International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Traffic Anomalies Detection Using Data Science and Deep Learning Methods

Yalaganti Sai Polaraju

Student, Rajam, Vizianagaram, 532127, India.

ABSTRACT

Data science is a field of study that works with enormous amounts of data using cutting-edge tools and methods to uncover hidden patterns, gather useful information, and make business decisions. The advancement of 5G has made it possible for Autonomous Vehicles (AVs) to have total control over every aspect of the operation. To be able to move independently, the AV takes autonomous actions and gathers travel information using a variety of smart devices and sensors. A computational data science strategy (CDS) is suggested for managing vast amounts of traffic data in various formats. The computational data science method was developed to find anomalies in traffic data that impair traffic efficiency. The integration of data science and cutting-edge AI methods, including deep learning, leads to a better degree of data anomaly identification, which reduces traffic jams and vehicle queues. The early identification of the variables that led to data abnormalities to prevent long-term traffic jams may be summed up as the primary contribution of the CDS technique. Additionally, CDS showed encouraging outcomes in many scenarios involving road traffic.

Keywords: Computational Data Science, Deep Learning, Autonomous Vehicle, AI, Data Anomalies.

1. Introduction

The development of the 5G network will offer new application options advancing the development of autonomous vehicles (AVs). A verification process is performed in which different types of data collected are checked for accuracy, integrity, availability and inconsistencies. Datasets affected by geographical factors are considered. Some road data arrive in an autonomous vehicle (AV) incomplete or falsified and can result in abnormal conditions on urban roads. Here, it is advised to use a deep learning model to control the dataset input. Deep learning methods have been applied in various research fields. A large dataset is needed to train such networks to develop effective models. This study focuses on the identification and classification of anomalies in road traffic. The traditional methods for surveying road conditions are very time-consuming and expensive. As an alternative, collaborative mobile sensing has been proposed which detects and automatically classifies road anomalies by applying data-mining approaches to data collected by smartphones. The dataset was analyzed based on a data science life cycle that contains the following phases: data collection, data preparation, data exploration, modelling and model evaluation and deployment. These raw data were then processed using various techniques and machine learning algorithms were applied to recognize road anomalies. The proposed computational data science approach combines artificial intelligence and process intelligent features based on goal-oriented agents such as autonomous, adaptive, mobile, and cooperative agents.

Model-based and data analysis-based anomaly detection methods are the two main categories that are frequently used. While approaches based on data analysis typically used statistical measurements, model-based approaches frequently use extremely accurate algorithms, such as machine learning schemes. Anomalies on urban roadways make drivers uncomfortable and hinder the flow of traffic. Traffic flow is abnormal when there are accidents and congestion. Traffic collisions, unfavourable weather, construction projects, and repeated lane changes are the main causes of abnormalities in AV networks.

Introduced visual surveillance to detect data anomalies in road traffic. Now some new challenges have yet to be considered, such as communication between AVs in heterogeneous wireless networks. Heterogeneity can cause delays in communication between AVs. Middleware is required for adoption in diverse wireless networks that offer different quality of services (QoS) for AV communication.

Finding road segments with a lot of noise is another difficulty for AVs. Various forecasting techniques have been suggested to detect anomalies in vehicular traffic. Recently, some deep-learning approaches have been introduced to predict urban traffic flow. In order to identify interference and other sources of abnormalities in road traffic, this work adopts a computational data science method and implements a deep learning scheme.

