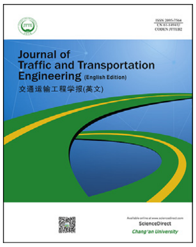
Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/jtte



Review Article

Modern data sources and techniques for analysis and forecast of road accidents: A review

Camilo Gutierrez-Osorio[*], *César Pedraza*

Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia

HIGHLIGHTS

- A state of art of machine learning techniques for road accident analysis and forecast is presented.
- The data sources used by the authors are classified according to its origin and characteristics.
- The best results are obtained when two or more machine learning techniques are combined.

ARTICLE INFO

Article history:
Received 24 October 2019
Received in revised form
11 May 2020
Accepted 15 May 2020
Available online 7 July 2020

Keywords:
Traffic engineering
Data analysis
Machine learning
Road accident forecasting
Traffic accident prediction

ABSTRACT

Road accidents are one of the most relevant causes of injuries and death worldwide, and therefore, they constitute a significant field of research on the use of advanced algorithms and techniques to analyze and predict traffic accidents and determine the most relevant elements that contribute to road accidents. The research of road accident prediction aims to respond to the challenge of offer tools to generate a more secure mobility environment, and ultimately, save lives. This paper aims to provide an overview of the state of the art in the prediction of road accidents through machine learning algorithms and advanced techniques for analyzing information, such as convolutional neural networks and long short-term memory networks, among other deep learning architectures. Furthermore, in this article, a compendium and study of the most used data sources for the road accident forecast is made. And a classification is proposed according to its origin and characteristics, such as open data, measurement technologies, onboard equipment and social media data.

For the analysis of the information, the different algorithms employed to make predictions about road accidents are listed and compared, as well as their applicability depending on the types of data being analyzed, along with the results obtained and their ease of interpretation and analysis.

The best results reported by the authors are obtained when two or more analytic techniques are combined, in such a way that analysis of the obtained results is strengthened. Among the future challenges in road traffic forecasting lies the enhancement of the scope of the proposed models and predictions by the incorporation of heterogeneous data sources, that include geo spatial data, information from traffic volume, traffic statistics, video, sound, text and sentiment from social media, that many authors concur that can improve the precision and accuracy of the analysis and predictions.

[*] Corresponding author. Tel.: +57 1 316 5000 (14076).
E-mail addresses: cgutierrez@unal.edu.co (C. Gutierrez-Osorio), capedrazab@unal.edu.co (C. Pedraza).

Peer review under responsibility of Periodical Offices of Chang'an University.
https://doi.org/10.1016/j.jtte.2020.05.002

1. Introduction

Traffic accidents have become one of the leading causes of death and injuries worldwide, with more than 1.25 million of deaths per year, between 20 and 50 million injured and a global proportion of 18 deaths per 100,000 inhabitants (World Health Organization, 2015). Most of these deaths occur in undeveloped countries (16%) and in developing countries (74%), which converts mortality from traffic accidents, not only into a public health problem, but also into a socio-economic development issue.

To address this public health and socio-economic problem, the United Nations proposes in the initiative The Global Goals for sustainable development (The Global Goals, 2017), in the numeral three, to reduce in a half the world number of deaths and injuries resulting from traffic accidents by 2020. Therefore, investigation on prediction and prevention of traffic accidents are highly relevant to fulfill this goal for sustainable development and to reduce the mortality on roads.

Several studies have been carried out to propose road accident prediction and analysis models, each one of them framed within the socioeconomic, cultural and development conditions of the country where it was proposed, and therefore, making evident the difficulty of proposing a single predictive or analytical model that works in all contexts. The main areas of interest that these models study are i) detection of problematic areas for circulation (Cao et al., 2015; Fawcett et al., 2017; Kumar and Toshniwal, 2016a,b); ii) real time detection of traffic incidents (D'Andrea et al., 2015; Gu et al., 2016); iii) road accident forecasting (Chen et al., 2016; Lin et al., 2015; Lu et al., 2017; Park et al., 2016; Shi and Abdel-Aty, 2015; You et al., 2017); and iv) prediction the severity of the consequences suffered by involved in a road accident (Kaplan and Prato, 2013; Mohamed et al., 2013; Zheng et al., 2019).

Current advances in analytic algorithms and machine learning methods allow researchers to propose models of complex issues in the study and prediction of road accidents, such as the mining of heterogeneous data sources (Moriya et al., 2018), the process and classification of real time data flows (Chen et al., 2018; D'Andrea et al., 2015), the need to process multi-dimensional data having strong prediction capabilities (Zheng et al., 2019), and to study, analysis and model the real-time information provided by on-road equipment such as loop inductors, cameras and radar equipment (Ozbayoglu et al., 2016; You et al., 2017). Therefore, the study of road accident prediction is a field of relevant and current scientific knowledge, open to innovation in the research of algorithms and data analysis techniques that respond to the challenge of generating a more secure mobility environment, which considers the particularities of each country or region, i.e., traffic composition, weather conditions, roads conditions, and demography.

In order to make an appropriate comparison of the results reported by the authors of each of the investigations cited in this paper, the following metrics were used (Powers, 2011).

- Accuracy: represent the rate of instances correctly classified over the total number of instances. The ideal value for accuracy is 1.00 (100% classification accuracy).
- Precision or confidence: defined as the proportion of predicted positive cases that are correctly predicted or labeled as real positives.
- Recall: calculated as the proportion of real positive cases that are accurately predicted positive.
- *F*-measure: weighted average of the precision and recall.
- Mean absolute error: also known as average prediction error, is the average of the difference between predicted and actual value in all the test cases. A low mean absolute error (MAE) indicates good predictive accuracy.
- Mean squared error: determined as the average of the squared differences between each computed value and its corresponding correct value.
- Root mean squared error (RMSE): RMSE is calculated as the square root of the MSE and is used as a measure of differences between valued predicted and the real values. A lower value of RMSE is an indicative of a higher prediction precision.

The objective of this paper is to present a review of the state of the art in the prediction of road accidents through algorithms and advanced techniques for analyzing information and the incorporation of new data sources, which were not present in published reviews on road accident prediction and forecasting (Halim et al., 2016; Taamneh et al., 2017). In this document emphasis was placed on documents published from year 2015 onwards and the review of novel methods to analyze and forecast traffic accidents. It must be considered that methods and algorithms, such as natural language processing (NLP) and deep leaning are widely employed in other fields of knowledge, such as sentiment analysis using NLP, or advance image processing employing deep learning algorithms, but their application in traffic accident investigation is more recent. In this paper a section is dedicated to review the most relevant data sources employed by researchers, including new information sources on traffic accidents, such as social media and open data provided by governments.

This paper is organized as follows: Section 2 gives a review and a classification of the most employed data sources available to perform analysis and predictive research.

Section 3 describes the algorithms employed to carry out data analysis on road accident, including their strengths and limitations. Section 4 gives a compilation of methods and algorithms employed on road accident forecasting and prediction. Section 5 provides the discussion and Section 6 covers the conclusions and future challenges.

2. Road accident data sources

Many authors agreed to refer to the information provided by public or private agencies that fulfill the functions of operational traffic control in each country, whether they are road security agencies or highway police (Castro and Kim, 2016; Kaplan and Prato, 2013; Kumar and Toshniwal, 2016a,b; Taamneh et al., 2017), while other authors use publicly accessible datasets on internet (Delen et al., 2006; Kononen et al., 2011). These data sets contain demographic information about those involved in the accident, a variable level of detail regarding road conditions and environmental settings, technical details about the vehicles involved and their geographical position, the degree of the severity of the resulting injuries on drivers, passengers and pedestrians, among other relevant variables. On the other hand, some authors gather their data by installing equipment on vehicles, for example satellite positional systems (GPS, GLONASS, Galileo), cameras and sensors, in order to gather data like acceleration, unexpected braking events, sudden lane changes and information about the driver behavior and status like drowsiness and level of stress (Cao et al., 2015; Zheng et al., 2014). Another emerging data source suitable for proposing models of road accident prediction is social media (Anantharam et al., 2015; Wang et al., 2016, 2017; Zhang et al., 2018). The ubiquity and availability of social media makes feasible to obtain in real time information reported by road users that cannot be found in other data sources, such as road infrastructure deterioration, parked vehicles, minor traffic incidents and incidents around the road. A summary of the most relevant data sources is shown in Table 1.

The incorporation of new sources of information such as mobile applications, internet of things (IoT) devices and more intelligent instrumentation and hard-ware available in new vehicles, such as on-board computers, GPS and sensors, it is expected that there will be a broader availability of data that may be susceptible to more extensive and in-depth analyzes, that involves additional information about drivers, passenger and pedestrians, and their habits and daily activity, and more detailed data about environmental, climatic and lighting conditions, as well as information about events and incidents around the road that could affect the safety of pedestrians, drivers and passengers.

