreover, CDS indicated a promoting results in various road traffic scenarios.

Keywords Computational data science · Deep learning · Autonomous vehicle · Traffic anomaly · Data analysis

1 Introduction

The development of a new generation of mobile telecommunication (5G) has led to AVs with new intelligent properties [1]. Fin the future, the AV will be able to collect data independently [30, 31, 33] and carry out data analysis based on modern artificial intelligence methodology. A verification process can be performed in which different types of collected data are checked for accuracy, reliability, integrity, availability and inconsistencies [53].

Datasets of input pertaining to road traffic can be affected input by geographical and human sources [28]. In this research, the datasets affected by geographical factors are considered. Some roads data arrive in an autonomous vehicle (AV) incomplete or falsified [20] and can result in abnormal conditions on urban roads [45]. A deep learning model is suggested here to manage the dataset input [2, 3, 5]. Deep learning (DL) methods have been applied in various research fields, due to the great capability of deep networks [9–14]. To train such networks, a very large training dataset is needed with a copious amount of data of various kinds, in order to develop effective models with several different parameters [1, 19, 22, 23].

The focus of this paper is on the detection and classification of road traffic anomalies. In regard to the identification of elements that characterize the road conditions, besides other surface anomalies, environment and health hazards have been identified [20]. The traditional methods for surveying the road conditions involve either the use of specialized vehicles equipped with laser sensors or manual surveys of the roads. However, these methods are very time consuming

✉ Jamal Raiyn, raiyn@qsm.ac.il | [1]Computer Science Department, Al Qasemi Academic College, Baqa Al Gharbiah, Israel.

and expensive, especially when processing large networks of roads in big cities [16]. As an alternative, the collaborative mobile sensing has been proposed. It is discussed as a promising mechanism for collecting large-scale real-world data [8].

Road anomaly detection and road condition survey systems based on collaborative mobile sensing typically detect and automatically classify road anomalies by applying data-mining approaches on data collected by smartphones [20, 25]. This is a challenging task, especially when the data are collected in real-world deployments. This work was carried out in the context of a joint industry/academia research and innovation project that aimed at building a short-term travel forecast scheme and detecting autonomous vehicle positioning [15]. The objective was to evaluate a road anomaly detection system based on a machine learning approach described in a previous work by the author [53]; a real world deployment was used i.e., an environment in which data were collected during normal driving which can reveal situations that were not anticipated during the system design process. The dataset was analyzed based on a data science life cycle concept that comprises the following phases: data collection, data preparation, data exploration, modeling and model evaluation and deployment [56]. The travel data used in this research was collected with smartphones. These raw data were then processed using various techniques to extract relevant features and machine-learning algorithms were applied to recognize road anomalies. The novel element in this research was the use of a computational data science approach that possesses intelligent features based on goal oriented agents, such as autonomous, adaptive, mobile, and cooperative agents. A goal-oriented agent system fulfills designed functions, such as detection of data anomalies, identification of their sources and the localization of the traffic data anomalies based on statistical analysis. The proposed computational data science combines artificial intelligence and data were evaluated and assigned a score based on the quality life cycle.

This paper is organized as follows: "Related work" section gives an overview of work related to AVs and data anomalies. "Computational data science approach (CDS)" section introduces the computational data science approach. "Computational artificial intelligence" section discusses the development of the term computational artificial intelligence. "Methodology" section describes the methodology and presents the simulations and results. Finally, "Conclusion and future work" section concludes the discussion, summarizing the work and pointing to directions for future research.

2 Related work

The field of anomaly detection has been widely researched [6, 18, 24, 27]. Anomaly detection approaches are typically divided into two types: model-based and data analysis-based. Model-based approaches mostly use algorithms that are very accurate, such as machine learning schemes [32], whereas the approaches based on data analysis usually used statistical measurements. On urban roads, anomalies cause discomfort to drivers and have a negative impact on traffic efficiency [6]. When traffic accidents occur and there is congestion, traffic flow is abnormal [26]. In AV network, the anomalies are caused by traffic accidents, bad weather, road work, and repeated lane changing attempts [19]. In addition, there are a number of other challenges [18] faced by AVs, such as noise and interference, which are further sources of anomalies in traffic flow. AVs collect various types of data via onboard devices and communication with devices on the Internet of Things [17, 35, 37]. Both the onboard devices and the AV communications protocols are affected by interferences and delays [41, 46]. The AV positioning data is collected by the GNSS [21], whose satellites are mainly located in medium earth orbits (MEO). The signals transmitted by a satellite propagates through the atmosphere, where they are subject to delays caused by ionospheric and tropospheric media. At ground level, multipath effects, namely the reception of signals that are reflected from obstacles such as buildings surrounding the receiver, can occur, causing one of the largest types of errors, one that is also difficult to model, as it depends strongly dependent on the receiver environment. Increased delays may affect the performance metrics of a positioning terminal, which are characterized in terms of availability, accuracy, and integrity. Delays may be caused by the weak performance of network equipment and the heterogeneity in equipment attributes. Many untargeted transmitted signals can even interfere with transmitted signals [38, 39]. The proposed solution is based on maintaining a minimal distance between AVs [44]. Dey et al. (2016) [4] used Het-Net to support connected vehicle applications based on vehicle-to-vehicle (V2V) and vehicle-to-Infrastructure (V2I) communications. Sepulcre and Gozalvez (2018) [36] presented an architecture for context-aware heterogeneous V2I communication in vehicular networks to improve the quality of service and satisfy vehicular application requirements. Brandl (2016) [7] introduced an initial proof of concept for future connected vehicular landscapes, focusing on the basics of "vehicle-to-everything" (V2X) communication, that is, from vehicle to vehicle and from vehicle to infrastructure. Weiss (2011) [38] introduced a field operational test simTD, which is the first of its kind to evaluate the effectiveness and benefits of applications based on vehicular communication. Jin et al. (2013) [39] proposed an improved multi-agent intersection management system, in which vehicle agents may form platoons using connected vehicles technologies. Compared to the conventional traffic signal control system, the proposed platoon-based multi-agent intersection management system can shorten average travel time and reduce fuel consumption. Gora and Rüb [40] proposed self-driving and connected vehicles that communicate with one another (through V2V technology) and with the road infrastructure (through V2I technology), and they designed a microscopic traffic simulation model for such vehicles, including a robust protocol for exchanging information. Li (2016) [41] proposed a cooperative traffic control algorithm based on vehicle-to Infrastructure (V2I) connections to reduce traffic delays and decrease fuel consumption. Bergenhem (2012) [42] described a V2V communication system that enables vehicles to drive in platoons. Jiang et al. (2010) [48] presented a practical model for characterizing V2V communication channel and the impact of inter-carrier interference (ICI) generated by orthogonal frequency division multiplexing (OFDM). Sun et al. (2016) [49] investigated the radio resource management problem for D2D-based V2V communication based on device-to-device (D2D) systems and proposed direct links as a possible enabler for V2V communications, where incurred intracellular interference and the stringent latency and reliability requirements are challenging issues. Du and Dao (2015) [43] proposed an analytical formulations for estimating delays in information propagation time delay via a V2V communication network serving a one-way or two-way road segment with multiple lanes. This technical view of platooning describes inter platooning interactions based on V2X communication and examines the maintaining of a platoon, for instance, while AVs are joining, leaving, or changing lanes. Furthermore, informative speed assistance and AV positioning estimation are essential for platoon control. Santhosh et al. [57–60] introduced the visual surveillance to detect data anomalies in road traffic.

Now there are some new challenges that have yet to be considered, such as communication between AVs in heterogeneous wireless networks. Heterogeneity can cause delays in communication between AVs. For AV communication in differentiated wireless networks that provide different quality of services (QoS), middleware is needed for adaptation. Another challenge faced by AVs is identifying road sections with a high degree of noise. To detect road traffic anomalies, various forecasting schemes have been proposed [27, 34, 48, 49]; recently, some deep learning approaches have been introduced to predict the urban traffic flow [24, 47]. This study takes a computational data science approach and introduces a deep learning scheme to detect interference and other sources of anomalies in road traffic. For each road section, the degree of anomaly is calculated based on various measurement of statistical error.

