computers

MDPI

Article

Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data

Camilo Gutierrez-Osorio *, Fabio A. González and Cesar Augusto Pedraza

Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Cra 45 No. 26-85, Bogotá 111321, Colombia

* Correspondence: cgutierrez@unal.edu.co

Abstract: Traffic accidents are a major concern worldwide, since they have a significant impact on people's safety, health, and well-being, and thus, they constitute an important field of research on the use of state-of-the-art techniques and algorithms to analyze and predict them. The study of traffic accidents has been conducted using the information published by traffic entities and road police forces, but thanks to the ubiquity and availability of social media platforms, it is possible to have detailed and real-time information about road accidents in a given region, which allows for detailed studies that include unrecorded road accident events. The focus of this paper is to propose a model to predict traffic accidents using information gathered from social media and open data, applying an ensemble Deep Learning Model, composed of Gated Recurrent Units and Convolutional Neural Networks. The results obtained are compared with baseline algorithms and results published by other researchers. The results show promising outcomes, indicating that in the context of the problem, the proposed ensemble Deep Learning model outperforms the baseline algorithms and other Deep Learning models reported by literature. The information provided by the model can be valuable for traffic control agencies to plan road accident prevention activities.

Keywords: machine learning; traffic accident risk prediction; traffic accidents

Citation: Gutierrez-Osorio, C.; González, F.A.; Pedraza, C.A. Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data. *Computers* **2022**, *11*, 126. https://doi.org/10.3390/computers11090126

Academic Editor: Paolo Bellavista

Received: 20 June 2022
Accepted: 29 July 2022
Published: 23 August 2022

1. Introduction

Traffic accidents are a major issue concerning the number of deaths, personal injury, and property damage. According to the World Health Organization (WHO) [1], 1.35 million people die each year as a result of road traffic crashes, 93% of the world's fatalities on the roads occur in low and middle-income countries, and as the year 2018, road accident injury is the leading cause of death for children and young adults aged 5 to 29 years. The call to governments made by the WHO was to improve their legislation on the key issues that directly affect the improvement of road safety, these being the control of speed on the road, driving while intoxicated, the use of helmets on motorcycles and mandatory use of seat belts and special seats for children. Another proposed road safety strategy was the improvement in the planning, design, and operation of roads, by configuring a star rating tool for road networks, as recommended by the International Road Assessment Program (iRAP) [2]. This strategy had an impact not only on the safety and wellbeing of drivers but also on other road users such as pedestrians and cyclists.

Given the precedents mentioned above, it is valuable to seek the application of machine learning methods for the prediction of road accidents, since it is useful not only to the public but to road users, transportation planners and governments. Considering the accessibility to the data of social media platforms that contain spatio-temporal data related to road accidents and the availability of Deep Learning algorithms suitable to analyze that kind of data, it is feasible to propose a road accident prediction model, which allowed us to integrate several data sources, extract spatio-temporal features and learn from time-series data and obtain meaningful patterns. On the other hand, we face the following issues when

designing the model: (i) Data scarcity. The proportion of road accidents is low against the case of no-accidents in a time series dataset; (ii) Data quality. Social media data may contain a high proportion of outlier data concerning coordinates and may contain duplicate reports of the same road accident; (iii) The inherent behavior of road accidents is complex and non-linear [3]; to tackle that obstacle the proposed model takes into account environmental conditions that have an impact on road accidents.

This paper presents a deep learning model that fuses different information sources to predict road accidents in the city of Bogota, using a dataset that contained road accident reports from May 2018 to June 2019 and climate information for the aforementioned time period and employing an ensembled deep Recurrent Neural Network (RNN). The road accident dataset was processed in order to remove outliers and noise and then it was transformed using a feature engineering process in order to obtain a representation of the data that can be useful to perform a machine learning process. Additionally, data from climatological information of the city of Bogota were used in order to enrich the scope of the study and enhance the prediction of road accidents. Regarding the proposed ensemble deep learning model, its design comprises a Gated Recurrent Network and Convolutional Neural Network architecture, devised to analyze road accident spatio-temporal data, road accident time patterns and climatological data. The results obtained by the model were compared against baseline models and with deep learning models reported by the literature.

The main contributions of this paper are summarized as follows:

1. The proposed ensemble deep learning model is designed to perform traffic accident prediction, using information about road accidents from social media and climate information from open data.
2. To our knowledge, this is the first research of this nature conducted in Bogota city using social media data.
3. The information provided by the model can be valuable for traffic control agencies to plan road accident prevention activities since it results can be applied to specific regions in the city.

The rest of the paper was organized as follows: Section 2 discusses the related work to this paper. Section 3 presents the methodology used, detailing the data analyzed, the data cleaning process, the feature engineering process designed to enhance the data, the deep learning methods employed and the design of the proposed model. Section 4 presents the results obtained and the comparison with baseline algorithms and deep learning models reported by the literature. Section 5 describes the conclusion and future work.

2. Related Work

Social media had become one of the main channels for public announcements and interaction, and among all the universe of topics that people discussed, one of the most relevant is the information about traffic status, road conditions, traffic accidents and other factors that may have an impact such as weather [4], and currently, the most studied and available data sources for road accident analysis and prediction are Waze, Inrix, Google Maps, Twitter and Sina Weibo, as reported by [5–8]. One of the most challenging tasks to carry out when working with social media information is the process of preparing the data to transform it into a source of structured information with adequate quality for analysis using machine learning or another technique as described by [9–12]. After transforming the raw data and obtaining a suitable data set, several research studies had been carried out, such as the detect traffic accidents reported on Twitter using deep learning algorithms in New York [13], a real-time monitoring system for traffic event detection [9], or a real-time road accident detector using Twitter [14].

The use of deep learning algorithms is a state-of-the-art technique that is useful to discover patterns and structures in high-dimensional data and generate learning patterns and discover relationships beyond immediate neighbors in the data [15]. Deep learning has been used in fields such as computer vision, natural language processing, signal processing and speech recognition, among other applications [16]. Regarding applications of deep

learning in the subject of urban road accidents prediction, there are several studies that are worth highlighting, such as the one carried out by [17], where the authors analyzed spatial and temporal data of traffic accidents from Beijing, between 2016 and 2017 using a Long Short-Term Memory model, and the research by [18], where the authors employed a novel approach by encoding a matrix of accident data into a grey image array that represented the weights of the traffic accident's features and used as an input for road accident severity prediction employing a Convolutional Neural Network model. Other innovative uses of state-of-the-art Machine Learning techniques related to transport are the prediction and identification of cargo theft in railway transport [19] and the prediction of pantograph failures of rail vehicles [20].

Concerning ensembled deep learning architectures applied to road accident analysis and prediction, Ref. [21] proposed a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory network (LSTM), called the CNN + LSTM model, that was employed to detect traffic events, including accidents, using a labeled dataset build of traffic-related information extracted from Twitter information. The authors in [22] proposed a method for real-time crash risk prediction on urban arterials using a combination of LSTM and CNN networks. Ref. [23] proposed a Stack Denoise Convolutional Auto-Encoder (SdAE) composed of eight hidden layers paired with a batch normalization method. [24] proposed a deep learning model called spatiotemporal convolutional long short-term memory or STCL-Net Model, which consisted of an ensembled model of LSTM layer, CNN layer and hybrid LSTM-CNN layer. The authors applied their model using several spatio-temporal configurations, using different time and spatial grid configurations to predict traffic accidents in New York City. Ref. [25] proposed a deep learning model called Deep Spatio-Temporal Graph Convolutional Network, DSTGCN, composed by a spatial layer, a spatio-temporal layer and an embedding layer.