2.1. *Government data sets and open data*

Government data refers to those data sets that are generated, collected, preserved, stored and made available to the public by government entities or those that are delegated to exercise functions of control, execution or reporting of information concerning road accidents (Cao et al., 2015; Kumar and Toshniwal, 2015a; Taamneh et al., 2017). Among these agencies can be included police bodies, traffic police and road concessionaires. Government data can be characterized as historical, since it contains data spanning several

Table 1 – Road accident analysis and prediction data sources including government data, open data, measurement technologies, vehicle onboard equipment and social media.

Type of data source	Author	Description
Government data	Zheng et al. (2019), Moriya et al. (2018), Taamneh et al. (2017), Castro and Kim (2016), Kumar and Toshniwal (2015a), Cao et al. (2015), Çodur and Tortum (2015), Scott-Parker and Oviedo-Trespalacios (2017), Tiwari et al. (2017), Hashmienejad and Hasheminejad (2017), Ghosh et al. (2017), Alkheder et al. (2017), Ren et al. (2017)	Data sets that are generated, collected, preserved, stored and made available to the public by government entities or those that are delegated to exercise functions of control, execution or reporting of information concerning road accidents
Open data	Australian Census (2019), Data. Gov. (2019), Data. Gov. UK (2019), Datos. Gov. Co (2019)	Open data catalogs are maintained by government agencies and are available to all public without restriction. The data must comply all legislation regarding privacy and confidentiality
Onboard equipment	Cao et al. (2015), Xiong et al. (2017)	Onboard equipment refers to all devices installed on a vehicle that can store or transmit data concerning the vehicle variables and driver conditions
Measurement technologies	Roshandel et al. (2015), Yuan and Abdel-Aty (2018)	Measurement technologies include all kind of equipment that is part of the road infrastructure, such as radar, cameras, or equipment embedded on the road itself, i.e., loop detectors
Social media	D'Andrea et al. (2015), Gu et al. (2016), Zhang et al. (2018), Sinnott and Yin (2015), Nguyen et al. (2016), Amin-Naseri et al. (2018), Dabiri and Heaslip (2019), Pandhare and Shah (2017), Salas et al. (2018)	Social media can be considered the newest developed data source in traffic and road accident related studies, and currently the most used data source comes from Waze, Inrix, Google Maps and Twitter streams

decades, and can be considered as reliable, because it is supported by the custody process of the entities responsible for the data. Government data is usually the technical support for the generation of public policy in each country regarding aspects like road infrastructure design and road security plans. One aspect to consider is that not all the variables of the data set may be available for public access, such as specific demographic information, sex and ethnicity, incompliance with the information privacy laws current in each country.

Open data can be defined, according to previous studies (Janssen et al., 2012; Veljković et al., 2014), as the data that is produced and funded with public money, that is make available and accessible without restriction to the public taking in to consideration privacy and confidentiality matters. The most used method to make available open data content to the general public is by means of dedicated web sites that enable search and expose interfaces, including web services, to allow automatic consumption of the data and integrate it to other web sites or applications. Road traffic information is usually one of the most available data, among other topics such as population, economic statistics and geographic information. Relevant examples of open data portals are the United States government open data catalog for traffic accidents (Data. Gov., 2019), that includes information from all the country, United Kingdom open data catalog (Data. Gov. UK, 2019) and Australia open data catalog (Australian Census, 2019), that includes real-time transit, traffic accidents and public transport time tables.

Regarding data quality and data preprocessing, as general steps, this kind of data sources require the handling of missing values and normalization and transformation of variables depending on the requirements of the type of analysis and the type of machine learning technique selected.

2.2. *Measurement technologies*

Measurement technologies include all kind of equipment that is part of the road infrastructure, such as radar, cameras, or equipment embedded on the road itself, i.e., loop detectors.

Loop detector, video surveillance, microwave and laser radar had been used in many studies, considering their availability and accessibility (Roshandel et al., 2015). Additionally, Bluetooth detectors and adaptive signal control datasets had been used to collect data from road intersections (Li et al., 2020; Yuan and Abdel-Aty, 2018). Loop detector data can be characterized as not highly dimensional, since it contains only variables related to type of vehicle, vehicle speed and time of the record, and information related to the loop itself, such as a loop localization and status. Loop detector arrays are not expensive, compared to other road equipment such as cameras or radars, and can be deployed along a main road or expressway. Conversely, loop detectors are not highly reliable (Ahmed and Abdel-Aty, 2012), since they tend to fail due to the harsh road infrastructure conditions, such as temperature, vibration and pavement variations.

On the other hand, video surveillance technologies are reliable and can gather additional information, such as vehicle occupation, vehicle model and vehicle make. Radar equipment and video equipment have improved over time in term of cost, reliability, accuracy, and ease of use.

2.3. *Onboard equipment*

Onboard equipment refers to all devices installed on a vehicle that can store or transmit data concerning the vehicle variables and driver conditions. The onboard equipment may include global positioning units, cameras set up to record road conditions or driver situations such as drowsiness or alert status (Zheng et al., 2014), accelerometers, vehicle condition recorders, such as change on vehicle speed, sudden braking event or lane changes, and finally, direction and acceleration in case of an impact or collision (Cao et al., 2015). One innovative approach was the proposed by Xiong et al. (2017) that used an advance system called chain road traffic incident to simulate vehicle collision settings based on PreScan platform.

On the subject of the data preparation and quality processes required to build a suitable dataset using data from onboard equipment, the most commons activities, as reported by authors (Wang et al., 2019), are i) remove outliers and hardware errors reported by GPS equipment; ii) remove unrelated data outside the area of study; iii) match captured GPS traces or data collected with the road segments defined in the analyzed area; and iv) filter data based on vehicle direction or other condition required.

2.4. *Social media*

Social media can be considered the newest developed data source in traffic and road accident related studies, and currently the most used data source comes from Waze, Inrix, Google Maps and Twitter streams as reported by Sinnott and Yin (2015), Salas et al. (2018), Nguyen et al. (2016), and Amin-Naseri et al. (2018).

Social media data can be labeled as unreliable, biased and difficult to interpret. Social media information is unreliable because is not easy to assess the trustworthiness of its origin or publisher; social media is difficult to interpret, because the users employ local jargon to post their content and the text may contain spelling and grammatical errors; and finally, it can be stated that social media data can be biased since, in the case of road accidents, not all the relevant information about accidents are report by the road users.

The unreliability of social media can be address by using methods that extract the time, localization, and subject of the report, as proposed by Lu et al. (2018) and Gu et al. (2016), in order to correlate the report to a real world incident; to deal with the deal with the grammatical complexities of social media, Dabiri and Heaslip (2019) proposed a method based on a deep learning architecture and Chen et al. (2017) described a method based on a convolutional recurrent network; both papers aim to over pass the limitation of the bag-of-words and predefined set of key words methods that are usually employed to process the content of the tweets. In order to tackle the bias of social media, Bao et al. (2017) proposed that the inclusion of human activity information reported in Twitter in spatial analysis of traffic accidents can improve the effectiveness and performance of a road accident model.

Regarding the data quality and preparation of the raw data obtained from social media, based on D'Andrea et al. (2015), Salas et al. (2018), Zhang et al. (2018) and Afzaal et al. (2018), the following steps can be summarized.

- Removing all non-relevant meta-information, such as hash tags, links, special characters, punctuation, and non-text characters.
- Tokenization or transformation of a stream of characters into a stream of processing units or token.
- Removal or filtering of stop-words that consist in eliminating common words such as articles, conjunctions, prepositions, and pronouns.
- Text transformation, using techniques such as stemming, in order to reduce each word to its stem or root form.

As a result of the aforementioned steps, the unstructured structure of social media data is converted into a structured database useful for feature analysis and classification tasks.

2.5. *Fusion of several data sources and big data*

Several authors had employed two or more of the aforementioned data sources, seeking to enhance the accuracy of the results obtained by their models and to have a better understanding of road accidents. Generally, the fusion of several data sources consists of the merger of heterogeneous data from road accidents, traffic information, weather condition, road infrastructure (including road geometry, and road lighting conditions), land use and demographic data related to the area of analysis, as studied by Fan et al. (2015), Yuan et al. (2017) and Bao et al. (2019). As a special case, Bao et al. (2019) proposed an useful classification of all the variables used in their model according to three categories, i) Type I variables, grouping variables spatially varied but temporally static, such as variables related to land use, population and road infrastructure, ii) Type II variables, grouping variables only temporally varied but spatially static, related to weather conditions, and iii) Type III variables, grouping variables both temporally and temporally varied, grouping the variables road crash risk and number of taxi trips collected in the area of study.