3 Computational data science approach (CDS)

A *computational data science* approach was used in this study to characterize anomalies in road travel data. Computational data science methodologies combine the concept of the data science lifecycle with advances in artificial intelligence methodology.

3.1 Data science lifecycle

The data science lifecycle involves several steps that constitute an analytical methods based on artificial intelligence methodologies. The lifecycle steps are data collection, data preparation, data exploration, and data validation,

3.1.1 Data collection

There is a wide range of technologies available to AVs for data collection. In contrast to traditional methods of collecting traffic data, which include human surveyors on site and the use of magnetic sensors, AVs use modern technologies to overcome the deficiencies of traditional methods. This paper introduces a method for data collection based on mobile services [26]. Figure 1 illustrates a comparison of data collection by mobile services with data collection conducted by magnetics sensors.

3.1.2 Data preparation

AVs receives data from various devices in different formats, and this can lead to problems with data quality. In the preparation phase, data are converted to a desired format, then the dataset is cleaned, and inconsistent, invalid, and corrupt data are removed.

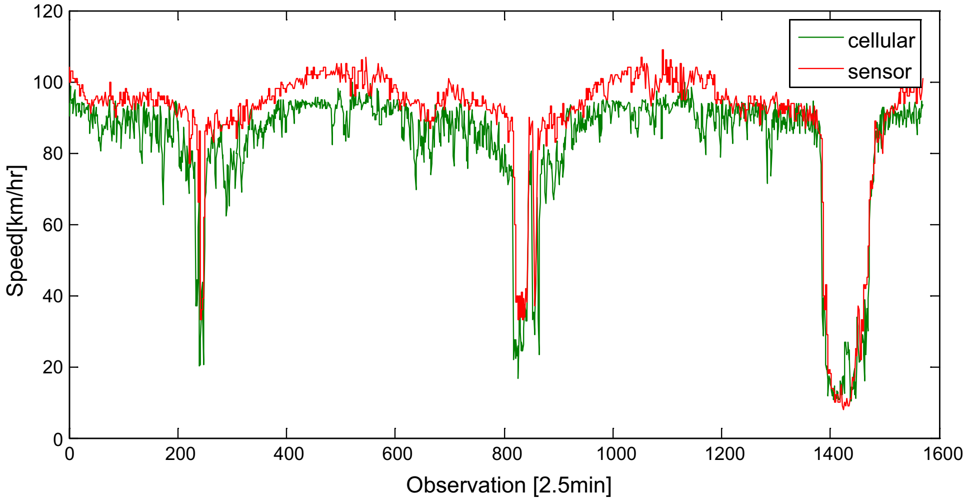


Fig. 1 Mobile service data versus sensor data

3.1.3 Data availability

The collected travel dataset consisted of statistical measurements; the analysis showed that there were missing data (see Fig. 2).

3.1.4 Data exploration

In this study, data exploration was aimed at understanding the behavior of the traffic flow on urban roads. Plotting is a tool that can be used to reveal hidden patterns within a dataset. Moreover, the multiple statistical measures, such as the mean and the standard deviation of a variable, and its interactions with other features, help to distinguish between normal and abnormal traffic flow and to explain the reasons for any differences. In an abnormal situation, the traffic load *(tt(t, k))* starts to decrease, and when the situation resolves, the travel traffic load starts to increase progressively. Furthermore, as illustrates Fig. 3 the standard deviation value (σ) changes when abnormal events occur. The record is considered to be abnormal when the traffic speed decreases to at least 30 km/h slower than the average speed of all

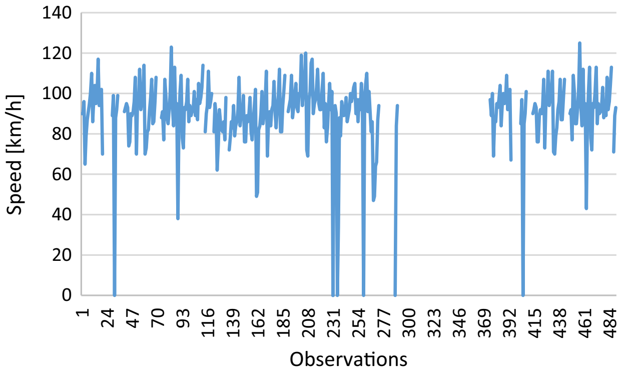


Fig. 2 Data collection based on magnetic sensors

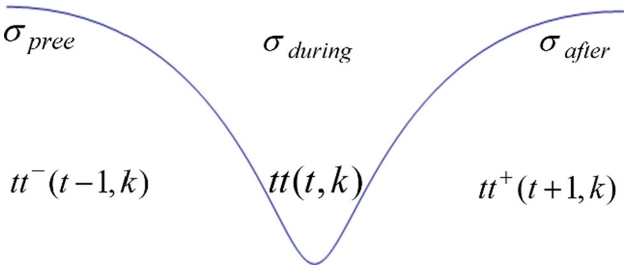


Fig. 3 Standard deviations of traffic flow before and after an abnormal condition

records for at the same time on the same day of the week. This threshold of 30 km/h is a symbolic value; it is the smallest speed change that people would consider "abnormal". The determination of when the threshold is reached depends on the travel observation data.

3.2 Mathematical description of an abnormal condition

Because the dataset for accidents included only contained information for the hour in which the abnormal event took place, but not the minute, it became necessary to plot the speed value of every accident in order to determine the actual time of the event [3]. The information was received from the nearest sensors closest to the event [15]. The speed of vehicles upstream from the event starts to decrease, while the speed downstream starts to increase as illustrates Fig. 4.

$$tt(k,t) - tt(k+1,t) \geq threshold = (K1) \tag{1}$$

When an incident occurs between stations *k and k + 1, the congestion causes a clear difference between* the occupants of the upstream and the downstream stations.

$$\frac{tt(k,t) - tt(k+1,t)}{tt(k,t)} \geq threshold = (K2) \tag{2}$$

This test checks for differences in speed in relation to the upstream speed. When this relative difference is larger than the threshold value *K2, a "bottleneck condition"* is detected. Not only is this speed difference large, it is greater than the upstream speed.

$$\frac{tt(k+1,t-1) - tt(k+1,t)}{tt(k+1,t-1)} \leq threshold = (K3) \tag{3}$$

The time test checks for an increase in the downstream speed over the last *t-1* min. This increase must be larger than a certain fraction of the previous speed. This test tells if the road in the downstream direction is becoming empty. The temporal variation of speed at a fixed location expressed as the coefficient of variation in speed [4, 18] (i.e., the standard deviation of the speed divided by the average speed) is given by the following equation:

$$CVS = \frac{1}{n}\sum_{i}^{n}\frac{\sigma_i}{\mu_i} \tag{4}$$

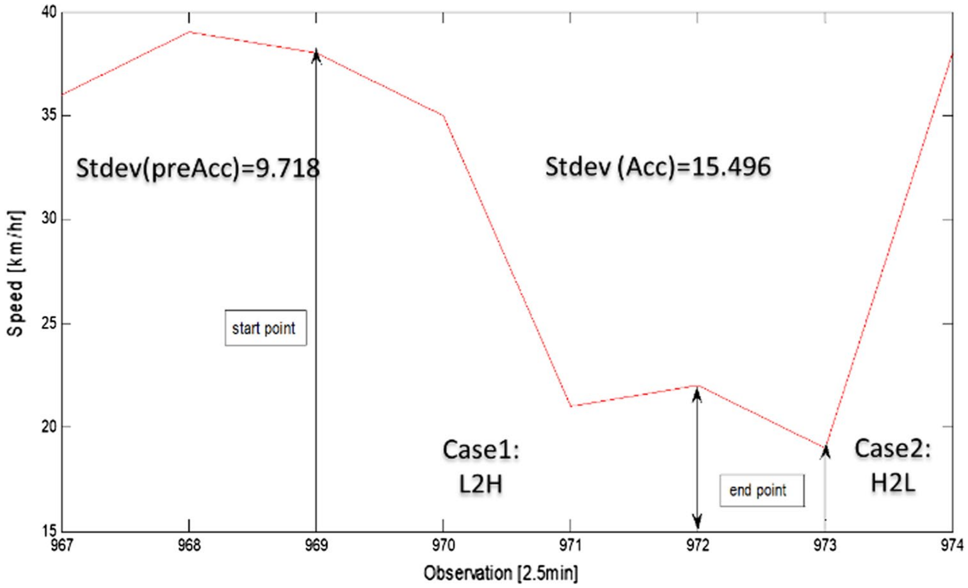


Fig. 4 Calculated standard deviation

The spatial variation in speed along road sections is expressed as the difference in speed between upstream and downstream.