Ensemble model architectures have been employed in other knowledge fields such as detection of speech patterns [26] and short team load forecasting [27]. The proposed model is a deep learning ensembled model designed to forecast road accidents risk, containing a GRU network to extract patterns from road accident time series, a CNN network to extract relations from road accidents related to holidays and traffic peak hours, and a CNN network designed to understand the underlying relationships between climatological conditions and road accidents. The analysis employed data from May 2018 to June 2019.

3. Materials and Methods

The proposed method comprised three main phases, as shown in Figure 1. The first phase was to perform a data cleaning and quality process, where the main objective was to deal with missing values, extreme values, and outliers. The second phase was a feature engineering process, with the objective of transforming the datasets in order to be used in Machine Learning, involving tasks such as estimating the frequency and probability of road accidents and generating time series data sets. The third phase was the prediction of road accidents, using an ensemble deep learning model, which uses as input the processed data coming from the feature engineering phase.

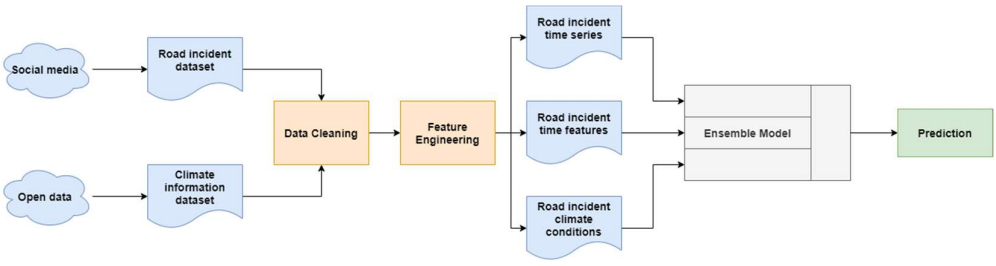


Figure 1. Overall architecture of the proposed method. The architectures detail the main phases, such as data cleaning, feature engineering and the design of the ensemble model.

The next subsections provide the description of the original data, the data quality and feature engineering process and the design and architecture of the ensemble deep learning model proposed.

3.1. Dataset Description and Data Cleaning Preprocessing

3.1.1. Road Accident Dataset

In order to collect the road accident data, a web crawler software was developed to collect data from the Waze application online live map from Bogota city. A similar approach developed to extract data from the Waze platform was developed by [28,29] to obtain information about traffic flow and traffic conditions, in Mexico City and Charlotte city (US), respectively.

The web crawler software extracted every 15 min all the information related to road accident reports, which included single-car collisions, collisions with a fixed object, vehicle fire, vehicle rollover, accidents with multiple vehicles involved and, finally any accident that involved pedestrians, bicycle, and motorcycle users. The collected data contains reports from May 2018 to June 2019 and each report consisted of time, latitude, and longitude as presented in Table 1. No information was captured regarding the vehicles involved, the severity of the accident, or the persons affected.

Table 1. Dataset extracted from Waze, example of data.

Timestamp	Latitude	Longitude
20 June 2019 15:41:33	4.682584	−74.04869
20 June 2019 15:41:33	4.681141	−74.0537
20 June 2019 15:41:57	4.71908	−74.07556
20 June 2019 15:42:22	4.651553	−74.07435
20 June 2019 15:43:19	4.694954	−74.08765
20 June 2019 15:43:23	4.676282	−74.08434
20 June 2019 15:43:44	4.629432	−74.08272

The first process that was conducted was a data cleaning and quality process. The main objective was to obtain an initial version of the dataset that was suitable for use in deep learning algorithms, considering the impact that missing values, extreme values and outliers can have on the results obtained.

It was sought to eliminate all the records that were outliers, which may contain noise and that could contain coordinates not related to Bogota city road infrastructure. With the intention of removing outlier coordinates, an official source was consulted, in this case, the Bogota city open data website that contains the information related to city road infrastructure (https://serviciosgis.catastrobogota.gov.co/arcgis/rest/services/Mapa_Referencia/Mapa_Referencia/MapServer/10 accessed on 20 February 2022). According to the information contained on the website, the values for the relevant maximum and minimum coordinates were obtained, and therefore, all the coordinates that were outside the specified range were deleted. The values for maximum and minimum reference coordinates are presented in Table 2.

Table 2. Bogota road infrastructure coordinates, as specified by Bogota open data website for maps, layer road infrastructure.

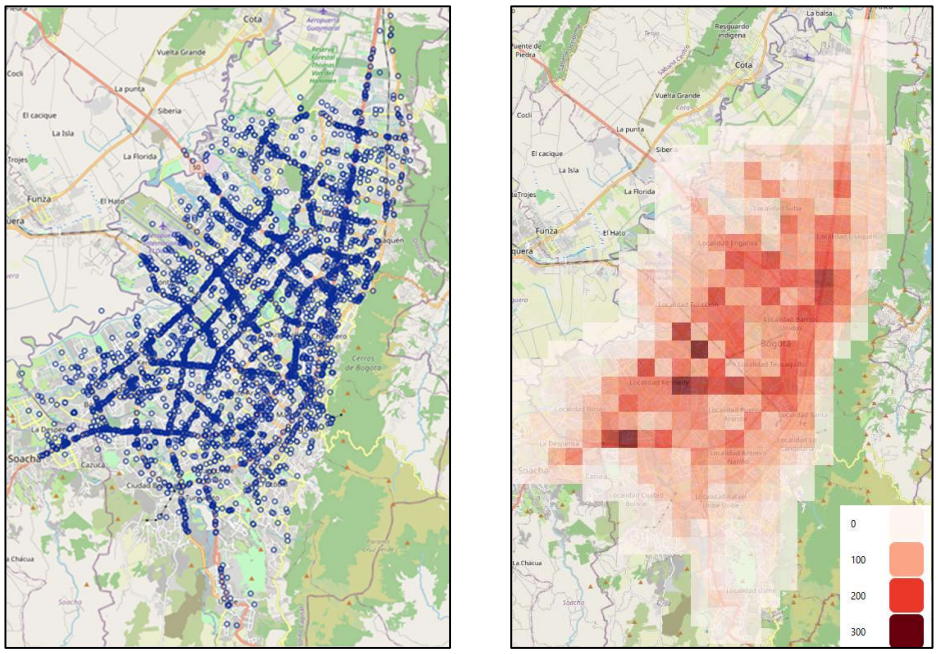
Description	Latitude	Longitude
Minimum values	3.7892	−74.3948
Maximum values	4.8366	−73.9913

The dataset was examined by searching for records with empty or null dates, null latitude, and null longitude values, and none were found. Finally, the clean version of the dataset consisted of 96,949 records, with no empty values, and no data outliers, as detailed

in Table 3. Finally, Figure 2 shows, (a) the geographical distribution of the data in Bogota city, remarking that all the reports are located inside Bogota's urban road infrastructure. Panel (b) of Figure 2 shows a heatmap of road accident distribution, to show the concentration of road accidents in areas of an approximate size of 1 km by 1 km. The heatmap uses a codification of four intervals, to show the number of road accidents, from 0 to 100, 101 to 200, 2001 to 300 and greater than 300 road accidents.

