Identifying Road Accident Black Spots using Classical and Modern Approaches

IOANNIS KARAMANLIS[1], ALEXANDROS KOKKALIS[1], VASSILIOS PROFILLIDIS[1], GEORGE BOTZORIS[1], ATHANASIOS GALANIS[2]
[1]Departmentof of Civil Engineering, Democritus University of Thrace,
Kimmeria Campus, 67100, Xanthi,
GREECE

[2]Department of Civil Engineering, International Hellenic University,
End of Magnesias Street, 62124, Serres,
GREECE

Abstract: - The utilization of conclusions from the data analysis of road traffic accidents is of high importance for the development of targeted traffic safety measures, which will effectively reduce the rate of road traffic accidents, thus promoting road safety. Considering the problems of time and money, it is not practical to improve road safety in all the places where road traffic accidents occur. Therefore, the process of identifying accident-prone locations, known as black spots, is a cost-effective and efficient way to analyze the causes of road accidents and reduce them. Identifying black spots is an effective strategy to reduce accidents. The core methods that may be used in the process of identifying the black spots of a road network are the sorting, grouping, and accident prediction methods. However, in practice, it is easy to overlook certain factors that significantly contribute to defining and characterizing a spot on the road network as black. Therefore, suggestions to carry out projects required to reduce security risks shall not be based on the above methods. Machine learning algorithms that in recent years have been widely used in the field of predicting a road traffic accident cover these weaknesses. They can effectively classify data sets and make a connection between factors and the severity of events. Machine learning algorithms include classification, regression, clustering, and dimensionality reduction. In this work, a study was conducted on road traffic accidents that took place on the national and provincial network of Northern Greece from 2014 to 2018, with the aim of determining the black spots. The study provided the general public access to a database of black spots on the road network of Northern Greece. At the same time, it created a point of reference for the recognition of the points in question located on the entire road network, and selected a black spot determination model, after having compared specific measures to determine the quality of a model, which resulted from the application of a logistic regression and machine learning algorithms.

Key-Words: -Black spot, road safety, traffic accidents, binary logistic regression, data set, machine learning

Received: August 28, 2022. Revised: April 19, 2023. Accepted: May 5, 2023. Published: May 29, 2023.

1 Introduction

In modern times, the improvement of human living standards and the economic development of a country undoubtedly depend on its transportation system, which constitutes an essential element of human civilization. Its improper management may lead to harmful situations for the citizens. One such situation is traffic accidents, which are included in the list of the most important social problems and form an ever-present threat to all road users. They are complex phenomena that researchers must deal with [1], given that they involve the simultaneous interaction of a multitude of factors. The most important factors are the road user, the vehicle, the road environment, and last but not least the road alignment and the various geometric characteristics, [2], [3], [4], [5]. The above-mentioned factors are inextricably linked, constituting a system with four hypostases, [6], [7]. In this particular system, the driver's behavior occupies a dominant position, [8].

Some authors argue that factors related to road traffic accidents fall within two categories: a) the road environment/infrastructure and the vehicle, b) the human factor identified with the concept of the driver to whom the majority of road accidents must be attributed, [9], [10].

In [11], the author supports that the human factor plays the most decisive and important role, having taken into account earlier studies, [12], [13], related to the interconnection degree of the factors that cause road traffic accidents (Table 1).

Table 1. Participation percentage of various factors in road traffic accidents, [6], [11]

Only human	65%
Human and road	24%
Human and vehicle	4.50%
Human, road, and vehicle	1.25%
Only road	2.50%
Road and vehicle	0.25%
Only vehicle	2.50%
Total	100%

The analysis/investigation of road traffic accidents is widely recognized as an important part of comprehensive and effective road safety management, [14]. Their study is a process that aims at developing prediction models, as well as formulating policies to improve road safety. The determination of their evolution over time is achieved through the development of methodologies that enhance the identification of road network locations in which an increased concentration of accidents is observed and the risk of collision is characterized as high, [15]. These specific locations are called "traffic accident black spots" or "dangerous road spots".

After identifying the black spots, short-term and long-term countermeasures are suggested for their restoration. Identifying them is a cost-efficient and effective way of analyzing the causes of road traffic accidents and reducing them, [16]. In addition, their further analysis and treatment are widely considered an effective approach to the prevention of road traffic accidents, [17], given that it may lead to the identification of problems concerning the design of the road network (such as wrong slope, lack of street lighting, incomplete vertical and horizontal markings), but also driving behaviors (such as distraction, excessive speed) that demonstrably lead to traffic accidents.