$$tt_{upstream}(t,k) - tt_{downloadstream}(t,k+1) \tag{5}$$

A mathematical description of the happened abnormal condition is based on the behavior of the traffic on urban roads [22]. The abnormal detection algorithm combines historical travel observations of abnormal travel conditions with real travel data. Equation (6) describes the downstream traffic flow and Eq. (7) describes the travel traffic in upstream flow. Table 1 includes a description of abbreviations and acronyms used in this paper.

$$tt^{F}_{accident}(t+1,k) = EMA^{H}_{accident}(t+1,k) + \delta\left(tt^{M}_{accident}(t,k) - tt^{H}_{accident}(t,k)\right) \tag{6}$$

$$tt^{F}_{accident}(t+1,k) = EMA^{H}(t+1,k) + \delta\left(tt^{M}_{accident}(t,k) - tt^{H}(t,k)\right) \tag{7}$$

3.3 Reliability of travel data

Unreliable travel data may cause congestion, which is a traffic condition characterized by slower speeds, longer travel times, and the occurrence of AV communication interruption. High levels of congestion increase the likelihood of unreliability. If roads are highly congested, the AV communications is interrupted and the delay is increased in V2V communication. Therefore, a lack of reliability in travel time is associated with delays caused by congestion. Travel time reliability is calculated from travel time data, which needs to be of good quality. It is calculated from the time a trip starts to the time when the destination is reached. Methods for measuring travel time include equipping vehicles with a global positioning system (GPS) and devices for cellular communication, as illustrated in Fig. 5. Travel time reliability is influenced by additional factors, such as the involvement of a variety of communication devices. To improve travel time reliability, a computational data science approach is proposed here, which detects and manages the incidents that cause traffic disruptions.

4 Computational artificial intelligence

The term *artificial intelligence* (AI) refers to any human-like intelligence exhibited by a computer, robot, or other machine. As for the term *software agent*, there is no definition that I accepted world-wide, and a clearly definition seems impossible. Franklin and Graesser define an *autonomous agent* as an agent that senses its environment and acts on it. Wooldrige and Jenning mention [54] *reactive agents* that perceive their environment and response in a timely fashion to changes that

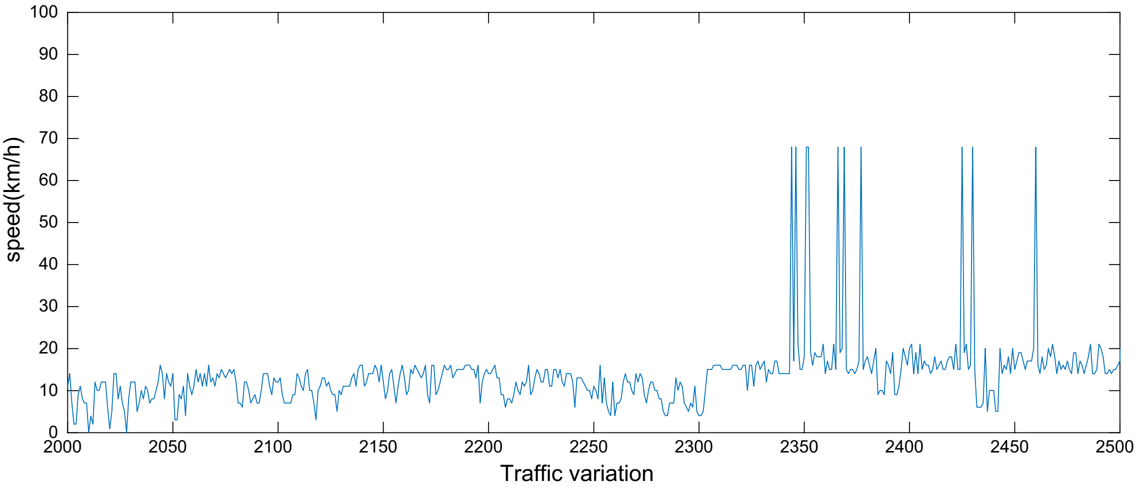


Fig. 5 Variation in travel time

occur in the environment, in other words, agents whose behavior is determined through a reasoning process based on beliefs and desires.

Russel and Norvig [55] propose that an agent is something that perceives and acts rationally to achieves its goals, based on both its own experience and the built-in knowledge used in contracting the agent for the particular environment in which it operates. However, there are many definitions, many of which are based on the functions and behaviors of the agent under the domain consideration, which means, an agent can be defined operationally in term of the environment in which it provides its services. An agent is supposed to have the following major characteristics such as autonomy, which means, it can take the initiative and exercises control over its actions, such as managing negotiations with human or other agents to update and to improve the basic rules. Bases on reasoning strategies, the agent has the ability to make a decisions and to draw conclusions. In general, *artificial intelligence* refers to the ability of a computer or machine to mimic the capabilities of the human mind. There is a relationship among the meanings of *artificial intelligence* (AI), *machine learning* (ML), and *deep learning* (DL), which is as follows: *Artificial intelligence* fields encompass anything relating to expert systems that makes decisions based on complex rules. *Machine learning* is a subset of AI application, involving systems that learn on their own; they actually re-program themselves as they digest more data to perform with increasingly greater accuracy the specific tasks they were designed for *Deep learning technologies* refers to a subset of machine- learning applications that teach themselves to perform a specific task with increasingly greater accuracy and without human intervention.

4.1 Deep learning technology

DL methods are one of the most useful tools in the area of machine learning [32]. The DL process inputs information in a hierarchical way, where each successive level of processing involves more abstract global and invariant features [52]. DL learns the features of a dataset and then combines them to achieve a specific goal.

The DL method is used to find solutions to complex problems, in this case, problems with intelligent transportation systems [50, 51]. In this paper, a DL method is proposed for the early detection of traffic anomalies. It is composed of two main phases: training and testing. DL methods are one of the most useful tools in the area of machine learning [32]. The DL process inputs information in a hierarchical way, where each successive level of processing involves more abstract global and invariant features [52]. DL learns the features of data and then combines them to achieve a specified goal. The system proposed here is composed of four phases: (1) preparation of the dataset, (2) a training phase, (3) a testing phase, and (4) a performance metrics phase. The dataset on travel speeds is obtained through smart phone services. The data pass through phases of preprocessing, such as cleaning and recovering missing values. In the training phase, features of the dataset extracted from the preprocessing phase are trained with DL.

4.2 Architecture of the DL concept

The DL, in general, is divided into three main parts, an input layer, a hidden layer, and an output layer [48].

4.3 The input layer

The input layer for DL consists of a large amount of data that is received from different sources. The big dataset for traffic modeling is diverse and comes from variety of sources, such as cameras, LIDAR, sensors, and GNSS. The devices installed in AVs provide near-real-time data and historical traffic data.