Concerning the scope of this review, the most relevant data sources of big data information relevant to road accident analysis and prediction are provided by intelligent transportation systems (ITS) and GPS traces from onboard equipment and cellular phones (Chen et al., 2016; Ren et al., 2017; Shi and Abdel-Aty, 2015). The GPS traces data can be characterized as composed by millions of records, with a temporal resolution of minutes (e.g., 5–10 min between each sample) and spanning several months or years of data. ITS systems installed on roadways, such as microwave vehicle detection systems (MVDS), continuously collect road data such as vehicle speed, volume, occupancy and vehicle type and can generate millions of records per week (Shi and Abdel-Aty, 2015). Another example of ITS system that generate big data information is a vehicle license plate recognition (VLPR) system, that can generate more than 7 million records each day (Chen et al., 2018).

By using big data information, large scale simulations (Sun and McIntosh, 2016) and general models of traffic accident risk models can be proposed (Chen et al., 2016). Furthermore, big data can help to design citywide predictions of accident risk (Chen et al., 2016, 2018; Fan et al., 2017) (Fig. 1).

3. Road accident analytic methods

By using analytic methods, researchers seek to characterize the information and variables of the road accident, in order to discover hidden patterns, profile behaviors, generate rules and inferences. These patterns are useful to profile drivers or drivers' behavior on the road, to delimitate unsafe areas for driving, to generate classification rules related to road accident data, to perform selection of variables to be fetched in real-time model of accidents and to select relevant variables to be used to train other methods, such as artificial neural networks and deep learning algorithms.

On the aspect of the algorithms and computational methods reported by the authors employed to analyze road

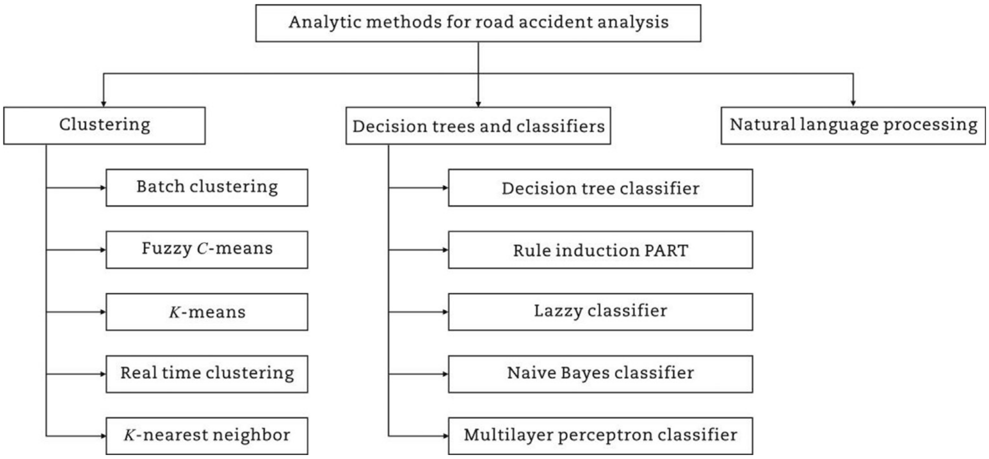


Fig. 1 – Analytic methods for road accident analysis.

Table 2 – Representative studies and methods on road accident data analysis.

Author	Research problem-computational method	Data source	Result
Cao et al. (2015)	Correlate abrupt braking events in real time-batch clustering, fuzzy *C*-means and real time clustering	Data from driving events for seven vehicles by the DAP platform from Ford Motor Company	Correlations that indicate potentially dangerous places for driving, according to the time of day
Kaplan and Prato (2013)	Determinate the variables that influence the severity of road accidents between cyclist and drivers (latent class clustering)	Data reported in Denmark (2007–2011), accidents involving cyclist and drivers	13 clusters showing specific patterns of urban and rural road accidents; obtaining a high classification accuracy, with all the clusters being correctly assigned for more than 80 percent of the observations, and reporting an entropy criterion of 0.86
Depaire et al. (2008)	Find patterns of severity of injuries resulting from road accidents in a heterogeneous data set (latent class clustering)	Accident data reported by the Belgian road police (1997–1999), 29 variables and 4028 accident records	7 clusters showing a high level of accidents for motorcyclist and cyclists under 19 years old
De Oña et al. (2013)	Identify the key factors that affect the severity of injuries caused by a rural road accident (latent class clustering, Bayesian networks)	3229 traffic accidents reported by the police in rural roads of Granada (Spain), occurred between 2005 and 2008	Results depends strongly on the initial data set analyzed and the techniques used
Zheng et al. (2014)	Determinate the variables that identify a driving style or driver with high risk of vehicular collision (cluster *K*-means)	31 vehicles with a GPS for 60 days, driving recorders and cameras to capture the road and driver's facial expression	3 clusters for road accident risk levels, a correlation between driving events and maximum deceleration
Kumar and Toshniwal (2015a)	Determinate the variables that influence the event of road accidents (cluster *K*-means, association rules model)	11,574 traffic events on the roads of Dehradun (India) (2009–2014)	6-cluster model as input to a model of association rules. Severity of accident, type of road, lighting and surrounding area affect the aggregation of the clusters
Taamneh et al. (2017)	Determination of the most important variables for severity prediction of traffic accident (J48 decision tree, rule induction PART, Naive Bayes)	5973 traffic accident records occurred in Abu Dhabi between 2008 and 2018	Age, gender, nationality, year of the accident affect the severity of the accident
Beshah et al. (2011)	Understand the interaction between the different actors that intervene in a road accident (CART and random forest)	14,254 traffic accidents with 48 attributes (May 2005–September 2008) Ethiopia	CART and RF behave similarly but RF has lower failure rates to predict the probability of someone emerging unscathed from a road accident
Ahmed and Abdel-Aty (2012)	Prediction of traffic accidents in real time with information provided by an automatic vehicle identification (AVI) (random forest)	Real-time data on speed, average speed and traffic volume obtained from AVI along 125 km highway at Orlando (FL) 2008	The model is sensitive to the distance between each tag to AVI, over 4 km nor statistically or predictive significant values were obtained

accident data, as summarized in Table 2, the most used are: i) clustering algorithms (Cao et al., 2015; Kumar and Toshniwal, 2015a; Moriya et al., 2018); ii) decision trees and classifiers (Castro and Kim, 2016; Gutierrez-Osorio and Pedraza, 2019; Scott-Parker and Oviedo-Trespalacios, 2017; Taamneh et al., 2017); iii) association rules (Ait-Mlouk et al., 2017; Ait-Mlouk and Agouti, 2019; Kumar and Toshniwal, 2015b) and iv) natural language processing algorithms (D'Andrea et al., 2015; Gu et al., 2016; Salas et al., 2018).

3.1. *Clustering algorithms*

Clustering is a method of partitioning and grouping objects into groups (clusters), so that objects grouped in each cluster share common characteristics, while looking for them to be clearly different from other objects grouped in other clusters. Common characteristics can be interpreted as the level of correlation of objects according to the characteristics on which clustering techniques are applied. Unlike classification methods, clustering does not require that the data be previously marked with any particular category in order to distinguish different groups within the data. The absence of these previous categories or classes indicates that the objective of clustering is to find an underlying structure in the information and achieve a more compact representation of it instead of discriminating future data into categories. The main advantages of clustering algorithms are that they do not require prior data processing, work well with large data sets, and their results are interpretable graphically. On the other hand, clustering algorithms are sensitive to the possibility of finding a local maximum instead of a global maximum on their optimization functions.

According to Jain et al. (2000), clustering algorithms can be classified according to the representation of their results and how they perform the grouping and partitioning on the data set. Clustering algorithms use a distance function to calculate the similarity in characteristics when they work with continuous elements and a measure of similarity for data with qualitative elements. Among the techniques based on similarity functions we can include *K*-nearest neighbor and *K*-means clustering (Kumar and Toshniwal, 2015a; Zheng et al., 2014). In the case of cluster techniques whose similarity function is based on distribution probabilities, their operation is based on the premise that each cluster has an underlying probability of distribution from which the data elements are generated. An example of this type of algorithm is latent class clustering (LCC) (De Oña et al., 2013; Depaire et al., 2008; Kaplan and Prato, 2013). For data sets with attributes both qualitative and quantitative, clustering techniques such as two-step clustering (Ma and Kockelman, 2006) may be used, in which a pre-allocation of clusters is performed using a function of logarithmic distance and then said pre-allocation is validated by comparing their distances to a given threshold value, then the clusters are joined if the distance value is greater than the defined threshold value.