The identification of black spots is difficult due to:

a) the lack of data and specifically the limited number of accidents in certain geographical areas,
b) the low quality of data due to their incorrect recording by competent government agencies,
c) the limitations of the applied methods to determine the black spots (e.g. GIS and statistical analyses), when it comes to the accuracy and completeness of the information provided, and
d) the road environment which, being dynamic, is constantly changing.

In general, improving road safety by eliminating or improving road network black spots involves the following three stages:

i. detection of black spots (known as dangerous or accident-prone locations),
ii. diagnosis of the factors that make the identified sites hazardous, and
iii. addressing the identified problems by establishing and implementing effective countermeasures.

2 Definition of a Black Spot and Identification Criteria

The black spots of road traffic accidents, also known as "blackspots", are spots on the road network where the frequency and consequences of road traffic accidents are significant for a certain period of time, [18]. So far, there is no globally accepted definition of a black spot. In the international literature, they are also called "hotspots", "accident-prone locations", "sites with promise", and "hazardous road locations". The most common assumption for identifying a blackspot is the environmental or road geometry issues that result in repeated accidents.

Officially, the criteria used and the methods applied to determine black spots vary from country to country. A summary can be found in Table 2, [19].

The conventional method used to identify black spots on road networks is based on fixed lengths of road segments, where the total length is divided into road segments of 300, 500, and 1,000 meters. The number of accidents that take place on every road section is then calculated and compared to the black spot criteria. However, the method is not precise, because the segment length is stable where accidents in each segment may not be related to each other. In addition, the method tends to overlook hazardous locations in which the section lengths must be large enough to cover all continuous accidents that may be related to each other.

3 Black Spots Identification Derived from Primary Data Sources

The Road Traffic Accident Reports (DOTA) are forms filled out by the Hellenic Police personnel

after an autopsy on the accident scene. Copies of the DOTA are sent to the Hellenic Statistical Authority (ELSTAT), where they are coded and registered in a database. Based on the above, DOTAs are the primary data sources of road traffic accidents.

Table 2. Summary presentation of criteria and methods for the identification of black spots

Country	Methodology	Sliding window (meters)	Threshold (accidents number)	Severity included	Time period (years)
Denmark	Poisson	Variable length	4	No	5
Croatia	Segment ranking	300	12	Yes	3
Switzerland	Weighted method	100	3	Yes	3
Hungary	Accident Indexing	100 (spot)/ 1,000 (segment)	4	No	3
Switzerland	Accident indexing	100 (spot)/ 500 (segment)	Statistical and critical values	Yes, with different critical values	2
Austria	Accident Rate	250	3	Yes	3
Germany	Weighted indexing	Traffic accident maps	4	No	5
Portugal	Weighted method	200	5	Yes	5
Norway	Poisson, statistical testing	100 (spot)/ 1,000 (segment)	4 (spot), 10 (segment)	Accident cost	5
Scotland	Accident frequency	200	3	No	3
Greece	Absolute count	1,000	2	No	N/A

For the purposes of this research, the collection of data on road traffic accidents that resulted in either the injury or the death of some of the persons involved and that took place on the national and provincial road network of Northern Greece (Nomenclature of Territorial Units for Statistics (NUTS) EL51, EL52, and EL53, according to Eurostat, Fig. 1) from 2014 to 2018 was required. The data came from portals available to the public, from government agencies, as well as from ELSTAT following a request for anonymized data. To further anonymize the data and eliminate any possibility of personalizing a road traffic accident, i.e., the possibility of a connection between a traffic accident and the people involved in it, we normalized the quantitative variables and categorized the qualitative ones. The creation of categories in the case of the "age" variable contributed in the same direction, as well as the creation of new variables (day/night) from the existing ones (time).

For each road traffic accident, the following data were used:

i. Accident location (kilometer point).
ii. Data concerning the incident and road environment (month, S/N, week of the year, dead, seriously injured, lightly injured, total injured, number of vehicles involved in the accident, type of road surface, atmospheric conditions, road surface conditions, marking of directions in road centerline, lane marking, with side safety guard left, with side safety guard right, road width (including the berm), narrowness, turn sequence, road gradient, straightness, right turn, left turn, boundary line marking left, boundary line marking right, accident severity, first collision accident type).
iii. Details of the drivers involved (gender of victims, age of victims).
iv. Details of the vehicles involved (type of vehicle, age of vehicle, mechanical inspection of vehicle).