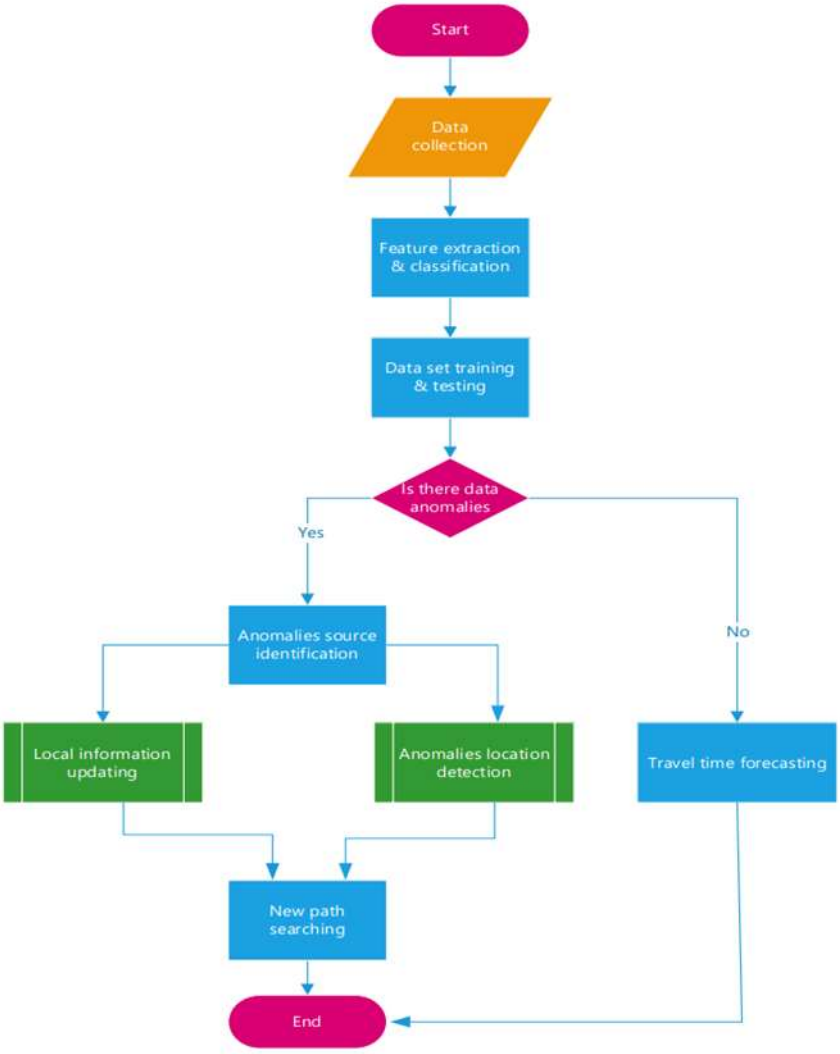

Fig-1: - CDS Algorithm

2. Literature Survey

In paper [1] Jamal Raiyn proposed that the advancement of a new era of mobile telecommunication(5G) has led AVs to be able to collect data independently & carry out data analysis based on modern AI technology. A CDS approach was used in this to characterize anomalies in road travel data. It combines the concept of the data science lifecycle which involves Data collection, Data preparation, Data availability, Data exploration, modelling and model evaluation and deployment. In this paper, a DL method is proposed for the early detection of traffic anomalies which consists of (1) the Preparation of the dataset,(2) a training phase,(3) a testing phase, and (4) a performance metrics phase.

- $y_{AD} = åx_n w_n +$ error, where x_n shows the input signal, w_n shows the weight corresponding to each input signal, and y_{AD} shows the output signal

The linear regression was applied to analyse the travel data based on mobile services. This paper proposes a novel system to detect traffic flow anomalies that lead to congestion. The DL concept makes possible the early detection of traffic congestion and accidents.

In paper [2] Y. Li, T. Guo, R. Xia and W. Xie introduced a theory by considering a large amount of uncertain information existing in traffic surveillance videos an algorithm was proposed for traffic anomaly detection for straight roads based on fuzzy theory. The traffic anomaly detection algorithm contributes:

- The fuzzy traffic flow
- The fuzzy traffic density
- The target's fuzzy motion state
- The traffic anomaly detection algorithm was proposed

Finally, experiments show that the algorithm can accurately detect traffic anomalies. It has a high accuracy rate and strong robustness.

In paper [3] Y. Yao et al. proposed an unsupervised method for traffic Video anomaly detection (VAD) based on future object localization. An ensemble method was introduced to combine object- and frame-level VAD methods to boost performance. DoTA is a large-scale dataset containing temporal, spatial and categorical annotations. Along with this a new spatial-temporal area under the curve (STAUC) metric to better evaluate VAD performance. Experimental results show that the DoTA dataset also enables research on video action recognition (VAR) and online action detection in driving scenarios. Future work will include early detection of traffic accidents and validation and verification of autonomous driving systems.

In paper [4] K. K. Santhosh, D. P. Dogra, P. P. Roy and A. Mitra proposed a colour gradient representation of vehicular trajectories which is extracted from videos recorded using a static camera. These trajectories are then classified and used to detect anomalies such as violations in lane driving, sudden speed variations, abrupt terminations, and vehicles moving in the wrong directions using a hybrid CNN-VAE detector. A high-level representation of object trajectories using a colour gradient has been proposed that can encode Spatio-temporal information of vehicular trajectories of varying lengths. A semi-supervised labelling technique has been proposed based on a modified Dirichlet Process Mixture Model(mDPMM) clustering to prepare training data for trajectory classification and anomaly detection. An autoencoder(VAE) cannot categorize an anomaly. On the contrary, CNN can be useful to classify trajectories. Thus, a hybrid CNN-VAE architecture has been used that can detect anomalies and classify them.