Cao et al. (2015) applied batch clustering, in combination with fuzzy *C*-means and real time clustering to study abrupt braking events in real time, considering the time and location to determine sectors where driving is dangerous. As a result of the analysis of batch clustering results, correlations were obtained that indicate potentially dangerous places for driving, according to the time of day. The investigation carried out by Kumar and Toshniwal (2016a,b) employed *K*-means clustering and association rules model in order to determinate the variables that influence the event of road accidents, obtaining a 6-cluster model, which was used as an input to a rules association model. It was found that the criteria that influenced the aggrupation of the clusters were accident severity, type of road, lighting present in the road and the type of surrounding area.

Moriya et al. (2018) analyzed real-traffic data from Tokyo city in order to predict the number of accidents on any road or intersection and to identify risk factors using clustering to group roads and finding risk patterns. The quantity of clusters was evaluated and selected using the Bayesian information criterion (BIC) (Fraley and Raftery, 1998) and Akaike information criterion (AIC) (Akaike, 1987). The authors identified three clusters, two of them that group risky locations and divide high-risk locations and one grouping low risk locations.

3.2. *Classification algorithms and decision trees*

A decision tree builds classification models in the form of trees or dendrogram, each node represents one of the input variables, and each node has several branches equal to the number of possible values of said input variable. Likewise, each leaf node is a value of the target attribute and represents the decision made based on the value of the input variables in its path from the root node to the leaf. Decision trees are useful tools in pattern classification applications. Its greatest utility is that its domain knowledge is not required for its construction, its method of analysis is exploratory and not inferential. They can be used with highly dimensional data and with data sets with incomplete information. Rule learners and classifiers do not require prior data processing, they work well with large data sets and rule learners and classifiers are interpretable graphically; however, their results are not as accurate compared to other methods.

Regarding classification algorithms and decision trees, the most reported are as follow.

- C4.5 is a decision tree that uses the concept of information gain or difference in entropy as a criterion of information division (Breslow and Aha, 1997). C4.5 can process numerical, nominal, missing attributes and noise data.
- PART rules classifier (Breslow and Aha, 1997) is based on an iterative process that follows a principle of divide and conquer. At each iteration, a subset of the training data set is used to generate rules using a decision tree. PART uses an algorithm that creates a partial C4.5 decision tree for a set of selected instances and chooses the leaf with the largest coverage to be a rule (Krishnaveni and Hemalatha, 2011).
- Rule induction is a rule-based algorithm that follows and iterative process that uses a divide-and-conquer approach and produces a set of if-then rules to classify the data. At each iteration, a subset of the training set is used to generate rules using one of the decision tree algorithms.

- CART algorithm or classification and regression trees (Gey and Nedelec, 2005), is a non-parametric technique that can select variables from a data set and determine hidden patterns, structures and relationships. Among its strengths is that it can use raw data, handle attributes with missing values, automatically manage nominal predictors and process massive data sets with a large number of predictors.
- Naive Bayes classifier (Wu et al., 2008) is a probabilistic classification algorithm that is based on Bayesian theorem, with the assumption on the independence the input attribute. Naive Bayes has the advantages that is fast and scalable and can be used for both binary and multiclass classification problems.
- Multilayer perceptron (MLP) can be used as a classifier algorithm and the most common architecture employed is a feed-forward artificial neural network, which uses back propagation algorithm. MLP consist on an input layer, one or more hidden layer and an output layer and each neuron in each layer is fully connected to every neuron in the adjacent layers.

Taamneh et al. (2017) employed decision tree classifier, rules induction PART, multilayer perceptron and Naive Bayes to determinate the most important variables suitable for the prediction of the severity of a traffic accident. By comparing the different ruled based models obtained, the authors concluded that the decision tree classifier and rules induction had the better accuracy, with a value of 0.08218. The variables that have more weight in accident fatality were age, gender, nationality, year of the accident and type of accident.

Tiwari et al. (2017) evaluated the performance of the classifier algorithms such as decision tree, lazy classifier, and multilayer perceptron, analyzing a dataset containing traffic accidents. It was reported that the best accuracy was obtained by the lazy classifier using clustered data, with an accuracy of 0.8235. The most relevant conclusion obtained by the authors was that the treatment of the dataset by using clustering algorithms, in this case hierarchical clustering, can lead to a better performance of classifiers, comparing the performance of the same group of classifier algorithms on a non-clustered dataset.

Scott-Parker and Oviedo-Trespalacios (2017) employed a hierarchical tree to segment road accident risk in young drivers, of ages 16 to 25, from Australia, New Zealand and Colombia. The main predictors of accident risk were mobile phone usage, the use of alcoholic drinks and driving with passengers.

3.3. *Association rules*

Association rule mining is a technique that extracts correlations among different attributes in a defined data (Zhang and Wu, 2011). As a result of applying this technique, it is obtained a set of rules that define the correlation between different set of an attributes in the data set. The quality of the rules is measure using support value, or the frequency of occurrence of a rule and confidence or reliability of a rule.

Ait-Mlouk and Agouti (2019) proposed a software framework based on association rules applied to the a dataset of road accidents occurred in Morocco, in order to extract meaningful relationships between variables related to a road accident, and then through multiple criteria analysis, select the most relevant rules. Finally, using the set of selected rules, the framework can predict death and injuries based on time series analysis.

3.4. *Natural language processing*

Natural language processing (NLP) algorithms are mainly used to process road accident events reported by social media and infer information such as geolocation, road accident features and relevant variables. NLP algorithms are used to perform classification of social media content, according to a predefined target classes, generally as a binary classification, being the social media content related or not to traffic accident.

D'Andrea et al. (2015) presented a real-time monitoring system for traffic event detection, including road accidents, captured from Twitter data stream. The objective was to label each tweet, as traffic event-related or not, using a support vector machine (SVM). The SVM was evaluated against seven different classification models, including Naive Bayes classifier, C4.5 decision tree, *K*-nearest neighbor (KNN) and PART classifier. The authors reported an accuracy value of 0.9575, which was a benchmark in this area of knowledge at the time of the publication of their paper. The authors also performed a 3-class classification model using a different dataset, in which every tweet was leveled as a) traffic due to external event, b) traffic congestion or accident-related and c) non-traffic related. The accuracy value for the 3-class classification model was 0.8889.

Gu et al. (2016) studied how to increase the effectiveness of road incident detection, using Twitter as data source, employing natural language processing. Their results showed that after processing the acquired data, only 5% of the data was useful, under the assumption that the tweets were traffic incident related and being able to geocode the data on a map. The results were validated against official traffic incident data sources, such as road condition report system and police traffic incidents reports. The authors affirmed that there was a strong pattern in the frequency of the information posting, having a peak on weekends. The authors reported an accuracy value of 0.9500 for the overall classification of the dataset as traffic incident related. On the other hand, the accuracy value for the process of obtaining geocoding information from the tweets dataset was 0.5200.

Salas et al. (2018) presented a methodology for retrieving, processing, and classifying information from Twitter related to traffic incidents, by combining natural language processing and support vector machine algorithm in order to perform text classification. The authors will include in their future work other techniques to improve their accuracy and to explore sentiment analysis within the content of the tweets.

4. Road accident forecasting

In the subject of road accident forecasting, machine learning methods had been employed widely, based on the ability of machine learning methods to process multi-dimensional data, flexible implementation and coding and strong predictive capabilities (Zheng et al., 2019). Considering the field of road accident forecasting, the researchers pursue to predict the occurrence of an accident using a defined set of conditions or variables, and to model road accident severity models or perform postmortem analysis (Halim et al., 2016).

Among the most used methods to predict traffic road accidentally, as shown in Table 3 and Fig. 2, the following can be taken into account: i) Bayesian networks (Castro and Kim, 2016; Ghosh et al., 2017; Yuan and Abdel-Aty, 2018); ii) genetic algorithms and evolutionary computing (Hashmienejad and Hasheminejad, 2017; Kunt et al., 2011); iii) support vector machines (Pandhare and Shah, 2017; Xiong et al., 2017); iv) artificial neural networks (Alkheder et al., 2017; Çodur and Tortum, 2015); and v) deep learning algorithms (Chen et al., 2018; Dabiri and Heaslip, 2019; Halim et al., 2016; Zhang et al., 2018; Zheng et al., 2019).