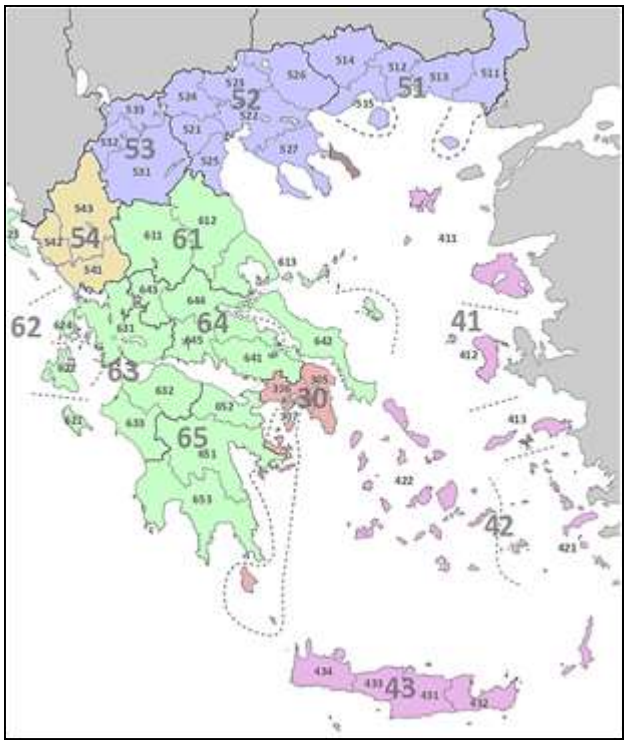

Fig. 1: The study area of Northern Greece (NUTS EL51, EL52, and EL53) (modified by the authors)

As illustrated in Table 3, 1,811 from a total of 13,426 accidents were analyzed, after removing the cases involving road traffic accidents that took place on highways, municipal, community, or other roads, level crossings, road sections with a central median strip, or a safety bar. In these cases, the incomplete

data with the impossibility of assumptions were removed, as well as the unclear data such as street code with 9999 (unknown) and mileage with 999.9 (unknown). The specific removals were made to ensure that the positions examined each time show uniformity in terms of traffic data, and the geometric and functional characteristics of the road that differ along the road axes.

Table 3. Traffic accidents and black spots by Regional Unit of Northern Greece

Regional unit	Traffic accidents (initial records)	Traffic accidents (final records)	Black spots
Rhodopi (EL 513)	390	108	11
Evros (EL 511)	435	120	11
Serres (EL 526)	379	150	12
Pella (EL 524)	275	99	8
Thessaloniki (EL 522)	8,852	369	32
Grevena (EL 531, south)	51	24	1
Kozani (EL 531, north)	220	65	4
Pieria (EL 525)	184	46	1
Kavala (EL 515, mainland)	513	124	4
Drama (EL 514)	359	33	4
Imathia (EL 521)	312	69	4
Chalkidiki (EL 527)	557	272	21
Florina (EL 533)	37	12	1
Xanthi (EL 512)	346	103	13
Kilkis (EL 523)	306	127	9
Thasos (EL 515, inland)	115	54	5
Kastoria (EL 532)	95	36	1
Total	13,426	1,811	142

In our case, the kilometer point where at least two road traffic accidents took place during the period in question was characterized as black spots. Given that in Greece, there are no specific criteria for classifying a spot of the road network as black, the threshold used by ELSTAT was adopted as the threshold for the above classification, which is nothing more than at least two accidents. By using the four classic methods of determining black spots (Poisson random distribution method, Quality control method, Accident frequency method, Severity index method) we confirmed, at a 90% significance level, the black spot characterization, as defined by the ELSTAT (two accidents at the same kilometer point).

The data provided by ELSTAT cannot be processed as such, in terms of statistical analysis methods. For this reason, preliminary work was performed to reflect the specific elements in an appropriate format. Thus, 35 variables were created with a predefined range of values, to reflect, as appropriate, the various ELSTAT data through categorical and quantitative variables. Table 4 presents the variables after said pre-processing.