Table 3. Clean version of the road accident dataset, detailing features statistics.

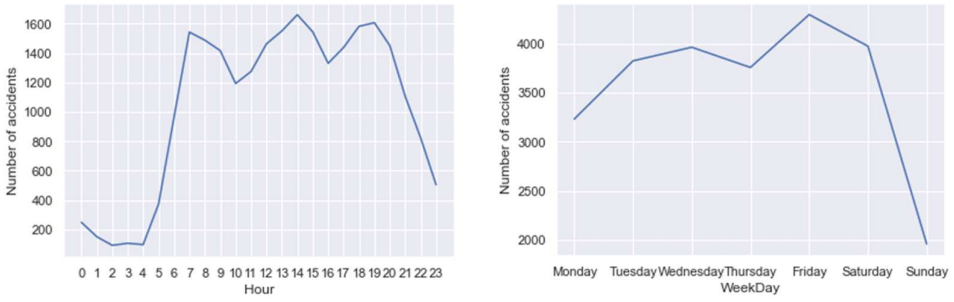
Value	Center Value	Dispersion	Min. Value	Max. Value	% Missing Values
Timestamp	2 December 2018	13 months	4 May 2018	19 June 2019	0%
Longitude	−74.1042	−0.00005	−74.2213	−74.0175	0%
Latitude	4.6557	0.01111	4.49278	4.82472	0%



(**a**) (**b**)

Figure 2. Geographical distribution of road accident data, Bogota city, May 2018 to June 2019: (**a**) Distribution of road accidents by latitude and longitude; (**b**) Heatmap of road accidents distribution, showing the spatial disparity of the distribution of road accidents.

Regarding the time patterns found in the data, they can be described using Figure 3. Figure 3a shows a clear pattern of traffic accidents that matches the pattern of traffic peak hours in Bogota, with a peak of accidents from 6 am to 8 am, a second peak between 12 to 3 pm and finally the third peak from 5 pm to 8 pm. The frequency of the peaks can be explained as matching the beginning and end of workdays and the peak in the middle is influenced by the end of the school day. Figure 3b shows the distribution of road accidents by day of the week, indicating that Wednesday and Friday are days with high numbers of road accidents. The small number of road accidents on weekends was expected since most people do not commute to work. Additionally, it is worth considering that the mobility of private vehicles and taxis is restricted by the "Pico y Placa" policy [30], which restricts the mobility of vehicles according to the date, the last figure of the vehicle plate and according to predefined time slots.



(**a**) (**b**)

Figure 3. Time patterns of road accidents in Bogota city: (**a**) Distribution of road accidents by hour of the day, showing a clear influence of peak hours; (**b**) Distribution of road accidents by day of the week, showing an increase on Wednesday and Friday, and a clear decrease in weekends.

3.1.2. Climate Information Dataset

In order to enhance the understanding of the underlying patterns behind the road accidents, climate information for Bogota city was gathered, since the weather conditions are a known fact related to road accidents [31]. Other climate information such as foggy days, visibility and lightning were not available from the Bogota City Open Data repository (https://datosabiertos.bogota.gov.co/ accessed on 20 February 2022) or other official data sources. The authors reviewed the police road accidents reports database, which could contain such information, and the result was that all the data related to climate conditions was not filled out in those reports. The information consists of reports of precipitation in the city of Bogota, measured in millimeters and with sampling intervals of every 5 min, discriminated by each of the city's meteorological stations, specifying the location by its coordinates, as shown in Table 4. The features of the dataset were the date (including hour and minute data) of the sampling, meteorological station code and location (including code of station and its latitude and longitude) and the amount of rain in millimeters. The information was gathered from the Colombian government's official open data site (https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg accessed on 15 October 2021). The content of the data is maintained and steward by Colombia's meteorological agency, "Instituto de Hidrología, Meteorología y Estudios Ambientales" (http://www.ideam.gov.co/ accessed on 15 October 2021) IDEAM.

Table 4. Precipitation information dataset for Bogota city, detailing features statistics.

Value	Center Value	Dispersion	Min. Value	Max. Value	% Missing Values
Date	12 January 2019	13 months	4 May 2018	19 June 2019	0%
Longitude	−74.1034	−0.0006	−74.2050	−74.0190	0%
Latitude	4.6630	0.0175	4.5120	4.8130	0%
Rain (mm)	0.1144	8.1452	0.0000	48.7000	0%

3.2. Feature Engineering

Feature engineering is the process to design the preprocessing pipelines and data transformations that result in the representation of the data that can be employed in machine learning algorithms [15]. The application of the feature engineering process used in this research is described in the following subsections.

3.2.1. Definition of Time Window

In order to manage the multiple instances of the same road accident reported by more than one person, the first step was to ensure that all the timestamps were converted using a 60-min window, by using the criteria that all the minute part was set to 00, i.e., the date 15 May 2019 07:38:00 was transformed to 15 May 2019 07:00:00. A multiples report can be

considered as a group of road accidents that had the same latitude, longitude and belong to the same 60-min window, and therefore only one report was considered using that grouping criterion. This decision seeks to consider that there may be reports of the same road accident made by multiple witnesses at the same time window. Other time periods were considered for the time window, such as 15 and 30 min, but they were not practical since they resulted in a small sample of data, not useful for calculation of traffic accident frequency and probability, making them not suitable for use in deep learning algorithms, as stated by [32].

3.2.2. Timestamp Transformation

The timestamp variable was initially represented as a datetime type, with format *yyyy-mm-dd hh:mm:ss*, and it was transformed to Linux Epoch or POSIX time (POSIX time as defined in https://pubs.opengroup.org/onlinepubs/9699919799/xrat/V4_xbd_chap04.html accessed on 15 October 2021). This representation ensures that the temporal resolution of the data was not lost and that it can be encoded back into the timestamp data type without risking losing information.

3.2.3. Definition of Spatial Matrix

The traffic accident data set was transformed using a two-dimensional matrix that represents the space variable. Two spatial resolutions were considered, a squared grid with elements of 1 km length and 1 km width and a squared grid with elements of 2 km length and 2 km width. Regarding the case of the 1 km by 1 km grid, the resulting matrix had a length (east to west) of 24 km and a width (north to south) of 42 km, and as a result, the proposed structure was a matrix of 24 by 42 elements. The 2 km by 2 km squared grid covered an area with a length of 26 km and a width of 44 km and comprised 13 by 22 elements.

The spatial matrix with 1 km by 1 km was generated using QGIS Software (version 3.16), using the following parameters:

- Geometry: Polygon
- SRC: EPSG: 3857-WGS 84/Pseudo-Mercator-Projected
- Extension: −8,262,533.5000, 496,384.8125: −8,238,532.5000,538,384.8750
- Unit: meters
- Number of objects: 1008

The selected spatial resolution dimension was the grid of 1 km and 1 km elements. This spatial resolution helps us to avoid the observation that the occurrence of traffic accidents became near zero and therefore, seeking to avoid the decrease of the performance of the predictive model, according to the results reported by [24] that iterated employing different configurations of spatial resolutions to analyze traffic accidents in New York City. Figure 4 shows the spatial grid with 24 by 42 elements, including the main administrative sub-division of the city, called localities.