4.4 The hidden layer

The hidden layer is responsible for processing the input data. It processes the attributes of the dataset and extracts useful information to construct new attributes that will be used as input for the DL model [52]. Each layer within the hidden layer is assigned rules that are focused on input data attributes, and these are updated in keeping with new data input. The size of the hidden layer is expressed in term of the number of neurons there. The neurons have an important influence on the learning ability of the algorithm; too few can lead to insufficient learning, and too many can lead to overfitting.

4.5 The output layer

The output layer is responsible for exporting the values, or the vectors of the values, that correspond to the format required for the problem [10], and it presents the visually results based on measurements of statistical error.

4.6 A DL method for road traffic flow

A *layer* of the network includes a combination of weights and multiplication and summing operations, Observable characteristics in the data, which can be input into a model, are called *features*, and a model's ability to perform well always depends on finding features that represent the data well. The learning algorithm for training constitutes the main part of DL. The number of layers differentiates a deep learning network from a shallow one. The greater the number of layers, the deeper is the network. Each layer can be specialized to detect a specific aspect or feature.

$$tt^{DL}(t,k) = tt_{normal}(k,t) + tt_{abnormal}(k,t) \tag{8}$$

4.7 Travel data on anomalies for connected AVs

On urban roads, traffic anomalies cause discomfort to drivers and degrade traffic efficiency. A road traffic anomaly is an event, such as accident, that results in a road section that is a typical in terms of traffic flow and congestion. In an AV network, anomalies are caused by traffic accidents, bad weather, road work, and failed lane- changing. A road traffic anomaly is a road section that is anomalous in terms of traffic flow. The source of road traffic anomalies is mostly accidents, weather, and road work. The detection of road traffic anomalies has been extensively studied, and many traditional tools have been applied, such as video surveillance, sensor networks and the RFID technique. The degree of anomaly in each road segment is calculated based on measurements of statistical error. The model is a releasing node for an early warning system for abnormal events in road traffic. To evaluate the influence of the anomalies on road traffic management, various scenarios have been considered.

4.8 The lane- changing process

A successful lane- changing process increases the efficiency of traffic flow. However, in some cases, if the lane- changing process is interrupted, anomalies result and increase congestion. To manage the lane- changing process, an AV that needs to communicate with other vehicles must be at a minimal distance from them. The positioning of each AV is estimated based on GNSS data and cooperation among vehicles must take place, within the coverage area. In other words, the vehicles should maintain a minimal distance and should remain within transmission range in the same zone. In V2X communication [29], quality of service (QoS) parameters should be controlled, especially in regard to delay. Delay is the second most prevalent cause of interference and leads to increased traffic congestion. The lane- changing process in platooning can be divided into three phases: informative speed assistance (ISA), AV positioning estimation, and cooperative communication based message exchange [34, 35]. It is worth mentioning that to hack the lane changing process, cyber attackers attempt to change the GNSS data, and increase delay to interrupt data transmission; however, cyber security is not considered in this paper.

4.9 Traffic accident detection

Accidents are a significant problem; they can cause traffic delays that last for hours. This problem has been widely considered, however, it is difficult to forecast traffic accidents caused by human factors early and in real-time. The short-term forecast model can be improved by considering variation in traffic flow caused by traffic accidents.

4.10 Noise in connected AVs

The quality of raw GNSS measurements (also called observables) is affected by several factors that originate from satellites, signal propagation and receivers. The signal transmitted by a satellite propagates through the atmosphere, where it is subject to delays caused by ionospheric and tropospheric media. At ground level, a multipath effect, namely, the multiple reception of signals that are reflected from obstacles such as buildings surrounding the receiver, can occur, causing one of the largest sorts of errors.

4.11 Urban road anomalies

Anomaly detection is defined as finding data that does not conform to the notions of normal behavior. The relevance of this issue to urban road travel comes from the need to act upon the discovery of the outliers. The proposed concept is based on the observation that a traffic anomaly can be detected by monitoring the changes in the behavior of individual AVs (e.g. in term of deceleration and lane changing). The proposed anomaly detection scheme can be divided into major stages as depicted in Fig. 6. The process starts with the feature extraction stage, which involves the conversion of the original traffic variables into features, and which contain all the essential information for the task of detection task. In this research, the feature extraction step is based on the use of a deep learning scheme. The DL architecture in this paper is composed of three layers: an input layer, a hidden layer, and an output layer. The first phase, the training phase, was based on raw travel dataset, which considered the attributes of time, road section and speed. The speed observations, based on mobile services, were made every 2.5 min based on mobile services. The optimal selection of neurons in the hidden layer was equal to the number of the urban road sections, each of which was 300 m long. In the second layer of the hidden layer the number of neurons was fewer.

5 Methodology

The process of anomaly detection faces many challenges. It is needed for discovering a patterns that do not conform to normal expected behavior in the data. The first step in the anomaly detection process is to define a normal traffic on a section of urban road and then to flag as anomalies any observations that are not fit with this normal pattern. The main challenge for anomaly detection in urban road traffic is to find these patterns. In comparison to traditional schemes, the CDS approaches can be applied to all types of normally trafficke roads.

5.1 The description of CDS

Many algorithms are proposed to detect the data anomalies in traffic road, their activities are directed toward influencing the data anomalies on road traffic which are expressed in traffic congestion. In other hand, the activities of CDS algorithm focused on data behavior which is influence by various factors. The cognitive data differentiates between the influence of internal factors, such as delay in vehicle –to-vehicle communication that are caused due to variation of requirement for QoS, and extern factors that influence the data such as cyber attackers, geographical factors, radio channel interference. The conventional algorithms, such ML mostly relay on providing a description of available structured traffic data, However, CDS are sensing the factors that caused anomalies and their locations.

The CDS steps are summarized as illustrates Fig. 7.