During the pre-processing of data, several problems related to the DOTAs content were identified. These include:

i. Discrepancies in the accuracy of the location of the accident. The specific discrepancies may be caused by non-updated road mileage, a lack of mileage markers, an undefined mileage starting point, or other factors that do not allow the exact location of the incident to be determined.
ii. Errors in data entry.
iii. Incorrect processing of raw data.
iv. Incomplete data.

Such problems were addressed through the complete removal of records, in consultation with the relevant Authorities for verification, as well as making cases. The latter was performed without excluding data, an action that would have led to fewer data available for processing, thus making their further analysis impossible. The on-site investigation was not chosen due to the objective difficulty of going to all the places where the road traffic accidents in question took place.

Table 4. The variables resulting from the pre-processing of ELSTAT data

Roadway type	Year
Traffic load	Lane divider
Daylight	Driver's age
Month	Road width
Week of year	Road narrowness
Day of week	Lane direction sign
Time	Sequential turns
Serious injured	Road gradient
Deceased	Straightness
Minor injured	Right turn
Totally injured	Left turn
Weekday	Right barrier
Atmospheric conditions	Left edge line
Roadside environment	Right edge line
Road surface conditions	Accident severity
Vehicle type	Vehicle age
Mechanical inspection	Driver's gender
Number of vehicles involved in the traffic accident	

4 Model Creation

4.1 Logistic Regression

The main pattern of the binary logistic regression model has the following form, [20]:

$$f(z)=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}} \quad (1)$$

where the input variable z represents the action of a set of independent variables, and f(z) determines the probability of a certain result occurring due to the above action. The variable z is defined as follows:

$$z=b_0+b_1\cdot x_1+b_2\cdot x_2+...++b_i\cdot x_i \quad (2)$$

where β_0 is the constant term (slope) of the regression line, and β_i are the regression coefficients, expressing the magnitude of contribution to the model of the corresponding variable. A negative value of β_i indicates that the independent variable reduces the probability of an event occurring, while a positive value of the explanatory variable indicates an increase in the above probability. A low value of the regression coefficient indicates a small effect of the independent variable on the probability of an event occurring or not, while a high value indicates a strong effect on the predicted probability.

4.2 K-nearest Neighbors

The most basic method of finding records that belong to the same class based on distance is the K-nearest neighbors' algorithm. In this algorithm, we assume that all instances correspond to spots in the n-dimensional space R^n. The nearest neighbors of an instance are defined by the Euclidean distance. More specifically, let's say that an instance x which is described by the feature vector $((a_1(x), (a_2(x), \ldots, (a_n(x))$ where $a_r(x)$ defines the value of the r-th feature of instance x. The distance between two instances x_i and x_j is equal to [21]:

$$d(x_i,x_j)=\sum_{r=1}^{n}a_r(x_i)-a_r(x_j)^2 \quad (3)$$

For the case of discrete values of form $f:R^n\rightarrow V$, where V is the finite set $\{u_1, u_2\ldots, u_s\}$, the algorithm approximates a discrete-valued "target" f with the estimate $\hat{f}$, which is the most common for instances f and k of the training set closest to x. The algorithm can be divided into two parts:

- training: saving of each training instance (x, f(x)) in the training list,
- sorting: to sort a random x_q instance, we have:

$$\hat{f}(x_q)\leftarrow \arg\max_{u\in V}\sum_{i=1}^{k}\delta(u,f(x_i)) \quad (4)$$

where $x_i, \ldots, x_k$ indicate the shortest k distances at x_q from instances in the training list, and d(a, b)=1 when a=b, otherwise δ=0. If we choose k=1, the algorithm inserts in $\hat{f}(x_q)$ the value $f(x_i)$ where x_i is the closest training instance at x_q. For larger k values, the algorithm inserts the most common value among the closest k instances of the training list.

4.3 The Gaussian Process

The Gaussian Process (GP) is a non-parametric Bayesian method used in machine learning to perform regression and classification. A GP represents a potentially infinite set of random variables ordered in space or time each finite subset of which jointly follows a Gaussian distribution. If we consider a function as an infinite set of spots in space, then we must say that a Gaussian Process is a distribution of functions over space.

A GP is uniquely described by a mean value function m(x) and a covariance (or kernel) function k (x, x′). The choice of kernel determines the smoothness of the function. The best-known kernel functions are RBF and Matérn kernel. We consider that we have a data set (x_i, y_i) consisting of N observations and assume that each observation y_i is derived from a function f(x) with the addition of some Gaussian noise, i.e., $y_i = f(x_i) + E_i$, with $E_i \sim N(0, \sigma^2)$.