4.1. *Bayesian networks*

Bayesian networks (BN) allow the modeling of phenomena through a set of random variables and the relationships that exist between those variables, represented as probability distributions. The variables represent the qualitative knowledge of the model by means of the directed acyclic graph (DAG) in which the variables are represented as nodes. The dependency and conditional independence relationships between the variables are represented as arcs between the nodes. Bayesian networks are used for the prediction of conditions that cause accidents and to propose accident severity models.

The research published by Castro and Kim (2016) employed Bayesian networks, J48 decision tree and an artificial neural network to determinate the most important variables to predict the severity of road accidents, finding that the Bayesian network generated the most accurate model, with an accuracy value of 0.8159, a precision value of 0.7239, a recall value of 0.7239 and a *F*-measure of 0.723. It was determined that the lightning conditions, the type of road and the type of maneuver that the vehicle was performing

Table 3 – Representative algorithms and methods used on road accident forecast.

Author	Research problem – computational method	Data source	Result
Castro and Kim (2016)	Determination of the most important variables to predict the severity of road accidents (Bayesian network, J48 decision tree, artificial neural network)	A set of data on road accidents reported by road safety accidents (England) was analyzed and a total of 451,462 traffic accidents were collected in a time interval from 2010 to 2012. A sample of 81,690 records was selected to be analyzed	The Bayesian network generated the most accurate model. The lightning conditions, the type of road and the type of maneuver are determining factors of the accident
Kononen et al. (2011)	The modelling of the severity of the injuries resulting from a traffic accident (logistic regression)	Data was taken from the national automotive sampling system (NAAS), crashworthiness data system, from the United States (1998–2008), with records from serious injuries to the vehicle's driver	The results of the study indicated that the delta-*V* (change in speed associated with primary direction from the force of the crash event), use of the seat belt and the direction of the collision were the most important predictors of the severity of the accident
Moghaddam et al. (2011)	The correlation between a set of variables to determine the severity of traffic accident that occurred on an urban road (artificial neural network)	Traffic accident data obtained from the information collection form used by the Tehran Police Department (Iran) (2003–2007) with a set of 52,447 accidents	The variables that had the most influence on the severity for a traffic accident were if the type of collision was frontal, the type or road was an urban road 18–22 m wide, involving motorcycles and bicycles, with the vehicle turning to left and including infraction of lane space and violation of the speed limit
Hashmienejad and Hasheminejad (2017)	A prediction model of the severity of the injuries resulting from traffic accidents that occurred in urban and rural roads (artificial neural networks, genetic algorithms, pattern search)	The data set consisted of 14,211 road accidents that occurred on rural and urban roads in Iran, reported by the police department between 2008 and 2013	A set of rules were used to generate decision trees (ID3, CART and C4.5), which were then validated by means of a test data set in order to contrast the veracity of the proposed rules

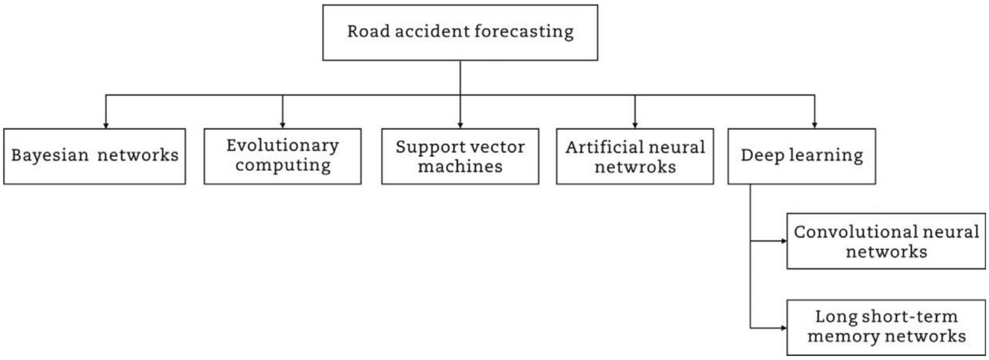

Road accident forecasting
Bayesian networks
Evolutionary computing
Support vector machines
Artificial neural netwroks
Deep learning
Convolutional neural networks
Long short-term memory networks

Fig. 2 – Representative algorithms and methods used on road accident prediction.

were determining factors of the accident, however, the severity of the accident was difficult to predict in a precise way.

4.2. *Genetic algorithms and evolutionary computing*

Genetic algorithms tend to converge on the optimal solution and not fall into local optimal values, generally the original data set must be processed, i.e., normalized, in order to be used as an input to the algorithm. Genetic algorithms represent their solution to a problem as a chromosome, a chromosome is composed by a set of genes, and each gene can be understood as the particular value of a variable in a particular set of genes called population. The process to obtain an optimal solution is attained by executing an iterative evolutionary process, which is initialized using a randomly chosen population, and each iteration is called a generation. To produce a new generation, the operator crossover and mutation are executed on different chromosomes. The process is repeated until the stop criteria were met, being that criteria either the number of iterations completed, or the value of fitness function is met.

Hashmienejad and Hasheminejad (2017) employed genetic algorithms and decision trees to propose a prediction model of the severity of the injuries resulting from traffic accidents that occurred in urban and rural roads. The model was generated by applying a set of training data to a multi-objective genetic algorithm in order to extract a set of classification rules to predict the severity of the road accidents. These set of rules were used to generate decision trees (ID3, CART and C4.5), which were then validated by means of a test data set in order to contrast the veracity of the proposed rules. The proposed method yields a precision value of 0.885, a recall value of 0.889, an accuracy of 0.8820 and an *F*-measure value of 0.8875, superior to other methods used by the authors for comparison such as SVM, Naive Bayes, KNN and ANN.

4.3. *Support vector machines*

Support vector machines are suitable for dealing with problems that involve linear classification tasks, as well as nonlinear classification tasks, concerning separable or non-separable problems. An SVM maps the data points from the input space to high-dimensional feature space, or hyper plane that represent the best margin between the sets, using a kernel function.

Xiong et al. (2017) proposed a framework to predict traffic collisions, based on an SVM that identified the driver maneuver as leaving lane (LL) or remaining in lane (RL) and a Gaussian-mixture hidden Markov model (HMM) to recognize the accident vs. non-accident pattern, yielding an accuracy of 0.8730, that according to authors shows a promising high performance on the predictions obtained by the model.

Pandhare and Shah (2017) employed a support vector machine classification and logistic regression classification to detect and classify road accidents and events reported online on Twitter.

Ghosh et al. (2017) presented a technique called Bayesian support vector regression, which combines support vector regression and Bayesian inference, to predict the duration of road incidents, in order to provide estimations useful in planning a real-time response to accidents.

4.4. *Artificial neural networks*

An artificial neural network is a forward-feeding neural network with one or more layers between the input and output layers. Each neuron in each layer is connected to each neuron in the adjacent layers, through lines called weight coefficients, and any change in the weight coefficients alters the function of the network, which, in summary, is the goal of the process of training the neural network, determine the values of the weight coefficients to obtain the desired output. The training or test vectors are presented to the input layer and processed by the hidden layers and the output layer (Kumar and Sahoo, 2012). The training process of an MLP consists in receiving the data in the input layer (first layer), which is assigned values of weight coefficients, which initially are random numerical values and are entered in the neurons of the second layer and propagated sequentially to the next layers.

Artificial neural networks are used for the prediction of conditions that cause accidents, to forecast road accidents and to propose accident severity models. To model a suitable ANN to approach the aforementioned problems implies the following steps: i) to define the independent and dependent variables inside the data set (Moghaddam et al., 2011); ii) to select of the type of ANN and its parameters, such as activation function, most suitable for the problem (Halim et al., 2016; Kunt et al., 2011); iii) to select the type of ANN architecture, by incorporating in the design the definition of the number of layers and neurons that will integrate the model; and iv) the definition of the performance and accuracy measures.

Çodur and Tortum (2015) proposed a model for highway accident prediction, based on an artificial neural network, taking as an input to the model, not only the data related to the accident such as driver, vehicle, time and hour, but including a highly detailed information about the road geometry and statistics regarding road traffic and volume. The performance of the proposed ANN was evaluated, reporting a correlation of 0.991, an *R*-squared value of 0.9824, a mean square error (MSE) of 4.115 and root mean square error (RMSE) of 2.0274. The authors found that the variable degree of vertical of curvature of the road was the most important parameter affecting the number of accidents on the analyzed highways.

Alkheder et al. (2017) modeled an artificial neural network to predict road accident severity, by preprocessing the road accident data using *K*-means clustering to classify the data and improve the prediction accuracy. The authors employed an ordered probit model to validate the results obtained, concluding that the ANN yield a better accuracy, reporting an accuracy value of 0.7460, over the 0.5990 obtained by the ordered probit model.