In paper [5] Tišljarić, Leo, Sofia Fernandes, Tonči Carić, and João Gama present a tensor-based method for the extraction of spatiotemporal road traffic patterns, to detect anomalies in the urban road network. This method is focused on two types of anomalies: The first one is sudden braking in transition which can be described as a bottleneck start, and the second type is suddenly reaching high speeds when leaving a congested region, which is known as intense acceleration in transition. The proposed method for the spatiotemporal road traffic patterns extraction includes STM computation, and the usage of the tensor composed of STMs to model the traffic patterns to address the spatiotemporal nature of the traffic data. The anomaly detection results are evaluated on the urban road network segments in a medium-sized European city.

3. Data Collection

AVs have access to a variety of tools for data collection. In contrast to conventional methods of gathering traffic data, which include using magnetic sensors and on-site surveyors, AVs make use of cutting-edge technology to address the shortcomings of conventional techniques. This paper describes a technique for gathering data using mobile services. AVs receive data from various devices in different formats, and this can lead to problems with data quality. In the preparation phase, data are converted to the desired format, then the dataset is cleaned, and inconsistent, invalid, and corrupt data are removed. The majority of the datasets used in this paper are publicly available.

4. Computational Data Science Approach (CDS): -

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

4.1 Data Science Lifecycle:

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

- Data collection:

AVs have access to a variety of tools for data collection. In contrast to traditional methods of collecting traffic data, which include human surveyors on site and therefore the use of magnetic sensors, AVs use modern technologies to beat the deficiencies of traditional methods. This paper introduces a way for data collection based on mobile services.

- Data preparation:

AVs receive data from various devices in several formats, and this will lead to problems with data quality. within the preparation phase, data are converted to the specified format, then the dataset is cleaned, and inconsistent, invalid, and corrupt data are removed.

- Data availability:

The trip dataset that was obtained includes statistical metrics.

- Data exploration:

during this study, data exploration was aimed at understanding the behaviour of the traffic flow on urban roads. Plotting may be a tool that can be used to reveal hidden patterns within a dataset. Moreover, the multiple statistical measures, like the mean and the standard deviation of a variable, and its interactions with other features, help to differentiate 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 things resolve, the travel traffic load starts to extend progressively. The record is taken into account to be abnormal when the traffic speed decreases to at least 30 km/h slower than the average speed of all records on the same day of the week at the same time. This threshold of 30km/h may be a symbolic value; it is the smallest speed change that people would consider “abnormal”. The determination of when the edge is reached depends on the travel observation data.

5. Computational Artificial Intelligence:

There are many definitions, many of which are based on the functions and behaviours of the agent under the domain consideration, which means, an agent can be defined operationally in terms 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 improve the basic rules. Based on reasoning strategies, the agent can make decisions and conclude. 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 make decisions based on complex rules. Machine learning is a subset of AI application, involving systems that learn on their own; they re-program themselves as they digest more data to perform with increasingly greater accuracy the specific tasks, they were designed for Deep learning technologies refer to a subset of machine-learning applications that teach themselves to perform a specific task with increasingly greater accuracy and without human intervention.

5.2 Deep Learning Technology

One of the most effective tools in the field of machine learning is deep learning (DL). The DL method inputs data in a hierarchical manner, involving increasingly abstract global and invariant properties at each level of processing. DL learns the features of a dataset and then combines them to achieve a specific goal. To solve complicated issues, in this example issues with intelligent transportation systems, the DL approach is applied. A DL approach is suggested in this research for the early identification of traffic irregularities. Training and testing make up its two primary components. One of the most effective tools in the field of machine learning is deep learning (DL). 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) the 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 smartphone services. The data pass through phases of pre-processing, such as cleaning and recovering missing values. In the training phase, features of the dataset extracted from the pre-processing phase are trained with DL.

5.3 Architecture of DL Concept

An input layer, a hidden layer, and an output layer are the three major components of the DL in general.

- The Input layer

The input layer for DL comprises a huge amount of data that is acquired from numerous sources. The big dataset for traffic modelling is diverse and comes from a variety of sources, such as cameras, LIDAR, sensors, and GNSS. Historical traffic statistics and near-real-time data are provided by the equipment placed in AVs.

- 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. Each layer within the hidden layer is assigned rules focused on input data attributes, which are updated in keeping with new data input. The size of the hidden layer is expressed in terms 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.

- 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, and it presents the visual results based on measurements of statistical error.