A GP can be used as the prior distribution and can be combined with our data to give us the posterior distribution of the function. Thanks to the posterior distribution we can predict the value of the function f at new x^*.

The main disadvantage of the Gaussian Process Regression is its cubic complexity. To solve this problem, many different approximate methods have been generally developed. For the purposes of our analysis, we will focus on such an approximate approach.

4.4 Extra Trees Classifier

Extremely Randomized Trees Classifier (Extra Trees classifier) is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is providing a random sample of k features from the feature set from which each decision tree must select the best feature to split the data based on some mathematical criteria

(typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to his/her choice.

4.5 Multilayered Perceptron (MLP)

It is a neural network where the mapping between inputs and output is non-linear. A Multilayer Perceptron has input and output layers and one or more hidden layers with many neurons stacked together. While in the Perceptron the neuron must have an activation function that imposes a threshold, like ReLU or sigmoid, neurons in a Multilayer Perceptron can use any arbitrary activation function.

Multilayer Perceptron falls under the category of feed-forward algorithms because inputs are combined with the initial weights in a weighted sum and subjected to the activation function, just like in the Perceptron. However, the difference is that each linear combination is propagated to the next layer. Each layer is feeding the next one with the result of their computation, their internal representation of the data. This goes all the way through the hidden layers to the output layer.

5 Prediction of Black Spots Through Pattern Recognition and Deep Learning Architectures

According to the literature, the phenomenon of black spots is complex and cannot be clearly defined. Usually, a black spot results from the accumulation of accidents in a “spot” over time. Identifying black spots is an effective strategy to reduce accidents. The core methods that can be used while identifying the black spots of a road network are the sorting, grouping, and accident prediction methods, [22]. However, in practice, it is easy to overlook certain factors that significantly contribute to the definition and characterization of a road network spot as black. Therefore, proposals to carry out projects required to reduce security risks shall not be based on the aforementioned methods. In addition, current research on road safety shows that applied statistical modeling fails when dealing with complex and highly non-linear data, [23], which could suggest that the relationship between influencing factors and outcomes of road traffic accidents is more complex and cannot be simply identified with the help of a single statistical approach. Most statistical methods are based on several strong assumptions, such as a priori determination and error distribution. In addition, a problematic issue is multi-co-linearity, namely a high degree of correlation between two or more independent variables. Furthermore, statistical models struggle when dealing with outliers, and missing or noisy data, [24].

The aforementioned weaknesses and limitations are covered and addressed by machine learning algorithms. These include Artificial Neural Networks (ANN), as well as deep learning models that have been applied to various traffic safety problems and have been used as data analysis methods due to their ability to work with huge amounts of multi-dimensional data. Due to its modeling flexibility, learning and generalization ability, as well as good predictive ability, machine learning has been considered a set of convenient and accurate mathematical models in the field of road safety.

In this study, it is attempted to analyze the spatiotemporal phenomenon called a road traffic accident, to predict black spots without the accumulation of accidents or over time using pattern recognition and deep learning methods learning.

Initially, four well-known machine learning methods were tested on the vectors resulting from the DOTA pre-processing, namely Nearest Neighbors, Gaussian Process, Extremely Randomized Trees, and Multilayer Perceptron Neural Network.

The results widely varied, thanks to the ability of each algorithm to find correlations and patterns in the data. Among these algorithms, some tend to perform better on linear data while others on data with non-linear correlations. Nevertheless, the main problem identified is that, in general, most methods cannot perform operations on categorical variables. The input vectors resulting from the DOTA pre-processing in many cases contain categorical variables with value fields, which are numeric but have no placement in the Euclidean space. For example, the variable 'road surface conditions' has a range of values [1 (Normal), 2 (Wet-wet), 3 (Slick, oils), 4 (Icy), 5 (Snowy), 6 (Other)]. This is problematic for methods that use vector distances to function, since there is no distance between the

"Wet-Wet" and "Other" values, while there are between 2 and 6. Thus, the distance between vectors is false and therefore algorithms that implemented such operations are not reliable.

This has led to the development of a method that represents vectors in a non-linear latent space, but has the meaning of properties in the projected space, even if it is not linear. This implies that transformation can be used by the above algorithms without questioning the correctness of their application and the results. Figure 2 illustrates the proposed method of this study in discrete steps.