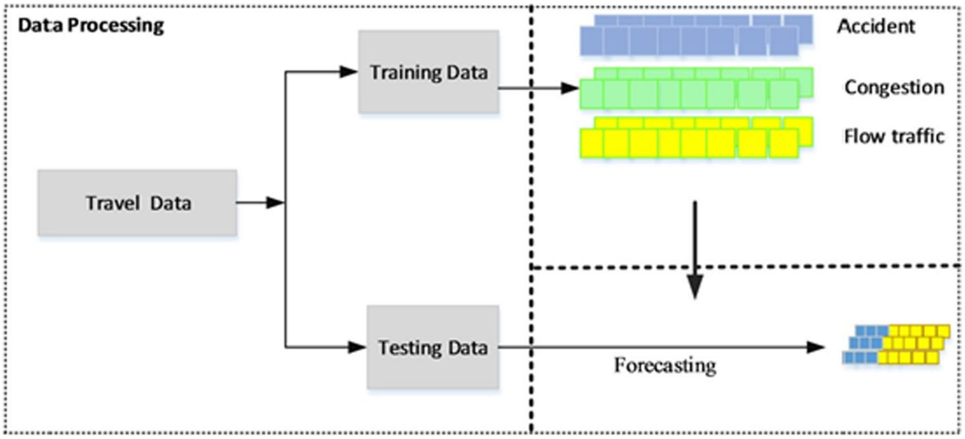


Fig. 6 Data processing strategy

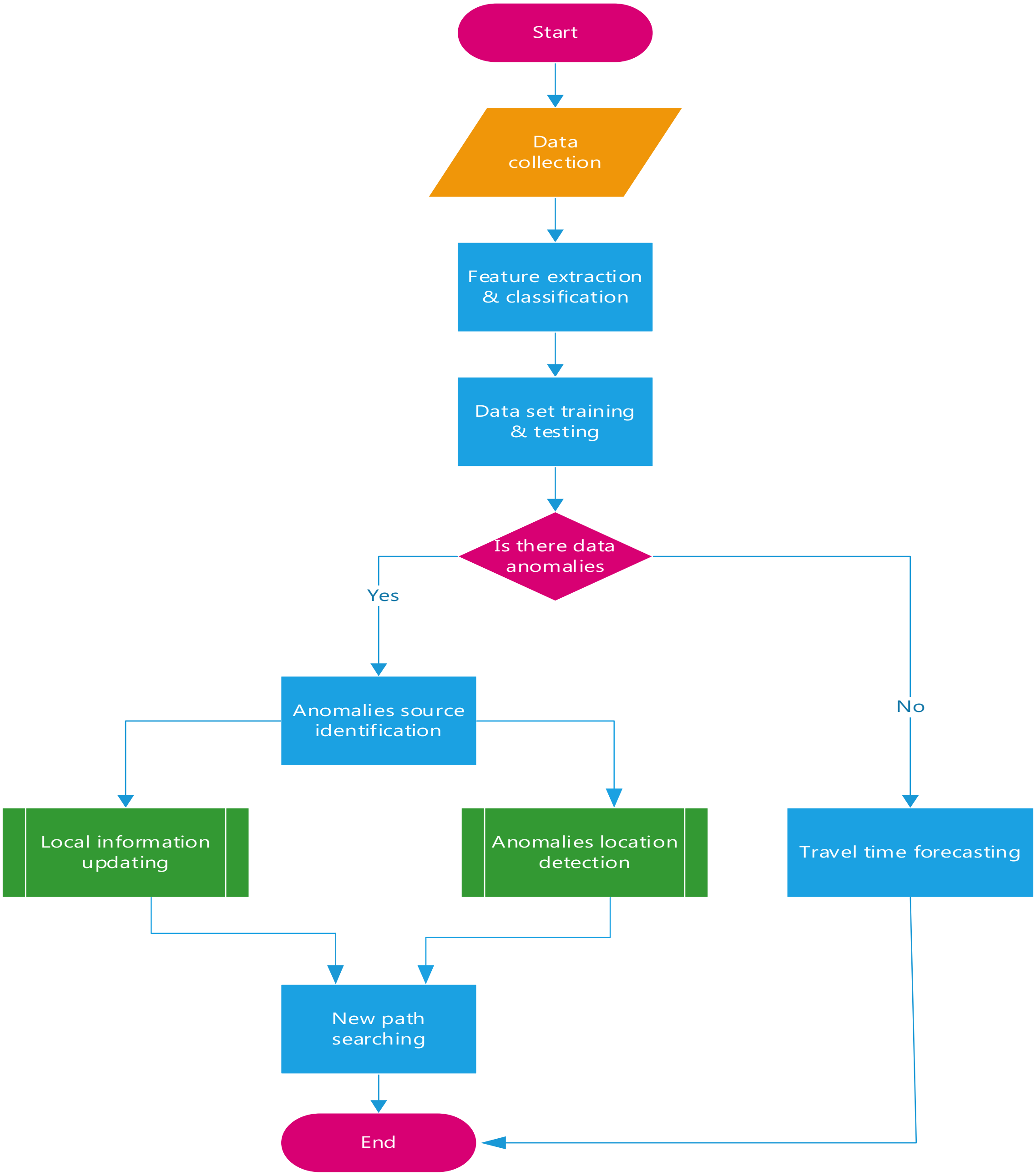

Start
Data collection
Feature extraction & classification
Data set training & testing
Is there data anomalies
Yes
No
Anomalies source identification
Local information updating
Anomalies location detection
Travel time forecasting
New path searching
End

Fig. 7 CDS algorithms

5.2 Data anomaly detection based on CDS

The travel data and AV positioning data were analyzed with various techniques. The most critical issue in collecting the data was finding the right input data set. The big data input went through several preparatory steps:

5.3 Recognition data

AVs collects various types of data from their installed devices. Classification tools are needed to identify the types of data sets. The data on vehicle positioning was collected from Ublox devices and smartphones as in Tables 2 and 3.

Table 1 Key mathematical symbols and abbreviation

Symbol	Definition
σ	Standard deviation value
μ	Mean value
tt^H	Historical travel time
tt^M	Current travel time
k	Road sections
EMA^H	Exponential moving average

Table 2 Unstructured GNSS data

```
$GNRMC,075035.00,A,4709.38620,N,00138.24936,W,0.004,,220916,,,D*72
$GNVTG,,T,,M,0.004,N,0.008,K,D*34
$GNGNS,075035.00,4709.38620,N,00138.24936,W,DA,18,0.63,30.1,48.1,,0000*4D
$GNGGA,075035.00,4709.38620,N,00138.24936,W,2,12,0.63,30.1,M,48.1,M,,0000*6B
$GNGSA,A,3,21,23,25,05,26,27,29,31,16,20,,,1.06,0.63,0.85*1E
$GNGSA,A,3,80,79,81,83,65,66,73,82,,,,,1.06,0.63,0.85*1A
$GPGSV,4,1,13,05,11,045,36,09,04,331,36,16,33,301,44,18,02,144,*79
$GPGSV,4,2,13,20,18,097,40,21,62,138,48,23,07,303,30,25,20,118,36*70
$GPGSV,4,3,13,26,64,304,49,27,13,254,43,29,39,061,47,31,47,209,50*79
$GNGRS,075035.00,1,–1.5,– 2.3,8.3,– 13.6,– 5.0,2.8,2.6,– 0.7,– 3.8,0.1,,*68
$GNGRS,075035.00,1,-6.3,14.8,-2.6,1.1,-6.3,8.8,-12.1,1.7,,,,*59
$GNGST,075035.00,28,,,,0.58,0.44,1.0*4B
$GNGBS,075035.00,0.6,0.4,1.0,73,,12.1,6.1*69
$PUBX,00,075035.00,4709.38620,N,00138.24936,W,78.277,D3,0.73,1.0,0.008,269.62,-0.002,,0.63,0.85,0.50,18,0,0*62
```

5.4 Structured vs. unstructured data

In general, AVs collect structured data. Vehicle positioning data is often represented by a matrix, whose columns represent distinct properties of these items in it. For instance, the set of positioning data input for the study contained one column for longitude and other columns for altitude. Some of the devices installed in AVs collect unstructured data as shown in Table 2, such as GNSS data, so our first step is to build a matrix to structure them.

5.5 Cleaning and formatting

An important step is cleaning and formatting the data. Travel data are collected by different devices and are influenced by human and environmental factors. *Data cleaning* is the process of modifying the data to ensure that the datasets are free of irrelevancies and incorrect or incomplete travel observations. The best computational data formats have several useful properties; for instance, they are easy for computers to parse, easy for people to read, and widely usable by other tools and systems. The travel data input for the study came from various smart devices in all kinds of formats. This dataset was cleaned to improve the efficiency of the data analysis and the quality of the results as shown in Table 2.

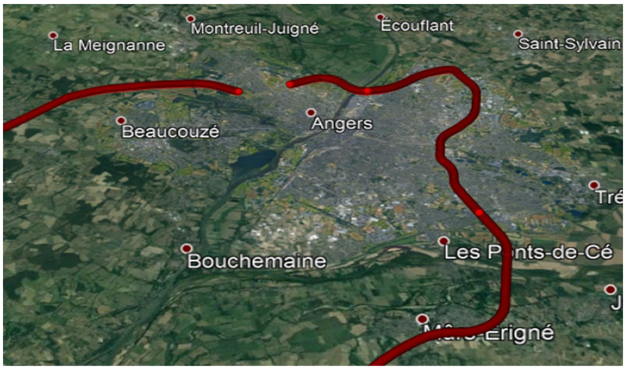


Fig. 8 Missing data

5.6 Visualizing data

An interactive exploratory travel dataset is useful for presenting errors in real-time. Due to noise and environmental factors that influence data communication, some of the datasets collected were not complete as illustrated in Fig. 8. An important aspect of data cleaning is identifying fields from which data are missing and then properly compensating for them.

5.7 Ranking of data

The historical data were scored and ranked with the use of functions that had been developed to detect lost data lost and data errors as illustrates Fig. 9. The road sections were scored according to the performance metrics of availability, accuracy, and integrity, which were characterized by the global performance of different devices.