Using the spatial matrix, the traffic accident probability was calculated, employing a method adapted from the proposals of [33,34], as follows:

The traffic accident probability is represented as s (X, D), with X being the number of road accidents in the dataset that occurs in a particular latitude and longitude and D the total amount of accidents on the same coordinates in that particular time.

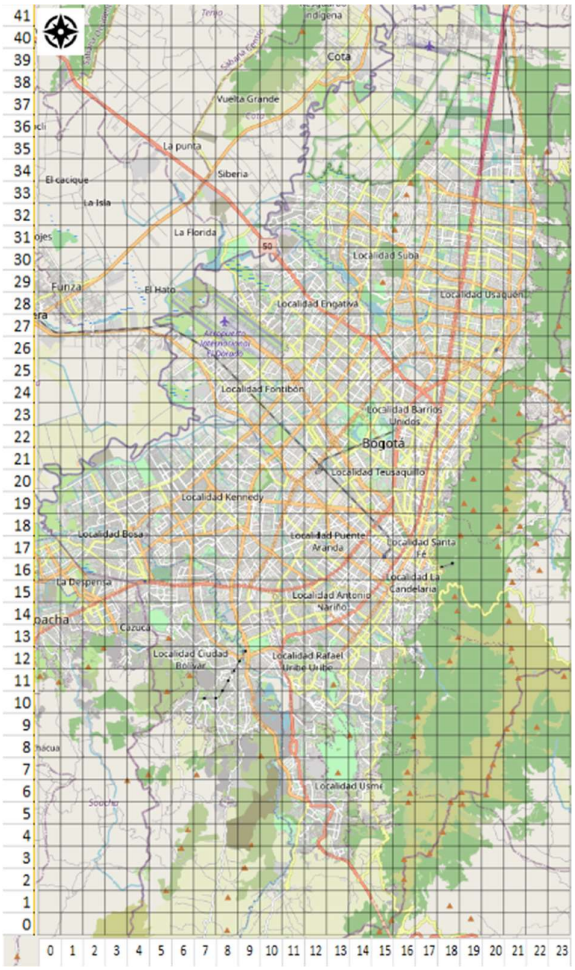


Figure 4. Spatial matrix details of Bogota city, resolution of 1 km × 1 km for each grid element.

3.2.4. Data Augmentation Procedure

In order to deal with the absence of negative reports of traffic accidents, defined data with zero frequency and zero probability in a given time and a specific coordinate, a data augmentation procedure must be employed, with the goal of reflecting the behavior and nature of traffic accidents in the real world. The method Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) [35] was evaluated since the SMOGN method is useful for regression problems when the synthetic data is of interest for the model. The results using SMOG were not useful, since its results do not ensure that the generated data belong to real world data, i.e., reports of non-accident data in the regions of study, that can be correlated to data from rain conditions reported by the official entities.

Our proposed procedure was created using an approach similar to that proposed by [36,37] and is shown in Algorithm 1. The idea behind the data augmentation procedure was to generate tuples with zero frequency and zero road accident probability in random hours and dates for a certain coordinate that was not reported in the original dataset. The procedure was executed for every one of the elements of the spatial matrix, the initial value for parameter k was the amount of traffic accidents reported for the current element x, y of the spatial matrix, meaning that it generated a zero-accident report with a set of k tuples.

Algorithm 1. Data augmentation procedure.

Input: Original Dataset
Output: Augmented dataset
Begin
For i = 1 → k
    select randomly an Epoch value from the dataset
    select randomly a X position from the grid
    select randomly a Y position from grid
    generate a tuple t_i ($time_i$, x position, y position)
    assign to the tuple t_i 0 probability and 0 frequency values
    save tuple t_i in the dataset Augmented dataset
End For
Return Augmented dataset
Delete from Augmented dataset duplicates tuples
Delete from Augmented dataset tuples that match a tuple from the original dataset
End

As a result, by using the union of the accident reports in every grid element and the non-accident report data generated for that grid element, a dataset with a proportion of 1:1 of positive reports (frequency and probability greater than zero) and negative reports (frequency and probability equal to zero) of road accidents was obtained for every grid element.

3.2.5. Resulting Set of Features

The resulting set of features obtained as output from the data engineering process is shown in Table 5. The specific features fed into every neural network of the ensemble model are explained in Sections 3.3.1–3.3.3.

Table 5. Resulting set of features obtained from data engineering process.

Feature	Description
Epoch	Date encoded as epoch
Latitude	Latitude of traffic accident report
Longitude	Latitude of traffic accident report
Hour of day	Hour of traffic accident
Weekday	Weekday of traffic accident
Rain mm	Quantity of rain fall in millimeters
Position X	Position relative to the spatial matrix in the X axis
Position Y	Position relative to the spatial matrix in the Y axis
Traffic accident probability	Traffic accident probability

3.3. Deep Learning Model

Deep learning models are Artificial Neural Networks with more than two hidden layers, that are designed to learn feature representations from data in an automatic way, rather than depend on human experience and prior knowledge [15]. Among the deep learning models, there are architectures that use models such as long-short memory networks (LSTM), gated recurrent units (GRU), convolutional neural networks (CNN) and a combination of them that are used to discover hidden relationships and structures in high dimensional data. The combination of models is called ensemble learning, and it is designed to make use of two or more machine learning models and combine their results to improve the results obtained. In the case of this research, the objective was to design a model that had the appropriate combination of Recurrent Neural Networks and Convolutional Neural Networks that were able to make predictions using data from road accidents and climatological information.

3.3.1. GRU Network for Road Accident Time Series

Gated Recurrent Unit (GRU) is a Recurrent Neural Network designed containing gating units that modulate the flow of information inside the unit, without having separate memory cells. The main difference with an LSTM network is that GRU does not have any mechanism to control the degree to which its state is exposed but exposes the whole state at each iteration without any control mechanism [38]. GRUs are suitable for learning from time-series data and sequence modeling.

The GRU network was designed to manage the information from the road accidents as a time series, where there is a time series defined for every member (x,y) of the spatio-temporal matrix defined in Section 3.2. The input of the GRU network is prepared with the shape (Sample, Time Slice, Features), meaning that the input to the network is a 3d array. The sample size was defined as 60% of the dataset for the training set and 40% for the testing set. The time window defined to group the series of data was defined as 6 h, since in a previous work regarding the characterization of road accidents in Bogota city [39], 6 h was the temporal pattern most relevant between traffic peak hours. The parameter Features was specified as 3, since the GRU network receives the grid position (x,y) and the calculated probability of road accident, as specified in Section 3.2. The output of the GRU network is the estimated probability of road accidents in the (x,y) position. The optimal parameters GRU were found by iterating all the possible combinations of parameters shown in Table 6 and selecting the best results obtained, using Mean Squared Error (MSE), Root Mean Squared Error RMSE and Mean Absolute Error (MAE) metrics.

Table 6. GRU neural network for road accident time series.