5.4 Urban Road Anomalies

Anomaly detection is defined as finding data that does not conform to the notions of normal behaviour. The relevance of this issue to urban road travel comes from the need to act upon the discovery of outliers. The proposed concept is based on the observation that a traffic anomaly can be detected by monitoring the changes in the behaviour of individual AVs (e.g., in terms of deceleration and lane changing). The proposed anomaly detection scheme can be divided into major stages. The process starts with the feature extraction stage, which involves the conversion of the original traffic variables into features, and contains 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 the 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 minutes based on mobile services. The optimal selection of neurons in the hidden layer was equal to the number of urban road sections, each of which was 300 meters long. In the second layer of the hidden layer, the number of neurons was fewer.

6. Methodology

There are numerous difficulties in the anomaly detection process. It is necessary for identifying patterns in the data that deviate from typical, expected behaviour. The definition of typical traffic on a stretch of urban road serves as the initial phase in the anomaly detection procedure. Any observations that

do not fit this typical pattern are subsequently flagged as anomalies. Finding these patterns is the fundamental difficulty in anomaly identification in metropolitan road traffic. The CDS methods can be used on all sorts of usually trafficked roadways, unlike previous schemes.

6.2 Description of DCS

To detect data anomalies in road traffic, many algorithms have been developed. These algorithms work to influence data anomalies in road traffic, which manifest as traffic congestion. On the other hand, the CDS algorithm's operations centred on data behaviour, which is influenced by a number of elements. The cognitive data distinguishes between the influence of external elements such as cyber attackers, geographical factors, and radio channel interference as well as the influence of internal factors such as the delay in vehicle-to-vehicle communication caused by changes in QoS needs. Conventional algorithms, like ML, mostly rely on describing the structured traffic data that is currently accessible; however, CDS are sensing the causes of anomalies and their locations.

6.3 Data Anomaly Detection Based on CDS

There were several different methods used to assess the travel data and AV location data. Finding the appropriate input data set was the most important challenge in the data collection. The massive data intake underwent a number of preparation processes:

- Data Recognition

The installed gadgets of AVs allow them to collect numerous types of data. To distinguish between different data set kinds, classification tools are required. Smartphones and Ublox devices were used to capture the vehicle positioning information.

- Structured vs Unstructured data

AVs typically collect organised data. A matrix is frequently used to display vehicle location data, with each column denoting a different property of the components in the matrix. A column for longitude and additional columns for height, for example, were included in the set of positioning data input for the study. Our initial step will be to create a matrix to arrange the unstructured data that some of the devices installed in AVs acquire.

- Cleaning and formatting

Deduplication and formatting are crucial steps. Different devices gather data about travel, and these data are affected by both environmental and anthropogenic influences. The act of editing data to remove irrelevant information and inaccurate or insufficient travel observations is known as data cleaning. The greatest computational data formats include a number of advantageous characteristics, such as being simple for computers to process, simple for people to read, and broadly applicable to different tools and systems. The study's input of travel data, in a variety of formats, came from numerous smart devices. This dataset was cleaned in order to increase the effectiveness of the data analysis and the calibre of the outcomes.

- Visualizing data

To display faults in real time, it is helpful to use an interactive exploratory journey dataset. Some of the datasets that were gathered weren't full because of noise and other environmental conditions that affect data transfer. Finding fields where data is absent and appropriately making up for them is a crucial part of data cleansing.

- Ranking of data

Utilizing tools created to identify missing data and data mistakes, the historical data were graded and ranked. The availability, accuracy, and integrity performance metrics—which measure the overall effectiveness of various devices—were used to grade the road segments.

6.4 Anomaly detection based on DL

The properties were extracted from the input data by the deep learning algorithm. The characteristics were categorised, and scores were given.

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

Here, x_n shows the input signal, wn shows the weight corresponding to each input signal, and y_{AD} shows the output signal. FDL is the activation function which means calculating the sum of data coming from the input.

6.5 Anomaly Detection Schemes Evaluation

AVs generally gather a range of traffic data. The raw traffic data were primarily collected in 2.5-minute cycles from smartphones. When the original data were statistically analysed, it was discovered that there were two significant flaws: the dataset was unfinished and the data contained noise.

Roads in cities are segmented, with each segment measuring 300 metres. The data underwent a number of processes, including the elimination of extraneous components like noise.

Data of poor quality might cause accidents and clogged roads. Furthermore, the urban noise created by network tunnels may prevent all of the data from being received. Statistical metrics are used to identify incomplete data.

7. Results and Discussion

For data analysis, linear regression is frequently employed. The travel data based on mobile services were analysed in this study using linear regression. Furthermore, it is advised to estimate and measure the impact of noise on travel data that is gathered by mobile services using linear regression and Pearson linear correlation. An analysis of statistics is shown in Table 1.