4.5. *Deep learning*

Deep learning architectures such as convolutional neural networks (CNN) and recurrent neural networks (RNN) and combination of both are used to discover hidden relationships and structures in high dimensional data, i.e., complex classification problems, text categorization, computer vision, image processing and speech recognition. CNN have a similar architecture to feed-forward artificial neural network, but they diverge in terms of i) connectivity patterns between neurons in adjacent layers, ii) the CNN reduces the parameter scale in the model by using a specialized layer called pool layer, and iii) the final layer is the only one that is fully connected. On the other hand, RNN is a neural network architecture that allows information to be propagated from one layer to another, by using hidden layers.

Ren et al. (2017) proposed a method to predict the risk of traffic accident, based on a deep learning method. The authors analyzed spatial and temporal data of traffic accident from Beijing, between 2016 and 2017, and presented the spatiotemporal correlation of the traffic incidents. The authors proposed a long short-term memory (LSTM) architecture model to predict traffic accident risk. The model comprises two input layers, the first consisting on the sequence of traffic accident frequency and the second input contains the coordinates of the region that is expected to predict. The LSTM model was compared to other baseline models, such as LASSO, support vector regression, decision tree regression and autoregressive moving average model, obtaining more accurate results, reporting a mean absolute error value of 0.014, a mean square error of 0.001 and a root mean squared error of 0.0340.

Zheng et al. (2019) presented a method for traffic accident severity prediction, by encoding the matrix of accident data in to a grey image that represented the weights of the traffic accident's features, and then, the grey images were used as an input for a severity prediction convolutional neural network. The results showed that the severity prediction using a convolutional neural network obtained better performance (precision and recall) than other methods such as *K*-nearest neighbor algorithm, Naive Bayes classifier, gradient boosting, support vector machines, neural network and long short-term memory recurrent neural network. The results reported for the prediction of slight severity accidents were an average precision value of 0.893 and an average recall value of 0.932, for the prediction of serious severity accidents were an average precision of 0.248 and an average recall of 0.167 and for fatal traffic accidents prediction an average precision of 0.063 and an average recall of 0.063.

Zhang et al. (2018) employed deep learning to detect traffic accidents from social media, using one year of Twitter content from Northern Virginia and New York City. The authors compared the performance of deep belief network (DBN) model against a long short-term memory (LSTM), an ANN with one hidden layer, a support vector machine (SVM) and supervised latent Dirichlet allocation (sLDA), reporting that the DBN model had an overall accuracy of 0.850, outperforming the other methods. The results were validated against road accident data from 15,000 loop detectors and accident logs from traffic authorities, finding that some accidents were reported in Twitter and were not documented by the police.

In the paper of Dabiri and Heaslip (2019), convolutional neural networks (CNN), long short-term memory (LSTM) and a combination of both models, called CNN + LSTM model were employed to detect traffic events, including accidents, using a labeled dataset build of traffic related information extracted from Twitter information. The authors obtained better results with the CNN model, when compared with other baseline models, such as the models proposed by D'Andrea et al. (2015) and Gu et al. (2016). The authors reported an accuracy value of 0.986 and an *F*-measure of 0.986.

Chen et al. (2018) collected 7 months of traffic accidents and 1.6 million GPS traces from Tokyo, in order to perform a study of the correlation of the human mobility and traffic accident risk. The authors develop a stack denoise autoencoder architecture to train a general deep learning model for real time simulation and prediction of traffic accident risk, obtaining better results than the models build employing decision trees, logistic regression and support vector machine algorithms, reporting a MAE value of 0.96, a mean relative error (MRE) value of 0.39 and a RMSE of 1.00. The authors acknowledge that there were limitations to their study and their results can be improved by the incorporation of geographical points of interest to their analysis.

5. Discussion

The researches reviewed, were limited by the lack of incorporation of other relevant factors and variables, such as traffic flow, human mobility and special events that can affect traffic and accident risk, i.e., massive events. Furthermore, in order to provide an effective forecast and analysis, the models output was coarse-grained, using data that comprise in spatial variables, road segments or city grids, and in temporal terms, day, or hours, that cannot be disaggregated. The results lacking predictions and analysis that can provide road segment level results and temporal analysis that cannot be drill down to minutes.

Considering the analytic methods for road accident analysis, the classification algorithms and decision trees are widely employed by their interpretability, but, in the other hand, they do not offer results with such high levels of precision and accuracy compared to other methods. Because of this, it can be considered that the approach proposed by Tiwari et al. (2017), as shown in Table 4, is valuable, since their research obtain better results by using clustering algorithms to preprocess the data set, in this particular case, hierarchical clustering and *K*-modes clustering were evaluated. The results obtained improved the performance of the classifiers methods. Regarding the natural language processing of social media information related to traffic accidents, the work by D'Andrea et al. (2015) was innovative and a baseline model to other authors, since their model was compared against other algorithms, such as Naive Bayes classifier, C4.5 decision tree, *K*-nearest neighbor (KNN) and PART classifier, and their model was put to test in task of classification of real-time twitter streams, with successful results.

Regarding road accident forecasting, as shown in Table 5, deep learning architectures, usually employed in the fields of signal and image processing, shows promising results to identify, analyze and forecast traffic accidents. The drawback of deep learning algorithms is their elevated computational requirements and the need of extensive data sets that can be subject to the possibility of produce over fitting models. The model proposed by (Ren et al., 2017) can be considered a baseline model for predicting traffic accident risk, since it incorporates big traffic accident data, as called by the authors, and proposed a novel deep learning architecture based on LSTM to predict the risk with accurate results. It can be remarked the novel approach proposed to model the data, using an encoding matrix that represents the spatial–temporal frequency of traffic accidents. Furthermore, the encoding matrix was developed using a heat map, which allowed visually highlighting the space-time zones with the highest road accident frequency values.

Table 4 – Representative results on road accident forecast methods.

Road accident analytic method	Author	Metric	Best result
Clustering algorithms	Cao et al. (2015), Kumar and Toshniwal (2016a,b), Moriya et al. (2018)	Bayesian information criterion (BIC), Akaike information criterion (AIC)	Moriya et al. (2018) with minimum values of AIC and BIC at 3 clusters
Classification algorithms and decision trees	Taamneh et al. (2017), Tiwari et al. (2017), Scott-Parker and Oviedo-Trespalacios (2017)	Accuracy, precision, recall and F-measure, using receiver operating characteristic (ROC) curve	Tiwari et al. (2017), with an accuracy of 0.8235
Natural language processing	Gu et al. (2016), D'Andrea et al. (2015), Salas et al. (2018)	Accuracy, precision, recall and F-measure, using receiver operating characteristic (ROC) curve	D'Andrea et al. (2015) reported and accuracy value of 0.9575

Table 5 – Representative metrics and results on road accident forecast.

Road accident forecasting method	Author	Metric	Best result
Bayesian networks	Castro and Kim (2016), Ghosh et al. (2017), Yuan and Abdel-Aty (2018)	Accuracy, precision, recall and F-measure, using receiver operating characteristic (ROC) curve	Castro and Kim (2016), accuracy is 0.8159, precision is 0.7239, recall value is 0.7239, F-measure is 0.723
Genetic algorithms and evolutionary computing	Kunt et al. (2011), Hashmienejad and Hasheminejad (2017)	Accuracy, precision, recall and F-measure	Hashmienejad and Hasheminejad (2017), precision is 0.885, recall is 0.889, accuracy is 0.8820, *F*-measure is 0.8875
Support vector machines	Xiong et al. (2017), Pandhare and Shah (2017)	Accuracy, precision, recall and F-measure	Xiong et al. (2017), accuracy is 0.8730
Artificial neural networks	Çodur and Tortum (2015), Alkheder et al. (2017)	Accuracy measure correlation, R-squared, MSE and RMSE	Alkheder et al. (2017), accuracy is 0.7460
Deep learning	Halim et al. (2016), Zheng et al. (2019) Zhang et al. (2018), Dabiri and Heaslip (2019), Chen et al. (2018), Ren et al. (2017)	Mean absolute error (MAE), mean relative error (MRE), mean squared error (MSE) and root mean squared error (RMSE)	Ren et al. (2017), MAE is 0.014, MSE is 0.001, RMSE is 0.0340

6. Conclusions and future challenges

This paper presented a review of algorithms and models used to analyze, characterize, and forecast road accidents. The algorithms and models comprise data mining and machine learning techniques, presented in the literature from the year 2015, for the sake of including emerging research approaches. Within the techniques and algorithms reviewed, neural networks can be highlighted for their accuracy to classify data, while deep learning methods are a relatively novel approach to predict road traffic accidents with high precision and incorporating the use of the fusion of several data sources, such as traffic accident reports, land use, road infrastructure, weather conditions and demographic information.