Hyperparameter	Values Evaluated	Optimal Value
GRU unit size	16, 32, 64, 128, 512	512, 128
GRU layers	1, 2, 3, 4	2

3.3.2. CNN Network for Road Accident Time Features

CNN have a similar architecture to a feed-forward artificial neural network, but they diverge in terms of (i) connectivity patterns between neurons in adjacent layers; (ii) the CNN reduce the parameter scale in the model by using a specialized layer called pool layer; (iii) CNN have a special layer called convolution, that consist in a series of filters that are convolved across the axis or dimensions of the input data or image; and (iv) the final layer is the only one that is fully connected [40]. CNN have been extensively used to extract spatial features from data and for perform object detection and semantic segmentation of high-resolution images. Regarding the architecture of CNN, there are different use cases, according to their dimensionality. One-dimensional CNN are best used for extracting features in one-dimensional data or signals such as sounds, 2-dimensional CNN are used for extracting patterns and analyzing images, grayscale or RGB images and both cases are considered to be 2-dimensional signals, and 3-dimensional CNN are used for 3-dimensional signals such as video frames, considering images as two-dimensional signals that vary during the time.

The CNN network chosen for analyzing the road accident time features is a 1-dimensional CNN or CONV1D, it was prepared to estimate the probability of road accidents in a given (x,y) position in the spatial matrix considering other time characteristics do not present in the GRU network time series, such as holidays and work hours. The input for this network was developed with the shape (Sample, Features). The sample used is 60% of the dataset for the training set and 40% for the testing set, and the features considered are X position, Y position, day of the week, hour of the day and road accident probability. The output of the CNN network is the estimated probability of a road accident in the (x,y) position. The optimal CNN network parameters were found by iterating all the possible combinations of parameters shown in Table 7 and selecting the best results obtained, using Mean Squared Error (MSE), Root Mean Squared Error RMSE and Mean Absolute Error (MAE) metrics.

Table 7. CNN network for additional time features.

Hyperparameter	Values Evaluated	Optimal Value
CNN unit size	32, 64, 128, 256, 512	512, 128, 64, 32
CNN layers	2, 3, 4, 5	4
Kernel size	1, 2, 3	1

3.3.3. CNN Network for Climate Data

The CNN network for climate data analysis is designed as a 1-dimensional CNN or CONV1D, that processes historical information on the amount of rain (in millimeters) in a given (x,y) position of the spatio-temporal matrix and correlates that information with the probability of road accidents. The input of the climate CNN network is a time series prepared with the shape (Sample, Time Slice, Features). The sample used is 60% of the dataset for the training set and 40% for the testing set, as stated previously, the time window for the data series was defined as 24 h, considering the nature of the climatological data. the features selected were X position, Y position, rain in millimeters, and road accident probability. The output of the CNN network is the estimated probability of a road accident in the (x,y) position. The optimal CNN network parameters were found by iterating all the possible combinations of parameters shown in Table 8 and selecting the best results obtained, using Mean Squared Error (MSE), Root Mean Squared Error RMSE and Mean Absolute Error (MAE) metrics.

Table 8. CNN network for climate data.

Hyperparameter	Values Evaluated	Optimal Value
CNN unit size	16, 32, 64, 128, 256	128, 128,128
CNN layers	1, 2, 3, 4	3
Kernel size	1, 2, 3	1

3.4. Ensemble Model Network

The proposed ensemble model architecture contains one GRU network for analyzing time-series road accident data, one CNN to analyze road accident additional time features and one CNN to analyze climate data related to road accidents, their output is combined by a concatenate layer and contains a dropout layer to avoid over-fitting as shown in Figure 5. The networks are processing data in parallel, and every network is working with separate input layers since each one of the networks works with its own input shape and subset of data, configured according to what was previously specified in previous Sections 3.3.1–3.3.3. The optimal hyperparameters for the additional layers required by the ensembled model are shown in Table 9. The final output of the model is the predicted road accident probability in the (x,y) selected zone of the spatio-temporal matrix.

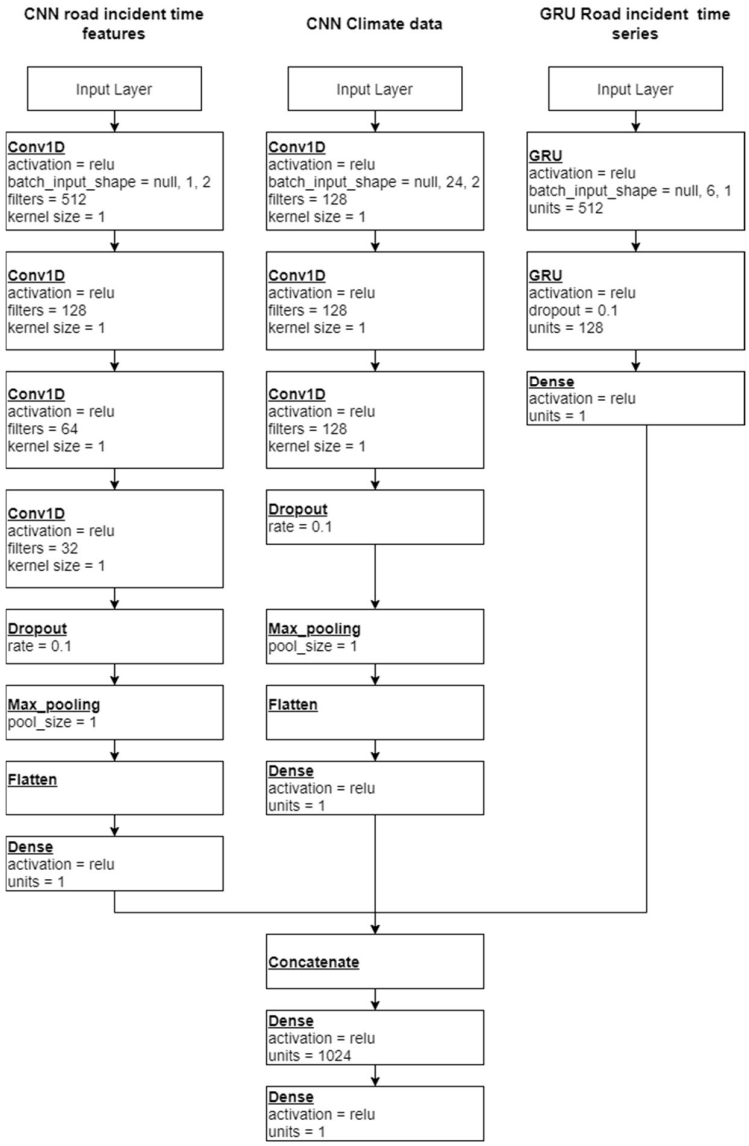


Figure 5. Ensemble network architecture.

Table 9. Ensemble model hyperparameters.