Regression Statistics		
	Without noise	noise
Multiple R	0.323592	0.940014072
R Square	0.104908	0.902426765
Adjusted R Square	0.100455	0.901694658
Stdev	0.032098	10.25988381
Skewness	0.610393861	0.934475831

Table 1: Regression Statistics

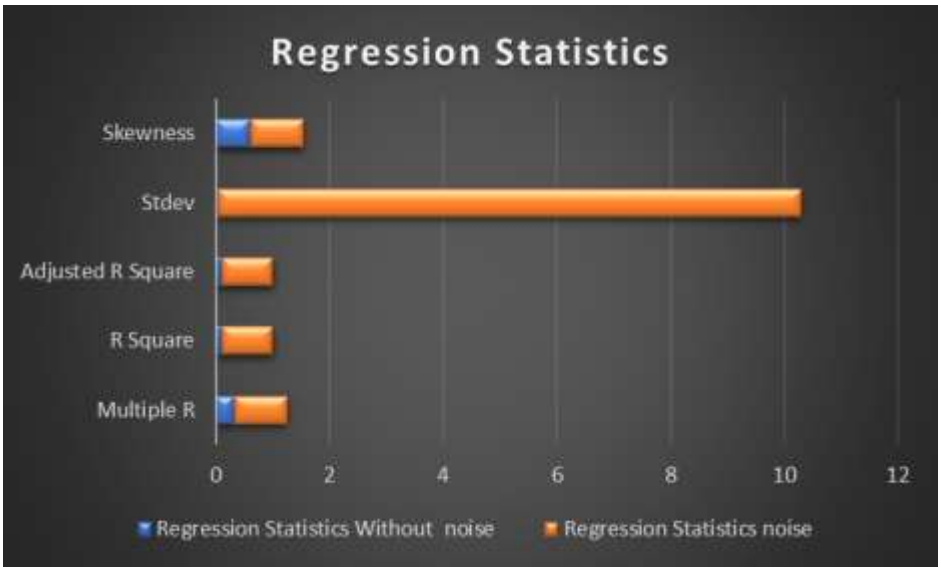

Fig 2: Graphical Representation of Regression Statistics

8. Conclusion

This study gives a solution to overcome the problems faced by Autonomous vehicles by pre-processing the original raw data. Missing information was compensated for with a data-cleaning process. 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 drivers' health and safety. The simulation results demonstrate that the forecasting system was improved by the use of the DL network.

REFERENCES

[1]. Jamal Raiyn. Detection of Road Traffic Anomalies Based on Computational Data Science, 11 January 2022, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1149975/v1]

[2]. Y. Li, T. Guo, R. Xia and W. Xie, "Road Traffic Anomaly Detection Based on Fuzzy Theory," in IEEE Access, vol. 6, pp. 40281-40288, 2018, doi: 10.1109/ACCESS.2018.2851747.

[3]. Y. Yao et al., "DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2022.3150763.

[4]. K. K. Santhosh, D. P. Dogra, P. P. Roy and A. Mitra, "Vehicular Trajectory Classification and Traffic Anomaly Detection in Videos Using a Hybrid CNN-VAE Architecture," in IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2021.3108504.

[5]. Tišljarić, Leo, Sofia Fernandes, Tonči Carić, and João Gama. 2021. "Spatiotemporal Road Traffic Anomaly Detection: A Tensor-Based Approach" *Applied Sciences* 11, no. 24: 12017, doi:10.3390/app112412017.

[6]. Zhanga, Q.; Yang, T. L.; Chenc, Z.; Li, P. A survey on deep learning for big data, Information Fusion 42 (2018). pp. 146– 157 DOI: 10.1016/j.inffus.2017.10.006.

[7]. Nam, V. H.; Dang, H. N. An Improvement of Traffic Incident Recognition by Deep Convolutional Neural Network, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2278-3075, Volume-8 Issue-1, November 2018. pp 10-14.

[8]. Liang, F.; Yu, A.; Hatcher, G. W.; Yu, W.; Lu, C. Deep Leaning-Based Power Usage Forecast Modeling and Evaluation, Procedia Computer Science, (2019). 154. pp.102-108.

[9]. Ancans, G.; Bobrovsa, V.; Ancansb, A.; Kalibatiene, D. Spectrum. Considerations for 5G Mobile Communication Systems, Procedia Computer Science, 104. (2017). pp. 509 – 516.

[10]. 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). International Journal of Advanced Computer Science and Applications, Vol. 10, No. 11, 2019. pp.475-483.