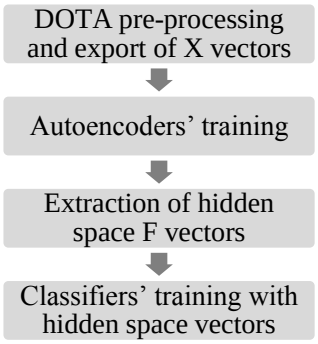

Fig. 2: Proposed methodology steps

The framework of self-supervised learning of autoencoders was used in order to represent vectors that have arisen during pre-processing of data from the primary DOTA source in a space that is appropriate for training pattern recognition and machine learning classifiers.

Autoencoders are closely related to the Principal Component Analysis (PCA) method. If the activation function used in the autoencoder is linear at each level, the hidden variables that are present at the bottleneck (the smallest level in the network) directly correspond to the principal components from the PCA. Generally, the activation function used in autoencoders is non-linear typical activation functions are ReLU (Corrected Linear Unit) and sigmoid.

The mathematics behind these networks is fairly self-explanatory and is presented briefly below. The main idea is that the network is divided into two parts: the encoder and the decoder. The encoder function, denoted by ϕ, maps the original data X into a hidden space F that exists at the bottleneck. The decoder function, denoted by ψ, maps the hidden space F to the output bottleneck. The output, in this case, is the same as the input function. Thus, the original input is regenerated after some generalized non-linear compression, [25]:

$$\begin{aligned} &\varphi : X \to F \\ &\psi : F \to X \\ &\varphi, \psi = \arg\min_{\varphi,\psi} \left\| X - (\psi \circ \varphi) X \right\|^2 \end{aligned} \tag{5}$$

The encoding network can be represented by the standard neural network function passed through an activation function, where z is the hidden dimension:

$$z = \sigma(Wx + b) \tag{6}$$

Similarly, the decoding network can be represented in the same way, but with different weights, biases, and possible activation functions used:

$$x' = \sigma'(W'z + b') \tag{7}$$

The error function can then be written in terms of these network functions. This is the loss function we will use in order to train the neural network through the standard backpropagation process, [25]:

$$L(x, x') = \left\| x - x' \right\|^2 = \left\| x - \sigma'(W'(\sigma(Wx + b)) + b' \right\|^2 \tag{8}$$

Since the input and output are the same vectors, it is not really supervised or unsupervised learning, therefore we usually call it self-supervised learning. The goal of the autoencoder is to choose our encoder and decoder functions in such a way that we need the minimum information to encode the input vector so that it is reproduced on the other side.

6 Results

Having a two-valued categorical variable as a dependent variable, to analyze its relationship with a set of independent variables, a binary logistic regression was qualified. This was performed using the statistical program IBM SPSS 25. Since logistic regression tends to violate the principle of co-linearity, a check was conducted at first by applying linear regression analysis. From the latter, co-linearity emerged between several variables. For this reason, an analysis of the core components was conducted, proving that the above variables are not capable of creating reliable factors. The final analysis resulted in the exclusion from the five-variable model. Then, because of the need to create a parsimonious and optimal model that will include the statistically significant variables, a binary

logistic regression was performed using the Backward LR (likelihood ratio) method.

The machine learning models and the corresponding experiments were carried out on a common platform, using Keras and Python in the context of deep learning.

To compare the performance of binary logistic regression with that of machine learning algorithms, three metrics were computed for the evaluation of models such as Accuracy, Recall, and F1 Score, [26].

Accuracy is an efficiency metric that calculates the fraction of total correct forecasted values divided by the total number of test examples (dataset):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP}) + (\text{TN} + \text{FN})} \tag{9}$$

where:

TP: the number of values correctly predicted as positive (true positive),

TN: the number of values correctly predicted as negative (true negative)

FP: the number of values wrongly predicted as positive (false positive),

FN: the number of values wrongly predicted as negative (false negative),

Recall or *Sensitivity* measures how many of the positive values of the dataset were correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

Precision is a measure of how many of the positively predicted values are actually positive and is determined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