Hyperparameter	Values Evaluated	Optimal Value
Number of neurons	128, 256, 512, 1024	1024
Number of layers	1, 2, 3, 4	1
Epoch number	10, 50, 100, 200, 500	100
Batch size	10, 20, 200, 500, 1000	10
Optimizer	Adam, RMSProp	Adam

4. Results

The data sets are divided into a training set (60%) and a test set (40%). The results obtained were evaluated using Mean Squared Error (MSE), Root Mean Squared Error RMSE and Mean Absolute Error (MAE) metrics [41]. MSE was calculated as the average of the squared differences between each computed value and its corresponding correct value and RMSE was calculated as the square root of the MSE and was used as a measure of differences between the value predicted and the real values. A low mean absolute error (MAE) indicates good predictive accuracy. A lower value of RMSE is indicative of a higher prediction precision. Both MSE and RMSE are widely used to evaluate deep learning

models applied in the field for road accident forecasting and prediction. Mean absolute error is also known as average prediction error and is calculated as the average of the difference between the predicted and actual value in all the test cases; a low MAE value indicates good predictive accuracy. The objective of using MSE, RMSE and MAE indexes to evaluate the model was to facilitate comparison with the results presented by other authors in similar or related investigations, as reported by [23–25,32,42,43].

The models were implemented in Python running on a system using an NVIDIA GeForce RTX 2060 GPU and an Intel Core i7-8750H processor.

In order to validate that the model can generate predictions of traffic accident probability with data from every region of the city we employed randomly generated test data. The results of the testing process using 20 randomly selected training data are shown in Table 10.

Table 10. Ensemble model performance using randomly selected test data.

Description	MSE	RMSE	MAE
Mean value	0.0049	0.225	0.160
Standard deviation	0.001	0.009	0.002
Minimum value	0.0047	0.217	0.156
Maximum value	0.0053	0.255	0.164

The next step in validating the model was selected to evaluate its performance, therefore, 20 regions were selected from the spatio-temporal matrix, as shown in Table 11. The selection criteria for the studied regions were based on quartiles, according to the number the values of the number of accidents reported, and the most representative regions of each quartile were taken. The geographical location of the regions is shown in Figure 6. As a relevant fact, the top five regions can be defined as areas with high traffic, all include intersections between main avenues and Bus Rapid Transit (TransMilenio) corridors. It is worth noting that according to [44] the BRT has an impact on the increase of road accidents, particularly those involving pedestrians around the busiest BRT stations at peak times.

The ensemble model reported reliable and stable results forecasting the probability of road accidents in the selected zone of Bogota city. It must be considered that the same level of results must not be attainable in other regions of the city, since they have an exceptionally low level of road accidents reported, and therefore, making it unfeasible to model their behavior using a time-series or other model that can lead to uncovering an underlying pattern. The results of the model are shown in Table 11. The first entry in the table corresponds to results obtained by the model using the union of all accidents, called the "Generalized Model". The Generalized model was employed as a basis to predict the probability of the other regions selected to be studied.

4.1. Model Comparison and Validation

In the next subsections, the results obtained are compared with baseline models and with results reported by other researchers that had employed deep learning models for the prediction of road accidents.

4.1.1. Results Comparison with Baseline Models

Regarding the algorithms employed to benchmark the results obtained by the ensemble model, the algorithms Ada Boost, Linear Regression, Random Forest Regression, and Support Vector Regression (SVR) were considered. The parameters and configuration of each algorithm are shown in Table 12.

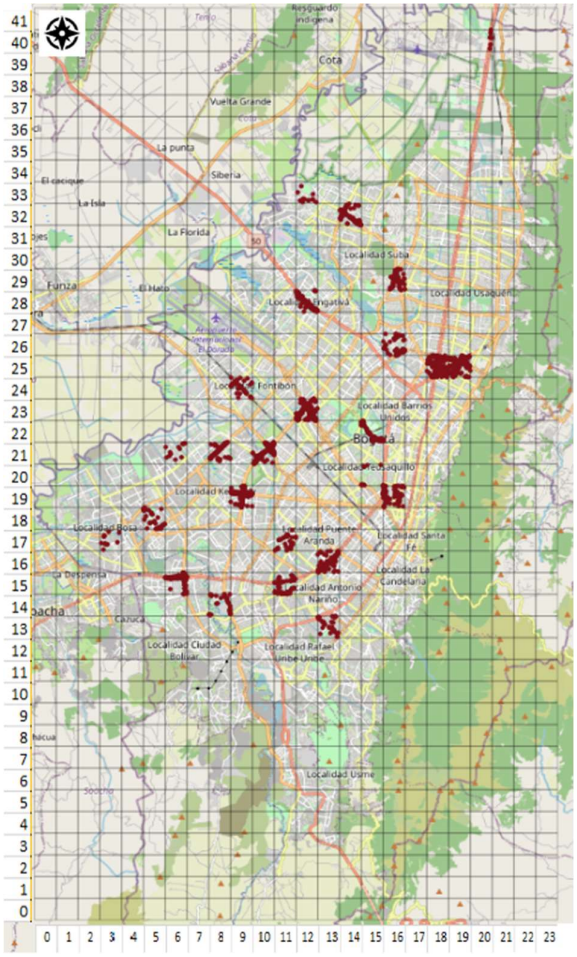


Figure 6. Geographical location of the selected regions used to evaluate the ensemble model.

Table 11. Results obtained by the ensemble model. The first entry in the table corresponds to the results obtained by the model using the union of all accidents, called the "Generalized Model".

Region	Number of Accidents	MSE	RMSE	MAE
Generalized model	3071	0.049	0.22	0.159
(X = 6 Y = 15)	321	0.07	0.264	0.183
(X = 9 Y = 19)	286	0.045	0.213	0.158
(X = 10 Y = 21)	277	0.05	0.223	0.163
(X = 18 Y = 25)	269	0.046	0.214	0.154
(X = 12 Y = 23)	248	0.038	0.196	0.153
(X = 19 Y = 25)	238	0.073	0.271	0.181
(X = 14 Y = 32)	155	0.057	0.239	0.167
(X = 12 Y = 28)	153	0.062	0.25	0.176
(X = 15 Y = 22)	152	0.048	0.218	0.153
(X = 16 Y = 29)	152	0.056	0.237	0.168
(X = 13 Y = 16)	150	0.053	0.23	0.171
(X = 16 Y = 19)	96	0.07	0.265	0.178
(X = 8 Y = 21)	96	0.038	0.194	0.149
(X = 11 Y = 15)	96	0.05	0.223	0.167
(X = 16 Y = 26)	96	0.053	0.229	0.154
(X = 11 Y = 17)	45	0.022	0.148	0.106
(X = 13 Y = 13)	44	0.032	0.179	0.144
(X = 5 Y = 18)	44	0.01	0.098	0.087
(X = 9 Y = 24)	44	0.014	0.12	0.103
(X = 8 Y = 14)	44	0.03	0.172	0.156

Table 12. Baseline algorithms parameters and configuration.

Algorithm	Parameters
AdaBoost	Base estimator: tree Number of estimators: 100 Algorithm (classification): Samme.r Loss (regression): Square
Linear Regression	Regularization: Lasso Regression (L1) with α = 0.0001
Random Forest Regression	Number of trees: 100 Maximal number of considered features: 2. Replicable training: Yes Maximal tree depth: unlimited Stop splitting nodes with maximum instances: 5.
Support Vector Regression	SVR type: SVR, C = 1.0, ε = 0.1 Kernel: Linear Numerical tolerance: 0.001 Iteration limit: 100

The results obtained by the model were a benchmark against the model detailed in Tables 12 and 13. It must be considered that the benchmark algorithms were evaluated using a consolidated dataset, shaped as (Sample, Features), being the Sample size the 60% of the dataset for the training set and 40% for the testing set; the Features contains X position, Y position, day of the week, the hour of the day, rain in millimeters and road accident probability.