5.8 Anomaly detection based DL

The deep learning scheme extracted the attributes from the inputs data. The attributes were classified and assigned scores.

$$y_{AD} = \sum x_n w_n + error \tag{9}$$

Here, x_n shows the input signal, w_n shows the weight corresponding to each input signal, and y_{AD} shows the output signal. F_{DL} is the activation function. the meaning of calculating the sum of data coming from the input. Figure 10 illustrates the deep learning layers.

5.9 Evaluation of anomaly detection schemes

In general, AVs collect a variety of traffic data. The raw traffic data used was mainly obtained from smart phones in 2.5 min cycles. When a statistical analysis of the original data was carried out, it was found that the original data contained two main defects; the dataset was incomplete, and the data contained noise.

Urban roads are divided into sections and each section is 300 m long. The data passed several steps, like the removal of unnecessary material such as noise [20].

Low quality data can lead to traffic congestion and collisions. Furthermore, data reception may be incomplete due to the urban noise produced by network tunnels as illustrated in Fig. 11. The detection of incomplete data is based on statistical measurements. Figure 12 illutsrates this appraoch, which acts like a radar that localizes the area affected by noise.

5.10 Data science analysis based ML schemes

In general, the linear regression is widely used for data analysis. In this research, the linear regression was applied to analyze the travel data based on mobile services. Furthermore, the linear regression and the Pearson linear correlation

Fig. 9 Positioning error identification

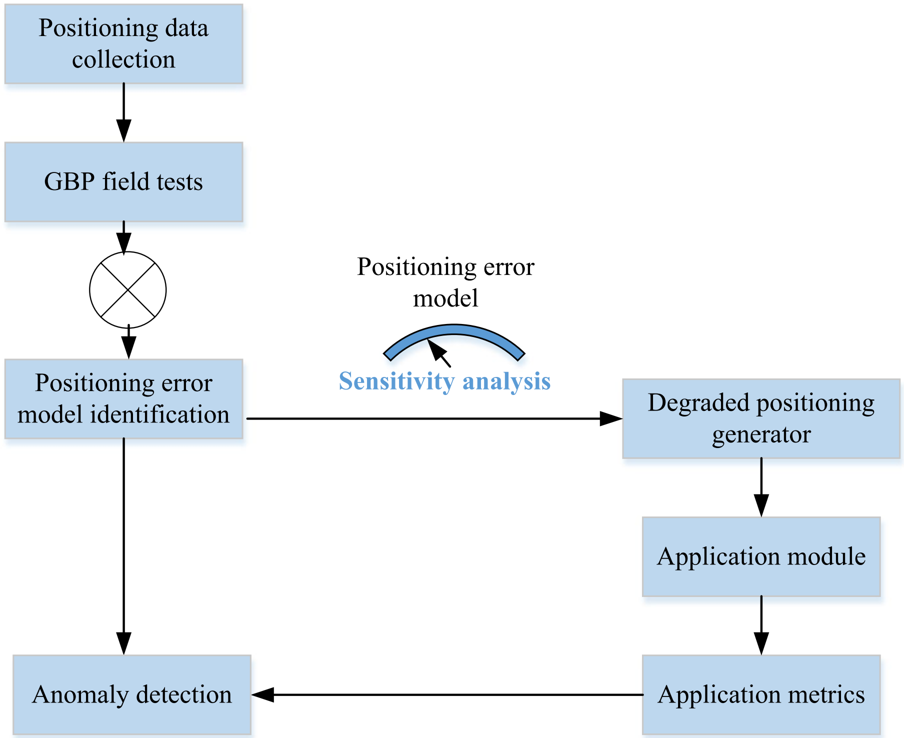
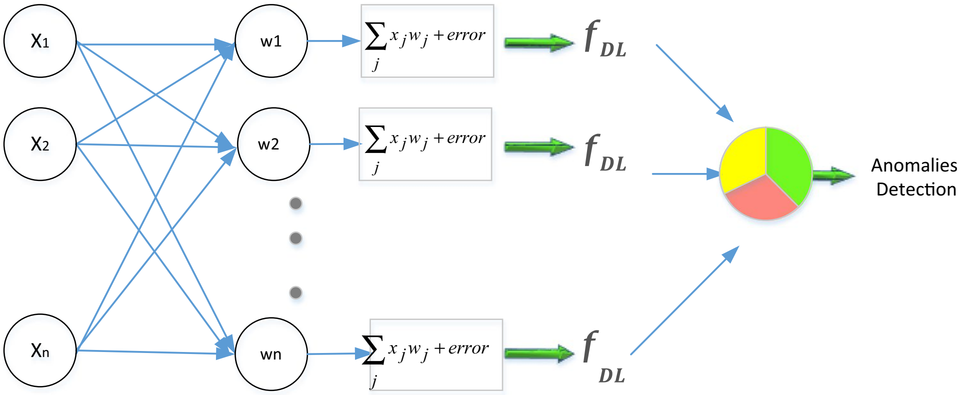


Fig. 10 DL for anomaly detection



are recommended for estimating and measuring the influence of the noise on travel data that is collected by mobile services. Table 4 presents a statistical description. The high value of the multiple indicates that the noise influences the speed and increases the congestions. Furthermore, the standard deviation of travel time becomes larger when the area is affected by noise.

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \tag{10}$$

where

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Fig. 11 Incomplete data

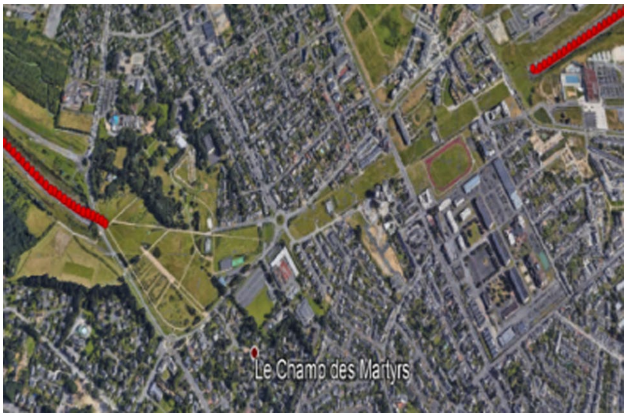


Fig. 12 Noise detection

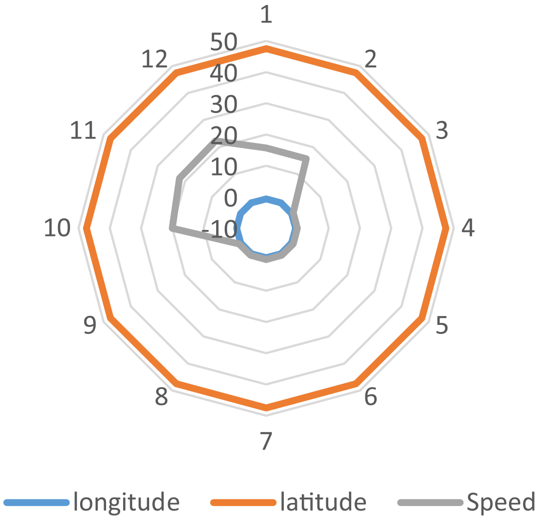


Table 3 Positioning data collection based on UBlox

Index	UTC	Lon	Lat	SoG	Speed	CoG
0	50:35.0	− 1.63749	47.15644	0	0	269.62
1	50:36.0	− 1.63749	47.15644	0	0.01	269.62
2	50:37.0	− 1.63749	47.15644	0	0	269.62
3	50:38.0	− 1.63749	47.15644	0.01	0.01	269.62
4	50:39.0	− 1.63749	47.15644	0	0.01	269.62
5	50:40.0	− 1.63749	47.15643	0.01	0.01	269.62
6	50:41.0	− 1.63748	47.15643	0.01	0.03	269.62
7	50:42.0	− 1.63748	47.15643	0	0.02	269.62
8	50:43.0	− 1.63748	47.15643	0	0.01	269.62
9	50:44.0	− 1.63748	47.15643	0.01	0.01	269.62
10	50:45.0	− 1.63748	47.15643	0	0.02	269.62

$$a = \frac{\sum y - b \sum x}{n}$$

The travel time $\hat{y}$ is influenced by $b_{n,}$ which is expressed in terms of weather, accidents, noise, delay, and other factors. Furthermore, the statistical analysis and the Pearson linear correlation coefficient showed that there was a strong relationship between travel time data and noise.