Finally, *F1 Score* is the harmonic mean of precision and recall and calculates how much a model perfectly classifies every value of the dataset in the right category:

$$\text{F1 Score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{12}$$

Table 5 summarizes the results of binary regression, as well as those obtained after training well-known machine learning algorithms using the vectors extracted by the autoencoder. The encoding of the input vectors resulted in: 1) the improvement of the accuracy of all algorithms except MLP, which drops from 79.61% to 28.65%, 2) the worsening of the Recall measure for all algorithms with the exception of MLP which rises from 10.67% to 91.93%, and 3) the improvement of the F1 Score measure for all algorithms. Especially, we remark that the logistic regression presents a better accuracy compared to the other machine learning algorithms, as opposed to the other indicators. The very low values of Recall and F1 Score for logistic regression show that the results of this particular model are not qualitative. On the other hand, although machine learning algorithms show slightly lower accuracy values compared to logistic regression, they excel by far in the rest of the model quality evaluation measures. In addition, it seems that decision-making algorithms that operate in non-linear spaces, such as Perceptron-type neural networks and decision-making trees, perform better than those that search for linear correlations in the data. That is to be expected because at the same time, the factors vary and it is complicated to ascertain the degree of participation of the variables that lead to the definition of an accident spot as a black spot. Nevertheless, with 81.26% success (Xtra Trees) it is possible to predict if a spot under certain conditions can be characterized as a black spot, without the accumulation of accidents over time.

Table 5. Comparative performance between binary logistic regression and machine learning algorithms

	Methodology	Accuracy	Recall	F1 Score
Original dataset	Nearest neighbors	68.31%	27.74%	16.05%
	Gaussian process	69.14%	17.74%	16.41%
	Xtra Trees	77.96%	29.03%	31.03%
	Multilayered Perceptron (MLP)	79.61%	10.67%	13.96%
	Binary logistic regression	82.90%	2.20%	4.25%
One-hot encoded	Nearest neighbors	71.62%	16.12%	16.26%
	Gaussian process	79.33%	17.75%	19.54%
	Xtra Trees	81.26%	25.81%	32.32%
	Multilayered Perceptron (MLP)	28.65%	91.93%	30.56%
	Binary logistic regression	82.90%	2.20%	4.25%

7 Conclusions

In conclusion, the identification of the black spots in the road network is decisive for the reduction of traffic accidents and, thus, the promotion of road safety. The thorough study of the spatiotemporal

phenomenon called road traffic accidents requires the collection of a large number of reliable data, as well as advanced analysis methods. This paper collected data related to road traffic accidents that took place on the provincial and national road network of Northern Greece from 2014 to 2018. This data, after being organized, cleaned, and anonymized, formed a database of factors that can lead to road traffic accidents and involve the characteristics of the driver, the vehicle, and the road environment. To solve the binary class problem, the study ended up proposing a four-step approach:

i. encoding of each variable,
ii. training an autoencoder,
iii. extracting hidden space vectors, and
iv. training and using a classifier.

The process of identifying the black spots of a road network is a difficult, complex, and demanding task, in terms of improving methodologies and continuous research. Nonetheless, the study overcomes difficulties related to the identification of black spots, such as the impossibility of evaluating the individual factors that contribute to causing road traffic accidents and the limited and at the same time poor quality data on road traffic accidents.

In conclusion, the study offers a database accessible to the public for further research, as well as a reliable method for the identification of black spots, but with room for improvement through the application of data enrichment techniques with virtual samples.

Acknowledgment:
We would like to thank the Hellenic Statistical Authority for providing the anonymized data related to road accidents that took place from 2014 to 2018 on the road network of Northern Greece.

References:

[1] Rolison, J.J., Regev, S., Moutari, S., Feeney, A., What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records, *Accident Analysis and Prevention*, Vol.115, 2018, pp. 11-24.

[2] Čičković, M., Influence of Human Behaviour on Geometric Road Design, Transportation Research Procedia, Vol. 14, 2016, pp. 4364-4373.

[3] Komackova, L., Poliak, M., Factors Affecting the Road Safety, *Journal of Communication and Computer*, Vol.13, 2016, pp. 146-152.

[4] Wagner, P., Hoffmann, R., Leich, A. Observations on the relationship between crash frequency and traffic flow, *Safety*, Vol.7, No.1, 2021, p. 18.

[5] Qin, X., Ivan, J.N., Ravishanker, N., Selecting exposure measures in crash rate prediction for two-lane highway segments, *Accident Analysis and Prevention*, Vol.36, No.2, 2004, pp. 183-191.