It could be stated that the best results are obtained when two or more analytic techniques are combined, in such a way that analysis of the obtained results is strengthened (Park et al., 2016). For instance, the use of a neural network combined with a rule system (Castro and Kim, 2016) or a genetic algorithm paired with a rule system and decision tree (Hashmienejad and Hasheminejad, 2017), shows improvement in the prediction reliability and precision of the results, when comparing with the results obtained by the execution of a single technique. Among the future challenges in road traffic forecasting lies the enhancement of the scope of the proposed models and predictions by the incorporation of heterogeneous data sources, that include geo spatial data, information from traffic volume, traffic statistics, video, sound and text and sentiment from social media, that many authors concur that can improve the precision and accuracy of the analysis and predictions. It can be acknowledged that the modeling of road accidents and the proposal of forecast models is difficult and limited by the inherent complexity of the subject, taking into account that there is a vast element of human behavior and psychology that cannot be effectively replicated in the currently available datasets, techniques and algorithms. To find a way to overcome this limitation, researches such as the study of Salas et al. (2017) seek to integrate the sentiment (positive or negative) and stress analysis, extracted from social media data, in order to enhance the detection of traffic events. Concerning the human behavior and external factors that can influence it, the research conducted by Lu et al. (2018) incorporated the adverse weather data and the sentiment related to it, in order to be integrated in city-level traffic safety and alerting system.

Conflict of interest

The authors do not have any conflict of interest with other entities or researchers.

Acknowledgments

This study was supported by the Universidad Nacional de Colombia, funding call “Convocatoria Nacional Para el Apoyo a Proyectos de Investigación y Creación Artística de la Universidad Nacional De Colombia 2017–2018”, project code 41614.

REFERENCES

Afzaal, M., Nazir, N., Akbar, K., et al., 2018. Real time traffic incident detection by using twitter stream analysis. In: Human Systems Engineering and Design: Future Trends and Applications. Springer International Publishing, Berlin, pp. 620–626.

Ahmed, M.M., Abdel-Aty, M.A., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Transactions on Intelligent Transportation Systems 13 (2), 459–468.

Ait-Mlouk, A., Agouti, T., 2019. DM-MCDA: a web-based platform for data mining and multiple criteria decision analysis: a case study on road accident. SoftwareX 10, 100323.

Ait-Mlouk, A., Gharnati, F., Agouti, T., 2017. An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety. European Transport Research Review 9 (3), 40–46.

Akaike, H., 1987. Factor analysis and AIC. Psychometrika 52 (3), 317–332.

Alkheder, S., Taamneh, M., Taamneh, S., 2017. Severity prediction of traffic accident using an artificial neural network. Journal of Forecasting 36, 100–108.

Amin-Naseri, M., Chakraborty, P., Sharma, A., et al., 2018. Evaluating the reliability, coverage, and added value of crowdsourced traffic incident reports from Waze. Transportation Research Record 2672, 34–43.

Anantharam, P., Barnaghi, P., Thirunarayan, K., et al., 2015. Extracting city traffic events from social streams. ACM Transactions on Intelligent Systems and Technology 6 (4), 1–27.

Australian Census, 2019. Australia's Regional Open Data Census - Open Data - Traffic Accidents. Available at: http://australia.census.okfn.org/dataset/traffic-accidents (Accessed 15 July 2019).

Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accident Analysis and Prevention 122, 239–254.

Bao, J., Liu, P., Yu, H., et al., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accident Analysis and Prevention 106, 358–369.

Beshah, T., Ejigu, D., Abraham, A., et al., 2011. Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: implications for improving road safety. In: The 2011 World Congress on Information and Communication Technologies (WICT 2011), Mumbai, 2011.

Breslow, L.A., Aha, D.W., 1997. Simplifying Decision Trees: A Survey. Navy Center for Applied Research in Artificial Intelligence, Washington DC.

Cao, G., Michelini, J., Grigoriadis, K., et al., 2015. Cluster-based correlation of severe braking events with time and location. Journal of Intelligent Transportation Systems 20 (6), 187–192.

Castro, Y., Kim, Y.J., 2016. Data mining on road safety: factor assessment on vehicle accidents using classification models. International Journal of Crashworthiness 21 (2), 104–111.

Chen, C., Fan, X., Zheng, C., et al., 2018. SDCAE: stack denoising convolutional autoencoder model for accident risk prediction via traffic big data. In: The 6th International Conference on Advanced Cloud and Big Data (CBD 2018), Lanzhou, 2018.

Chen, Q., Song, X., Yamada, H., et al., 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. In: The 30th AAAI Conference on Artificial Intelligence (AAAI 2016), Phoenix, 2016.

Chen, Y., Lv, Y., Wang, X., et al., 2017. A convolutional neural network for traffic information sensing from social media text. In: IEEE Conference on Intelligent Transportation Systems, Yokohama, 2018.

Çodur, M.Y., Tortum, A., 2015. An artificial neural network model for highway accident prediction: a case study of Erzurum, Turkey. Promet-Traffic-Transportation 27 (3), 217–225.

D'Andrea, E., Ducange, P., Lazzerini, B., et al., 2015. Real-time detection of traffic from twitter stream analysis. IEEE Transactions on Intelligent Transportation Systems 16 (4), 2269–2283.

Dabiri, S., Heaslip, K., 2019. Developing a twitter-based traffic event detection model using deep learning architectures. Expert Systems with Applications 118, 425–439.

Dacos Gov Co, 2019. Datos.Gov.Co - Colombia - Open Data Catalog- Road Accidents. Available at: https://www.datos.gov.co/browse?qRegistronacionaldeaccidentesdetransitosortByrelevance (Accessed 15 July 2019).

Data Gov, 2019. Data.Gov - United States - Open Data Catalog - Traffic Accidents. Available at: https://catalog.data.gov/dataset?q=traffic+accidents&sort=views_recent+desc&tags=crash&as_sfid=AAAAAAXHjZkDY7gFA5iMx_28NUE0FLt7GCD6A_wjSzainkj_rspLB-fqUew5h3LiHfKwq25Q1jllDf64k8tuEJ03xVdCKo4_qW6HRpHe_XBlCPYQhLUOwC0CkWT-WHXEHYKSTII%3D&as_fid=be93db12e7584b (Accessed 15 July 2019).

Data. Gov. UK, 2019. Data.Gov.UK - United Kingdom - Open Data Catalog - Traffic Accidents. Available at: https://data.gov.uk/search?q=traffic+accidents (Accessed 15 July 2019).

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention 38 (3), 434–444.

De Oña, J., López, G., Mujalli, R., et al., 2013. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks. Accident Analysis and Prevention 51, 1–10.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40 (4), 1257–1266.

Fan, X., He, B., Brézillon, P., 2017. Context-aware big data analytics and visualization for city-wide traffic accidents. In: International and Intedisciplinary Conference on Modeling and Using Contest, Paris, 2017.

Fan, X., He, B., Wang, C., et al., 2015. Big data analytics and visualization with spatio-temporal correlations for traffic accidents. In: The 15th International Conference on Algorithms and Architectures for Parallel Processing, Zhangjiajie, 2015.

Fawcett, L., Thorpe, N., Matthews, J., et al., 2017. A novel Bayesian hierarchical model for road safety hotspot prediction. Accident Analysis and Prevention 99 (Part A), 262–271.

Fraley, C., Raftery, A.E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal 41 (8), 578–588.

Gey, S., Nedelec, E., 2005. Model selection for CART regression trees. IEEE Transactions on Information Theory 51 (2), 658–670.

Ghosh, B., Asif, M.T., Dauwels, J., 2017. Bayesian prediction of the duration of non-recurring road incidents. In: IEEE Region 10 Annual International Conference, Singapore, 2017.

Gu, Y., Qian, Z., Chen, F., 2016. From twitter to detector: real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies 67, 321–342.

Gutierrez-Osorio, C., Pedraza, C.A., 2019. Characterizing road accidents in urban areas of Bogota (Colombia): a data science approach. In: The 2nd Latin American Conference on Intelligent Transportation Systems (ITS LATAM 2019), Bogota, 2019.

Halim, Z., Kalsoom, R., Bashir, S., et al., 2016. Artificial intelligence techniques for driving safety and vehicle crash prediction. Artificial Intelligence Review 46 (3), 351–387.

Hashmienejad, S.H., Hasheminejad, S.M.H., 2017. Traffic accident severity prediction using a novel multi-objective genetic algorithm. International Journal of Crashworthiness 22 (4), 425–440.