Table 4 Statistical description

Regression Statistics		
	Without noise	Noise
Multiple R	0.323891	0.950014086
R Square	0.104905	0.902526764
Adjusted R Square	0.100452	0.901693659
Stdev	0.032198	10.25998381
Skewness	0.610392862	0.934485831

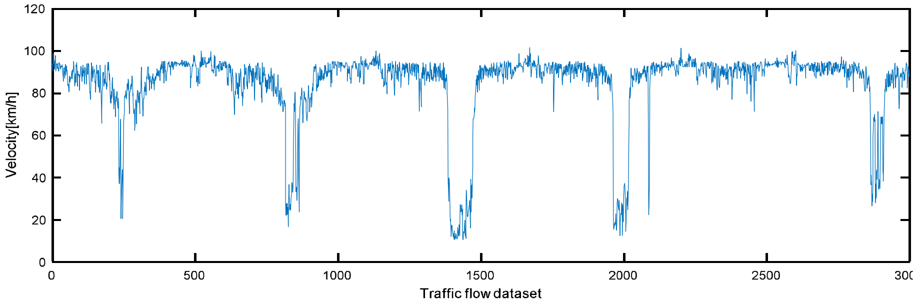


Fig. 13 Positioning Data Anomaly Detection Based on DL

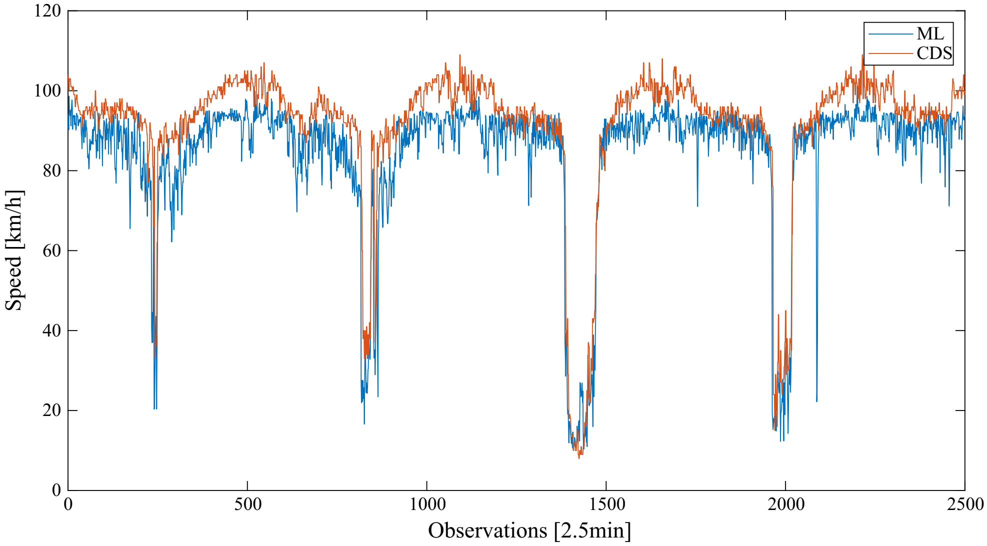


Fig. 14 Comparison between CDS and ML

$$r_p = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{11}$$

5.11 DL-based detection of positioning data anomalies

Each device on an AV was subjected to a threshold of sensitivity analysis, which involved field tests of the GNSS-based positioning terminal (GBPT) carried out under real conditions to identify positioning errors. The field tests were specialized tests with an embedded reference trajectory measurement system for delivering the ground truth. This method had the advantages of expanding the number of field tests that could be executed under real operational conditions (SaPPART). Figure 13 illustrates an anomaly detected in the data. Figure 14 illustrates the comparison between CDS and ML. The CDS detected the data anomalies early which helps to avoid traffic congestion and log term vehicular queueing.

6 Conclusion and future work

An AV uses V2V, V2I, and V2X communications to improve driving safety and traffic efficiency and to provide information and entertainment to the driver. However, AV communications can be affected by co-channel and adjacent interference, road reflection and multipath effects due to the environment. Noise causes data anomalies that can increase road accidents and congestion. Furthermore, AVs can experience bad propagation conditions for extended periods of time. The effects of these forms of interference are only partially compensated for by global models in single frequency receivers. To overcome the problems mentioned above, the original raw data was preprocessed. Missing information was compensated for with a data cleaning process that involved estimating the data from the moment before the gap. This reduced or eliminated unwanted features attributed to noise in the original data. The processed dataset was divided into training and testing subsets to carry out supervised learning. In this paper, a novel system is proposed to detect anomalies in traffic flow that lead to congestion. The DL concept, based on statistical measurements, makes possible the early detection of traffic congestion and traffic accidents. Thus, the proposed system may have a direct and significant positive impact on driver's health and safety. The simulation results demonstrate that the forecasting system was improved by the use of the DL network.

Data availability The majority of the datasets used in this paper are publicly available. Private datasets can be given upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ancans G, Bobrovsa V, Ancansb A, Kalibatiene D. Spectrum. Considerations for 5G mobile communication systems. Procedia Comput Sci. 2017;104:509–16.
2. Aqib M, Mehmood R, Alzahrani A, Katib I, Albeshri A. A deep learning model to predict vehicles occupancy on freeways for traffic management. Int J Comput Sci Netw Secu. 2018;18(12):1–8.
3. Alrajhi M, Kamel M. A deep-learning model for predicting and visualizing the risk of road traffic accidents in Saudi Arabia: a tutorial approach. IJACSA. 2019;10(11):475–83.
4. Dey KC, et al. Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication in a heterogeneous wireless network—Performance evaluation. Transp Res Part C. 2016;68:168–84.
5. Berman SD, Buczak LA, Chavis SJ, Corbett LC. A survey of deep learning methods for cyber security. Information. 2019;10:122.
6. Qin H, Yan M, Ji H. Application of Controller Area Network (CAN) bus anomaly detection based on time series prediction. Veh Commun. 2021;27:100291.
7. Brandl O. V2X traffic management. Elektrotech Inftech. 2016;133(7):353–5.
8. Borzacchielo TM. The use of data from mobile phone networks for transportation applications. In: TRB 2010 Annual Meeting. 2010.
9. Chen C, Liu Y, Sun X, Cairano-Gilfedder DC, Titmus S. Automobile maintenance prediction using deep learning with GIS data. Procedia CIRP. 2019;81:447–52.
10. Liang F, Yu A, Hatcher GW, Yu W, Lu C. Deep leaning-based power usage forecast modeling and evaluation. Procedia Comput Sci. 2019;154:102–8.
11. Polson GN, Sokolov OV. Deep learning for short-term traffic flow prediction. Transp Res Part C. 2017;79:1–17.
12. Nguyen H, Kieu L-M, Wen T, Cai C. Deep learning methods in transportation domain: a review. IET Intell Transp Syst. 2018;12(9):998–1004.
13. Suhao L, Jinzhao L, Guoquan L, Tong B, Huiqian W, Yu P. Vehicle type detection based on deep learning in traffic scene. Procedia Comput Sci. 2018;131:564–72.
14. Yan M, Li M, He H, Peng J. Deep learning for vehicle speed prediction. Energy Procedia. 2018;152:618–23.
15. Raiyn J. Speed adaptation in urban road network management. Transp Telecommun. 2016;17(2):11–121.
16. Jan B, et al. Deep learning in big data analytics: a comparative study. Comput Electr Eng. 2017;75:1–13.
17. Jawhar I, Mohamed N, Usmani H. An overview of inter-vehicular communication systems, protocols and middleware. J Netw. 2013;8(12):2749–61.