[6] Sabey, B.E., Taylor, H., *The Known Risks we Run: the Highway*. Transport and Road Research Laboratory, Department of the Environment and Transport, Supplementary Report 567, Crowthorne, Berkshire, United Kingdom, 1980. Available from: https://trl.co.uk/uploads/trl/documents/SR567.pdf

[7] Wicaksono, A., Ambarwati, L., Indriastuti, A.K., A comparison of accident characteristics in highland and lowland regions: a case study in Malang District, Indonesia, *Journal of the Eastern Asia Society for Transportation Studies*, Vol.8, 2010, pp. 2160-2172.

[8] Evans, L., *Traffic Safety*, Science Serving Society, Inc., Bloomfield Hills, MI, 2004.

[9] Chliaoutakis, J.E., Demakakos, P., Tzamalouka, G., Bakou, V., Koumaki, M., Darviri, C., Aggressive behavior while driving as predictor of self-reported car crashes, *Journal of Safety Research*, Vol.33, No.4, 2002, pp. 431-443.

[10] Gregersen, N.P., Berg, H.Y., Lifestyle and accidents among young drivers, *Accident Analysis and Prevention*, Vol.26, No.3, 1994, pp. 297-303.

[11] Yannis, G., *Road Accidents, Traffic Behaviour and Safety*, Road Safety Workshop on Road Accidents and Safe Traffic Behaviour: Are they prevented or are they just cured? The American College of Greece – Institute of Public Health, Athens, Greece, 2018. Available from: https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-cp329.pdf

[12] Treat, J.R., Tumbas, N.S., McDonald, S.T., Shinar, D., Hume, R.D., Mayer, R.E., Stansifer, R.L., Castellan, N.J., *Tri- level study of the causes of traffic accidents.* Institute for Research in Public Safety, Indiana University, Bloomington, USA, 1979. Available from:

https://deepblue.lib.umich.edu/handle/2027.42/64993

[13] Rumar, K., The basic driver error: late detection. *Ergonomics*, Vol. 33, 1990, pp. 1281-1290.

[14] Wang, J., Yan, M., Application of an Improved Model for Accident Analysis: A Case Study. *International Journal of Environmental Research and Public Health*, Vol.16, No.15, 2019.

[15] Elvik, R., A survey of operational definitions of hazardous road locations in some European countries, *Accident Analysis and Prevention*, Vol.40, No.6, 2008, pp. 1830-1835.

[16] Zhang, C., Shu, Y., Yan, L., A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China, *IEEE Access*, Vol.7, 2019, pp. 140197-140205.

[17] Ghadi, M., Török, Á., A comparative analysis of black spot identification methods and road accidents segmentation methods, *Accident Analysis and Prevention*, Vol.128, 2019, pp. 1-7.

[18] Elvik, R., *State-of-the-art approaches to road accident black spot management and safety analysis of road networks.* Norwegian Centre for Transport Studies, Institute of Transport Economics, 2007.

[19] Kokkalis, A., Kalpakis, F., An applied assessment of the procedures and criteria for black spot determination, *International Journal of Transportation*, Vol.5, No.2, 2017, pp. 15-32.

[20] Profillidis, V.A., Botzoris, G.N., *Modeling of Transport Demand*, Elsevier, 2018.

[21] Mitchell, T.M., *Machine Learning*, McGraw-Hill, 1997.

[22] Ghadi, M., Török, A., Comparison Different Black Spot Identification Methods, *Transportation Research Procedia*, Vol.27, 2017, pp. 1105-1112.

[23] Karlaftis, M.G., Vlahogianni, E.I., Statistical methods versus neural networks in transportation research: Differences, similarities and some insights, *Transportation Research Part C: Emerging Technologies*, Vol.19, No.3, 2011, pp. 387-399.

[24] Principe, J.C., Euliano, N.R., Lefebvre, W.C., *Neural and Adaptive Systems: Fundamentals through Simulations*, Wiley, 1999.

[25] Launay, H., Ryckelynck, D., Lacourt, L., Besson, J., Mondon, A., Willot, F, Deep Multimodal autoencoder for crack criticality assessment, *International Journal of Numerical Engineering*, Vol.123, No.6, 2022, pp. 1456-1480.

[26] Kolidakis, S.Z., Kotoula, K.M.A., Botzoris, G.N. (2022). School mode choice classification model exploitation though artificial intelligence classification application. *Mathematical Modelling of Engineering Problems*, Vol.9, No.6, 2022, pp. 1441-1450.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.