MOBILITY & VEHICLE MECHANICS

https://doi.org/10.24874/mvm.2025.51.03.05
UDC: 656.1.08:004.414.23

ROAD TRAFFIC ACCIDENTS PREDICTION USING MACHINE LEARNING METHODS

Vesna Ranković[1], Andrija Đonić[2], Tijana Geroski[3]*

Received in August 2024 *Accepted in October 2024*

RESEARCH ARTICLE

ABSTRACT: Road traffic accidents are identified as a significant societal issue based on extensive and comprehensive research in public health and traffic safety. Such incidents lead to significant negative outcomes, including human casualties, economic consequences, vehicle damage, and significant medical costs. Developing predictive models is crucial for identifying risk factors associated with accidents, thereby improving understanding of accident causes and enabling more effective prevention interventions. The occurrence of traffic accidents is influenced by a multitude of factors, including driver behavior, vehicle characteristics, weather conditions, road volume, road geometry, type of road, road conditions, speed limits, frequency of police controls, etc. In this paper, machine learning (ML) techniques are used to develop the traffic accident prediction model due to the non-linear relationship between input and output variables. The research investigates the influence of certain input variables on the number of traffic accidents, as their optimal choice significantly affects the prediction performance. The random forest, support vector machine, and neural networks are employed for data preprocessing and model development. Statistical indicators are used to evaluate the performance of the developed models. Based on the obtained performances, the developed ML models accurately predict the number of traffic accidents.

KEY WORDS: *road safety, road accident, prediction, machine learning, feature*

[1]*Vesna Ranković, Faculty of Engineering, University of Kragujevac, Sestre Janjić 6, Kragujevac, Serbia, vesnar@kg.ac.rs, https://orcid.org/0009-0004-0172-3216, (*Corresponding author)*
[2] *Andrija Đonić, Department of Informatics and Quantitative Methods, Faculty of Economics, University of Kragujevac, Liceja Knezevine Srbije 3, 34000 Kragujevac, Serbia, andrija.djonic@ef.kg.ac.rs, https://orcid.org/0000-0003-1624-3536*
[3] *Tijana Geroski, Faculty of Engineering, University of Kragujevac, Sestre Janjić 6, Kragujevac, Serbia, tijanas@kg.ac.rs https://orcid.org/0000-0003-3459-3011*

PREDVIĐANJE SAOBRAĆAJNIH NEZGODA KORIŠĆENJEM METODA MAŠINSKOG UČENJA

REZIME: Saobraćajne nezgode su identifikovane kao značajan društveni problem na osnovu opsežnih i sveobuhvatnih istraživanja u oblasti javnog zdravlja i bezbednosti saobraćaja. Takvi incidenti dovode do značajnih negativnih ishoda, uključujući ljudske žrtve, ekonomske posledice, oštećenja vozila i značajne medicinske troškove. Razvoj prediktivnih modela je ključan za identifikaciju faktora rizika povezanih sa nezgodama, čime se poboljšava razumevanje uzroka nezgoda i omogućavaju efikasnije preventivne intervencije. Na pojavu saobraćajnih nezgoda utiče mnoštvo faktora, uključujući ponašanje vozača, karakteristike vozila, vremenske uslove, obim puta, geometriju puta, vrstu puta, uslove na putu, ograničenja brzine, učestalost policijskih kontrola itd. U ovom radu, tehnike mašinskog učenja (MU) se koriste za razvoj modela predviđanja saobraćajnih nezgoda zbog nelinearnog odnosa između ulaznih i izlaznih varijabli. Istraživanje istražuje uticaj određenih ulaznih varijabli na broj saobraćajnih nezgoda, jer njihov optimalan izbor značajno utiče na performanse predviđanja. Za prethodnu obradu podataka i razvoj modela koriste se slučajna šuma, mašina vektora podrške i neuronske mreže. Statistički indikatori se koriste za procenu performansi razvijenih modela. Na osnovu dobijenih performansi, razvijeni MU modeli tačno predviđaju broj saobraćajnih nezgoda.