18. Park CR, Homg JE. Urban traffic accident risk prediction for knowledge-based mobile multimedia service. Pers Ubiquit Comput. 2020. https://doi.org/10.1007/s00779-020-01442-y.
19. Raiyn J. Road traffic congestion management based on search allocation approach. Transp Telecommun. 2017;18(1):25–33.
20. Raiyn J. Developing vehicle locations strategy on urban road. Transp Telecommun. 2017;18(4):253–62.
21. Ramm K, Schwieger V. Mobile positioning for traffic state acquisition. J Location Serv. 2007;1(2):133–44.
22. Lv Y, Tang S. Real-time highway traffic accident prediction based on the K-nearest neighbor method. International conference on measuring technology and mechatronics automation. IEEE: Piscataway; 2010.
23. Xiaoqiang Z, Ruimin L, Xinxin Y. Incident duration model on urban freeways based on classification and regression tree. In: 2nd international conference on intelligent computation technology and automation, TRB annual meeting. 2010; 2: 526–528.
24. Wang J, Cehn R, He Z. Traffic speed prediction for urban transportation network: a path based deep learning approach. Transp Res Part C. 2019;100:372–85.
25. Wang Z, Murray-Tuite P. Modeling incident-related traffic and estimating travel time with a cellular automaton model. In: proceedings of transportation research board's 89th annual meeting CD-ROM. DC; 10–14 Jan 2010.
26. Chrobok R, Kaumann O, Wahle J, Schreckenberg M. Different methods of traffic forecast based on real data. Eur J Oper Res. 2004;15:558–68.
27. Mishra A, Cohen A, Reichherzer T, Wilde T. Detection of data anomalies at the edge of pervasive IoT Systems. Computing. 2021. https://doi.org/10.1007/s00607-021-00927-9.
28. Chang W, Jung BC. Optimal transmission strategy without capacity loss at a primary user in cognitive radio networks over inter-symbol interference channels. IEEE Commun Lett. 2014;18(3):411–4.
29. Li J, Li S, Zhao F, Du R. Co-channel interference modeling in cognitive wireless networks. IEEE Trans Commun. 2014;62(9):3113–27.
30. Adnan A, Nordina S, Bahruddinb MAB, Alic M. How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. Transp Res Part A. 2018;118:819–36.
31. Martinez-Diaz M, Soriguera F. Autonomous vehicles: theoretical and practical challenges. Transp Res Procedia. 2018;33:275–82.
32. Xiang W, Huang T, Wan W. Machine learning based optimization for vehicle-to-infrastructure communications. Futur Gener Comput Syst. 2019;94:488–95.
33. Ndashimye E, Ray KS, Sarkar N, Gutiérrez AI. Vehicle-to-infrastructure communication over multi-tier heterogeneous networks: a survey. Comput Netw. 2017;112:144–66.
34. Silva N, Shah V, Soares J, Rodrigues H. Road anomalies detection system evaluation. Sensors. 2018;18:1–20.
35. Wang T, Zhao J, Li P. An extended car-following model at unsignalized intersections under V2V communication environment. PLoS ONE. 2018;13(2):e0192787.
36. Sepulcre M, Gozalvez J. Context-aware heterogeneous V2X communications for connected vehicles. Comput Netw. 2018;136:13–21.
37. Choudhury A, Maszczyk T, Math CB, Li H, Dauwels J. An integrated simulation environment for testing V2X protocols and applications. Procedia Comput Sci. 2016;80:2042–52.
38. Weiß C. V2X communication in Europe—from research projects towards standardization and field testing of vehicle communication technology. Comput Netw. 2011;55:3103–19.
39. Jin Q, Wu G, Boriboonsomsin K, Barth M. Platoon-based multi-agent intersection management for connected vehicles. In: 16th international IEEE conference on intelligent transportation systems. 2013.
40. Gora P, Rüb I. Traffic models for self-driving connected cars. Transp Res Procedia. 2016;14:2207–16.
41. Li J, Dridi M, El-Moudni A. A cooperative traffic control of vehicle-intersection, (CTCVI) for the reduction of traffic delays and fuel consumption. Sensors. 2016;16:1–20.
42. Bergenhem C, Hedin E, Skarin D. Vehicle-to-vehicle communication for a platooning system. Procedia Soc Behav Sci. 2012;48:1222–33.
43. Jiang T, Chen H-H, Wu H-C, Yi Y. Channel modeling and inter-carrier interference analysis for V2V communication systems in frequency-dispersive channels. Mobile Netw Appl. 2010;14:4–12.
44. Sun W, Ström EG, Brännström F, Sou KC, Sui Y. Radio resource management for D2D-Based V2V communication. IEEE Trans Veh Technol. 2016;65(8):6636–50.
45. Xinran L, Xingwu L, Yuanhong W, Juhua P, Xiangliang Z. Detecting anomaly in traffic flow from road similarity analysis. Lecture notes in computer science. Berlin: Springer Nature; 2016. p. 92–104.
46. Du LL, Dao H, Li X-X. Information dissemination delay in vehicle-to-vehicle communication networks in a traffic stream. IEEE Trans Intell Transp Syst. 2015;16(1):66–80.
47. Jing Bai J, Chen Y. A deep neural network based on classification of traffic volume for short-term forecasting, Hindawi. Math Probl Eng. 2019. https://doi.org/10.1155/2019/6318094.
48. Nam VH, Dang HN. An improvement of traffic incident recognition by deep convolutional neural network. Int J Innov Technol Explor Eng (IJITEE). 2018;8(1):10–4.
49. Li R, Pereira FC, Ben-Akiva ME. Overview of traffic incident duration analysis and prediction. Eur Transp Res Rev. 2018;10(22):1–13. https://doi.org/10.1186/s12544-018-0300-1.
50. Lu N, Cheng N, Zhang N, Shen X, Mark JW. Connected vehicles: solutions and challenges. IEEE Internet Things J. 2014;1(4):289–99.
51. Iliopoulou C, Kepaptsoglou K. Combining ITS and optimization in public transportation planning: state of the art and future research paths. Eur Transp Res Rev. 2019;11(27):1–16. https://doi.org/10.1186/s12544-019-0365-5.
52. Zhanga Q, Yang TL, Chenc Z, Li P. A survey on deep learning for big data. Inf Fus. 2018;42:146–57. https://doi.org/10.1016/j.inffus.2017.10.006.
53. Raiyn J. Road traffic anomaly detection based on deep learning technology. In: 7th international conference on vehicle technology and intelligent transport systems (VEHITS). Apr 2021.
54. Wooldrige M, Jennings RN. Intelligent agents: theory and practice. Cambridge: Cambridge University Press; 2009.
55. Russel S, Norvig P. Artificial intelligence, a modern approach. 4th ed. London: Pearson Education Limited; 2021.
56. Raiyn J. Classification of road traffic anomaly based on travel data analysis. Int Rev Civ Eng (IRECE). 2021. https://doi.org/10.15866/irece.v12i6.20530.
57. Santhosh KK, Dogra DP, Roy PP. Anomaly detection in road traffic using visual surveillance: a survey. ACM Comput Surv. 2020;6(53):1–26.

58. Dogra DP, Roy PP, Mitra A. Video trajectory classification and anomaly detection using hybrid CNN-VAE Architثcture. *IEEE transactions on Intelligent Transportation Systems*, 2021.
59. Santhosh KK, Mohapatra S, Debi MS, Dogra DP, Roy PP, Kim Mitra GB. Computer vision-guided intelligent traffic signaling for isolated intersections. Expert Syst With Appl. 2019;134:267–78.
60. Santhosh KK, Dogra DP, Roy PP. Queuing theory guided intelligent traffic scheduling through video analysis using Dirichlet process mixture model. Expert Syst With Appl. 2019;118:169–81.