Jain, A.K., Murty, M.N., Flynn, P.J., 2000. Data clustering: a review. ACM Computing Survey 31 (3), 264–323.

Janssen, M., Charalabidis, Y., Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of open data and open government. Information Systems Management 29 (4), 258–268.

Kaplan, S., Prato, C.G., 2013. Cyclist-motorist crash patterns in Denmark: a latent class clustering approach. Traffic Injury Prevention 14 (7), 725–733.

Kononen, D.W., Carol, A.C.F., Wang, S.C., 2011. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Analysis and Prevention 43 (1), 112–122.

Krishnaveni, S., Hemalatha, M., 2011. A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications 23 (7), 40–48.

Kumar, S., Toshniwal, D., 2016a. A data mining approach to characterize road accident locations. Journal of Modern Transportation 24 (1), 62–72.

Kumar, S., Toshniwal, D., 2015a. A data mining framework to analyze road accident data. Journal of Big Data 2, https://doi.org/10.1186/s40537-015-0035-y.

Kumar, S., Toshniwal, D., 2015b. Analyzing road accident data using association rule mining. In: 2015 International Conference on Computing, Communication and Security (ICCCS), Pamplemousses, 2015.

Kumar, S., Toshniwal, D., 2016b. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). Journal of Big Data 3, https://doi.org/10.1186/s40537-016-0046-3.

Kumar, Y., Sahoo, G., 2012. Analysis of Bayes, neural network and tree classifier of classification technique in data mining using WEKA. Computer Science and Information Technology (CS & IT) 2, 359–369.

Kunt, M.M., Aghayan, I., Noii, N., 2011. Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. Transport 26 (4), 353–366.

Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accident Analysis and Prevention 135, 105371.

Lin, L., Wang, Q., Sadek, A.W., 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies 55, 444–459.

Lu, W., Luo, D., Yan, M., 2017. A model of traffic accident prediction based on convolutional neural network. In: 2017 2nd IEEE International Conference on Intelligent Transportation Engineering, Beijing, 2017.

Lu, H., Zhu, Y., Shi, K., et al., 2018. Using Adverse weather data in social media to assist with city-level traffic situation awareness and alerting. Applied Sciences 8 (7), https://doi.org/10.3390/app8071193.

Ma, J., Kockelman, K., 2006. Crash modeling using clustered data from Washington state: prediction of optimal speed limits. In: IEEE Intelligent Transportation Systems Conference, Toronto, 2006.

Moghaddam, F.R., Afandizadeh, S., Ziyadi, M., 2011. Prediction of accident severity using artificial neural networks. International Journal of Civil Engineering 9 (1), 41–49.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., et al., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. Safety Science 54, 27—37.

Moriya, K., Matsushima, S., Yamanishi, K., 2018. Traffic risk mining from heterogeneous road statistics. IEEE Transactions on Intelligent Transportation Systems 19 (11), 3662—3675.

Nguyen, H., Liu, W., Pivera, P., et al., 2016. TrafficWatch: real-time traffic incident detection and monitoring using social media. In: The 20th Pacific-Asia Conference on Know Ledge Discovery and Data Mining (PAKDD 2016), Auckland, 2016.

Ozbayoglu, A.M., Kucukayan, G., Dogdu, E., 2016. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. In: 2016 IEEE International Conference on Big Data, Washington DC, 2016.

Pandhare, K.R., Shah, M.A., 2017. Real time road traffic event detection using twitter and spark. In: International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2017.

Park, S., Kim, S., Ha, Y., 2016. Highway traffic accident prediction using VDS big data analysis. The Journal of Supercomputing 72 (7), 2815—2831.

Powers, D.M.W., 2011. Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. Flinders University, Adelaide.

Ren, H., Song, Y., Wang, J., et al., 2017. A deep learning approach to the citywide traffic accident risk prediction. In: The 21st International Conference on Intelligent Transportation System (ITSC), Maui, 2017.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. Accident Analysis and Prevention 79, 198—211.

Salas, A., Georgakis, P., Nwagboso, C., et al., 2017. Traffic event detection framework using social media. In: IEEE International Conference on Smart Grid and Smart Cities, Singapore, 2017.

Salas, A., Georgakis, P., Petalas, Y., 2018. Incident detection using data from social media. In: IEEE Conference on Intelligent Transportation Systems, Yokohama, 2018.

Scott-Parker, B., Oviedo-Trespalacios, O., 2017. Young driver risky behaviour and predictors of crash risk in Australia, New Zealand and Colombia: same but different? Accident Analysis and Prevention 99 (Part A), 30—38.

Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transportation Research Part C: Emerging Technologies 58, 380—394.

Sinnott, R.O., Yin, S., 2015. Accident black spot identification and verification through social media. In: 2015 IEEE International Conference on Data Science and Data Intensive Systems, Sydney, 2015.

Sun, H., McIntosh, S., 2016. Big data mobile services for New York City taxi riders and drivers. In: 2016 IEEE International Conference on Mobile Services, San Francisco, 2016.

Taamneh, M., Alkheder, S., Taamneh, S., 2017. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. Journal of Transportation Safety and Security 9 (2), 146—166.

Tiwari, P., Dao, H., Nguyen, N.G., 2017. Performance evaluation of lazy, decision tree classifier and multilayer perceptron on traffic accident analysis. Informatica 41 (1), 39—46.

The Global Goals, 2017. The Global Goals for Sustainable Development. Available at: http://www.globalgoals.org (Accessed 15 August 2017).

Veljković, N., Bogdanović-Dinić, S., Stoimenov, L., 2014. Benchmarking open government: an open data perspective. Government Information Quarterly 31 (2), 278—290.

Wang, D., Al-Rubaie, A., Clarke, S.S., et al., 2017. Real-time traffic event detection from social media. ACM Transactions on Internet Technology 18 (23), 1—23.

Wang, J., Luo, T., Fu, T., 2019. Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach. Accident Analysis and Prevention 133 (S), 105320.

Wang, S., Li, F., Stenneth, L., et al., 2016. Enhancing traffic congestion estimation with social media by coupled hidden markov model. In: ECML PKDD: Joint European Conference on Machine Leaning and Knowledge Discovery in Databases, Riva del Garda, 2016.

World Health Organization, 2015. Global Status Report on Road Safety, 2015. World Health Organization, Geneva, 2015.

Wu, X., Kumar, V., Quinlan, J.R., 2008. Top 10 algorithms in data mining. Knowledge and Information System 14, 1—37.

Xiong, X., Chen, L., Liang, J., 2017. A new framework of vehicle collision prediction by combining SVM and HMM. IEEE Transactions on Intelligent Transportation System 19 (3), 1—12.

You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. Journal of Modern Transportation 25 (2), 116—123.

Yuan, J., Abdel-Aty, M., 2018. Approach-level real-time crash risk analysis for signalized intersections. Accident Analysis and Prevention 119, 274—289.

Yuan, Z., Zhou, X., Yang, T., 2017. Predicting traffic accidents through heterogeneous urban data: a case study. In: The 6th ACM KDD International Workshop on Urban Computing (UrbComp 2017), Halifax, 2017.

Zhang, S., Wu, X., 2011. Fundamentals of association rules in data mining and knowledge discovery. WIREs Data Mining and Knowledge Discovery 1 (2), 97—116.

Zhang, Z., He, Q., Gao, J., et al., 2018. A deep learning approach for detecting traffic accidents from social media data. Transportation Research Part C: Emerging Technologies 86, 580—596.

Zheng, M., Li, T., Zhu, R., et al., 2019. Traffic accident's severity prediction: a deep-learning approach-based CNN network. IEEE Access 7, 39897—39910.

Zheng, Y., Wang, J., Li, X., 2014. Driving risk assessment using cluster analysis based on naturalistic driving data. In: The 17th International IEEE Conference on Intelligent Transportation Systems (ITSC14), Qingdao, 2014, pp. 2584—2589.

Camilo Gutierrez-Osorio graduated in computer science from the Universidad de Antioquia (Colombia) and received his MSc degree from the Universidad de Murcia (Spain). Currently he is a PhD student in the Universidad Nacional de Colombia. His current research focuses on machine learning and intelligent transportation systems.

César Pedraza received the PhD degree in informatics engineering from the Rey Juan Carlos University in Madrid, Spain. And he received a MSc degree from the Universidad de Los Andes, Bogota Colombia. He is currently a professor of parallel computing and operating systems at the Universidad Nacional de Colombia. His current researches are bio-inspired algorithms, intelligent transportation systems and parallel computing.