Based on the information in Tables 13 and 14, it can be ensured that the proposed model presents a better performance to predict road accidents in areas where there is a greater number of reports and its performance declines in areas with fewer reports, as can be seen in the results. This behavior is expected since Deep Learning methods require a significant amount of data in order to obtain better results [16].

Table 13. Ensemble model results comparison with baseline algorithms using RMSE.

	Generalized Model	(x = 6, y = 15)	(x = 9, y = 19)	(x = 10, y = 21)	(x = 12, y = 23)	(x = 18, y = 25)
Ensemble model	**0.215**	**0.264**	**0.213**	**0.223**	**0.196**	**0.214**
AdaBoost	0.365	0.337	0.264	0.292	0.273	0.262
Linear Regression	0.229	0.271	0.206	0.230	0.220	0.244
Random Forest Regression	0.228	0.280	0.214	0.244	0.245	0.239
SVR	8.209	0.305	0.259	0.315	0.326	0.250

Table 14. Ensemble model results comparison with baseline algorithms using MSE.

	Generalized Model	(x = 6, y = 15)	(x = 9, y = 19)	(x = 10, y = 21)	(x = 12, y = 23)	(x = 18, y = 25)
Ensemble model	**0.046**	**0.070**	**0.045**	**0.050**	**0.038**	**0.046**
AdaBoost	0.133	0.113	0.070	0.085	0.075	0.069
Linear Regression	0.052	0.074	0.042	0.053	0.048	0.060
Random Forest Regression	0.052	0.078	0.046	0.060	0.060	0.057
SVR	67.400	0.093	0.067	0.099	0.106	0.062

4.1.2. Results Comparison with Deep Learning Methods Reported by Literature

In order to make a comparison with other methods reported in the literature, papers were selected that had used or designed deep learning models for the prediction of road accidents, with a spatio-temporal resolution that was equivalent to that one described in this document. The description of the investigations selected for the comparison of

results is illustrated in Section 2, Related Work. Regarding the obtained results used for comparison, the results obtained by the generalized model, as shown in Table 10 were used, and therefore, the reference numbers calculated were an RMSE value of 0.220 and an MSE value of 0.049 and an MAE value of 0.159.

The comparison with other deep learning models was better or concurrent with those reported in the literature as shown in Table 15. It should be considered that each model was designed using different input data and that each model reflects the situation of road traffic in cities with unique characteristics of development, population, and infrastructure.

Table 15. Model comparison results with Deep Learning methods using RMSE and MSE. Empty cells indicate that the authors did not report the result.

Method	RMSE	MSE
Ensemble model	0.215	0.046
[23] SdAE	1.00	-
[24] STCL-Net Model grid 8x3 hourly prediction	-	0.378
[24] STCL-Net Model grid 15x3 hourly prediction	-	0.119
[24] STCL-Net Model grid 30x10 hourly prediction	-	0.003
[25] DSTGCN	0.343	-

5. Conclusions

This study applied an ensembled deep learning model GRU-CONV1D to predict road accident probability using spatio-temporal information, employing social media and meteorological information as sources. In order to make the raw data suitable for working with Machine Learning techniques, we described a data quality process by which the data was cleaned of outlier values and noise, followed by a feature engineering process, which allowed us to enhance the original information, create the means to model the data as a spatio-temporal matrix, calculate road accident probabilities and model the features as a time series. Next, we design the deep learning ensemble model, that allowed us to integrate the deep learning models more suitable to extract the features available in every data source, such as a GRU network to extract patterns from a road accident time series, a CNN network to extract relations from road accidents and holidays and traffic peak hours, and finally, a CNN network designed to understand the underlying relationship between climatological conditions and road accidents.

The results imply that the proposed ensemble deep learning model performed better than the benchmark models, with the aforementioned advantage of integrating several data sources that allowed us to consolidate various viewpoints of road accident analysis. The comparison with other deep learning models yields interesting results and the results obtained are better or concurrent with those reported in the literature. However, it must be considered that the predictions obtained by the models are highly dependent on the input data and it must be considered the influence of the particularities of the situation of traffic accidents of each city, which were influenced by elements such as development, the conditions of the road infrastructure and the culture of drivers and pedestrians, which at the present time cannot be incorporated into the models very precisely.

The information provided by the model can be valuable for traffic control agencies to plan road accident prevention activities since the results obtained showed problematic regions regarding road accidents, which can be characterized as areas with high traffic, which include private vehicles, buses and trucks, and the regions included intersections between main avenues and Bogota city Bus Rapid Transit (TransMilenio) corridors and stations. A limitation of the present study is that there were zones in the city that have a scarce level of road accidents reported, and therefore, it is not realistic to model their behavior using a time-series model and using a deep learning model to analyze them. In future work, it is planned to integrate other relevant data such as traffic flow, land use, average traffic speed and additional road accident features, such as accident severity, to improve the predictions obtained.

Author Contributions: Conceptualization, C.A.P.; methodology, C.A.P. and F.A.G.; software, C.G-O; validation, C.A.P. and F.A.G.; investigation, C.G-O and C.A.P.; data curation, C.G-O; writing—original draft preparation, C.G-O; writing—review and editing, C.G-O and C.A.P.; supervision, C.A.P. and F.A.G.; project administration, C.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to acknowledge the support by the research group PLaS (Programming Languages and Systems) of the Universidad Nacional de Colombia.

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

References

1. World Health Organization. *Global Status Report on Road Safety 2018*; World Health Organization: Geneva, Switzerland, 2018.
2. 3 Star or Better-iRAP. Available online: https://irap.org/3-star-or-better/ (accessed on 21 March 2022).
3. Choi, J.Y.; Lee, B. Combining LSTM Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting. *Math. Probl. Eng.* **2018**, *2018*, 2470171. [CrossRef]
4. Lv, Y.; Chen, Y.; Zhang, X.; Duan, Y.; Li, N.L. Social media based transportation research: The state of the work and the networking. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 19–26. [CrossRef]
5. Amin-Naseri, M.; Chakraborty, P.; Sharma, A.; Gilbert, S.B.; Hong, M. Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from Waze. *Transp. Res. Rec.* **2018**, *2672*, 34–43. [CrossRef]
6. Moriya, K.; Matsushima, S.; Yamanishi, K. Traffic Risk Mining From Heterogeneous Road Statistics. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3662–3675. [CrossRef]
7. Gutierrez-Osorio, C.; Pedraza, C. Modern data sources and techniques for analysis and forecast of road accidents: A review. *J. Traffic Transp. Eng. Engl. Ed.* **2020**, *7*, 432–446. [CrossRef]
8. Lu, H.; Zhu, Y.; Shi, K.; Lv, Y.; Shi, P.; Niu, Z. Using Adverse Weather Data in Social Media to Assist with City-Level Traffic Situation Awareness and Alerting. *Appl. Sci.* **2018**, *8*, 1193. [CrossRef]
9. D'Andrea, E.; Ducange, P.; Lazzerini, B.; Marcelloni, F. Real-Time Detection of Traffic from Twitter Stream Analysis. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2269–2283. [CrossRef]
10. Afzaal, M.; Nazir, N.; Akbar, K.; Perveen, S. Real Time Traffic Incident Detection by Using Twitter Stream Analysis. In Proceedings of the International Conference on Human Systems Engineering and Design: Future Trends and Applications, Reims, France, 25–27 October 2018; Springer: Cham, Switzerland, 2018.
11. Jones, A.S.; Georgakis, P.; Petalas, Y.; Suresh, R. Real-time traffic event detection using Twitter data. *Infrastruct. Asset Manag.* **2018**, *5*, 77–84. [CrossRef]
12. Suat-Rojas, N.; Gutierrez-Osorio, C.; Pedraza, C. Extraction and Analysis of Social Networks Data to Detect Traffic Accidents. *Information* **2022**, *13*, 26. [CrossRef]
13. Zhang, Z.; He, Q.; Gao, J.; Ni, M. A deep learning approach for detecting traffic accidents from social media data. *Transp. Res. Part C Emerg. Technol.* **2017**, *86*, 580–596. [CrossRef]
14. Pandhare, K.R.; Shah, M.A. Real time road traffic event detection using Twitter and spark. In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10–11 March 2017; pp. 445–449.
15. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *Pattern Anal. Mach. Intell. IEEE Trans.* **2013**, *35*, 1798–1828. [CrossRef]
16. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1–21. [CrossRef]
17. Ren, H.; Song, Y.; Wang, J.; Hu, Y.; Lei, J. A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2017.
18. Zheng, M.; Li, T.; Zhu, R.; Chen, J.; Ma, Z.; Tang, M.; Cui, Z.; Wang, Z. Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access* **2019**, *7*, 39897–39910. [CrossRef]
19. Lorenc, A.; Kuźnar, M.; Lerher, T.; Szkoda, M. Predicting the probability of cargo theft for individual cases in railway transport. *Teh. Vjesn.* **2020**, *27*, 773–780.
20. Kuźnar, M.; Lorenc, A. A method of predicting wear and damage of pantograph sliding strips based on artificial neural networks. *Materials* **2022**, *15*, 98. [CrossRef]

21. Dabiri, S.; Heaslip, K. Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert Syst. Appl.* **2019**, *118*, 425–439. [CrossRef]
22. Li, P.; Abdel-Aty, M.; Yuan, J. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accid. Anal. Prev.* **2020**, *135*, 105371. [CrossRef]
23. Chen, C.; Fan, X.; Zheng, C.; Xiao, L.; Cheng, M.; Wang, C. SDCAE: Stack Denoising Convolutional Autoencoder Model for Accident Risk Prediction Via Traffic Big Data. In Proceedings of the Sixth International Conference on Advanced Cloud and Big Data (CBD), Lanzhou, China, 15–15 August 2018; IEEE: New York, NY, USA, 2018; pp. 328–333.
24. Bao, J.; Liu, P.; Ukkusuri, S.V. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accid. Anal. Prev.* **2019**, *122*, 239–254. [CrossRef]
25. Yu, L.; Du, B.; Hu, X.; Sun, L.; Han, L.; Lv, W. Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing* **2020**, *423*, 135–147. [CrossRef]
26. Zhang, Z.; Robinson, D.; Tepper, J. Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network. In Proceedings of the European Semantic Web Conference, Hersonissos, Greece, 29 May–2 June 2018; Springer: Cham, Switzerland, 2018; pp. 1–10.
27. Wu, L.; Kong, C.; Hao, X.; Chen, W. A Short-Term Load Forecasting Method Based on GRU-CNN Hybrid Neural Network Model. *Math. Probl. Eng.* **2020**, *2020*, 1428104. [CrossRef]
28. Pérez-Espinosa, A.; Reyes-Cabello, A.L.; Quiroz-Fabián, J.; Bravo-Grajales, E. Trafico CDMX system: Using big data to improve the mobility in Mexico City. In Proceedings of the 2018 International Conference on Big Data and Computing, Shenzhen, China, 28–30 April 2018; pp. 13–17.
29. Parnami, A.; Bavi, P.; Papanikolaou, D.; Akella, S.; Lee, M.; Krishnan, S. Deep Learning Based Urban Analytics Platform: Applications to Traffic Flow Modeling and Prediction. In *ACM SIGKDD Workshop on Mining Urban Data (MUD3)*; ACM: London, UK, 2018.
30. Bonilla, J.A. The More Stringent, the Better? Rationing Car Use in Bogotá with Moderate and Drastic Restrictions. *World Bank Econ. Rev.* **2019**, *33*, 516–534. [CrossRef]
31. Roshandel, S.; Zheng, Z.; Washington, S. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accid. Anal. Prev.* **2015**, *79*, 198–211. [CrossRef]
32. Zhou, Z.; Wang, Y.; Xie, X.; Chen, L.; Liu, H. RiskOracle: A minute-level citywide traffic accident forecasting framework. *arXiv* **2020**, preprint. [CrossRef]
33. Geurts, K.; Wets, G.; Brijs, T.; Vanhoof, K. Profiling of High-Frequency Accident Locations by Use of Association Rules. *Transp. Res. Rec.* **2003**, *1840*, 123–130. [CrossRef]
34. Kumar, S.; Toshniwal, D. A data mining approach to characterize road accident locations. *J. Mod. Transp.* **2016**, *24*, 62–72. [CrossRef]
35. Branco, P.; Ribeiro, R.P.; Torgo, L.; Krawczyk, B.; Moniz, N. SMOGN: A Pre-processing Approach for Imbalanced Regression. *Proc. Mach. Learn. Res.* **2017**, *74*, 36–50.
36. You, J.; Wang, J.; Guo, J. Real-time crash prediction on freeways using data mining and emerging techniques. *J. Mod. Transp.* **2017**, *25*, 116–123. [CrossRef]
37. Wen, Q.; Sun, L.; Song, X.; Gao, J.; Wang, X.; Xu, H. Time Series Data Augmentation for Deep Learning: A Survey. *arXiv* **2020**, arXiv:2002.12478.
38. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
39. Gutierrez-Osorio, C.; Pedraza, C.A. Characterizing road accidents in urban areas of Bogota (Colombia): A data science approach. In Proceedings of the 2nd Latin American Conference on Intelligent Transportation Systems (ITS LATAM), Bogota, Colombia, 19–20 March 2019; pp. 1–6.
40. Fan, X.; He, B.; Wang, C.; Li, J.; Cheng, M.; Huang, H.; Liu, X. *Big Data Analytics and Visualization with Spatio-Temporal Correlations for Traffic Accidents*; Springer: Cham, Switzerland, 2015; p. 9529.
41. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
42. Albertengo, G.; Hassan, W. Short Term Urban Traffic Forecasting Using Deep Learning. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 4–5. [CrossRef]
43. Çodur, M.Y.; Tortum, A. An Artificial Neural Network Model for Highway Accident Prediction: A Case Study of Erzurum, Turkey. *PROMET-Traffic Transp.* **2015**, *27*, 217–225. [CrossRef]
44. Bocarejo, J.P.; Velasquez, J.M.; Díaz, C.A.; Tafur, E.L. Impact of bus rapid transit systems on road safety: Lessons from Bogotá, Colombia. *Transp. Res. Rec.* **2012**, *2317*, 1–7. [CrossRef]