KLJUČNE REČI: *bezbednost u saobraćaju, saobraćajna nezgoda, predviđanje, mašinsko učenje, funkcija*

ROAD TRAFFIC ACCIDENTS PREDICTION USING MACHINE LEARNING METHODS

Vesna Ranković, Andrija Đonić , Tijana Geroski

INTRODUCTION

Road traffic accidents constitute a serious global issue with significant impacts on human lives and national economies. According to the World Health Organization's Global Road Safety Report 2023 [20], the number of road traffic deaths in 2021 of 1.19 million worldwide represents a 5 % decrease compared to the number of deaths recorded in 2010. Besides the loss of human lives, traffic accidents also impose significant economic consequences. In 2019, the total socio-economic costs of traffic accidents in the Republic of Serbia amounted to 8.8 % of the gross national product [12]. Countries around the world have revised their road safety policies and strategies while incorporating new technologies to mitigate road accidents and their consequences. These efforts are aimed at enhancing traffic safety and reducing the severity of accidents [6].

Data-driven road safety models are crucial for evaluating the effectiveness of applied safety measures and policies, enabling their continuous improvement in order to reduce the number of traffic accidents. In the literature, various techniques for predicting, classifying, and analyzing traffic accidents have been proposed [9]. The factors influencing the number of traffic accidents are numerous and complex. They include technical aspects of roads and vehicles, vehicle speed, traffic density, as well as human factors such as driver behavior, level of concentration and fatigue, and weather conditions [2].

Models based on classical statistical methods, such as the Poisson regression model and negative binomial regression model, have limitations in predicting the number of traffic accidents due to the complex non-linear relationships between input and output variables [11].

To overcome the limitations of traditional statistical methods, various machine learning and deep learning techniques have been applied to traffic safety analysis [5]. These techniques are suitable for modeling flexibility, ability to learn and generalize from data, and high predictive accuracy. Consequently, machine learning models are considered robust and accurate tools in traffic safety research. Artificial intelligence techniques enable the discovery of patterns and relationships in traffic data that were previously difficult to detect using conventional methods. Ali et al.[3] conducted a comprehensive review of machine learning-based models designed to predict different aspects of traffic accidents. The models encompass predicting crash occurrences, forecasting crash frequencies, and estimating the severity of injuries resulting from crashes. Analysis indicates that a significant amount of research has focused on machine learning-based models for predicting injury severity.Almamlook et al. [4] developed a model to binary classify the severity of traffic accidents using various machine learning methods, including AdaBoost, logistic regression, naive Bayes, and random forests.

Gorzelanczyk [10] employed neural network time series prediction using a multilayer perceptron to forecast the number of road accidents. Raja et al. [16] utilized different recurrent and feedforward neural network architectures for classifying severity levels in accidents (slight, serious, and fatal) and used the long short-term memory (LSTM) model for time series forecasting of the number of accidents. The dataset includes various attributes such as personal details, specifics of the accidents, environmental factors, vehicle

information, and road characteristics. Singh et al. [19] used a multilayer perceptron with four hidden layers for predicting road accident frequencies on highways. The sixteen input variables belonging to road geometry, traffic, and road environment were considered as potential inputs to the model. García de Soto [7] developed two neural networks with one hidden layer each to predict the number of accidents with light injuries, severe injuries and fatalities, on open roads and in tunnels. The models utilized continuous input variables such as annual average daily traffic, percentage of heavy traffic, average curve radius, mean positive slope, mean negative slope, longitudinal evenness rating, and surface adhesion rating, as well as categorical variables like posted speed limit and the number of lanes per direction.

This paper utilizes machine learning models to accurately predict the number of road accidents. The performance of feedforward neural networks, support vector machines, and random forest is compared, due to the fact that these models have been shown as the most appropriate in similar problems. The proposed models are applied to selected sections of IB order roads in the Republic of Serbia. Ten features are considered, and the feature importance values are calculated using the RF classifier.

1 METHODOLOGY

Feedforward neural network, support vector machine and random forest are supervised learning algorithms that require a labeled training dataset. The training dataset consists of *m* samples $\left\{\left(\boldsymbol{x}^{(k)}, y_k\right)\right\}_{k=1}^{m}$, where $\boldsymbol{x}^{(k)} \in \mathbb{R}^n$ represents the input variables of the *k*th element of the training data set and $y_k \in \mathbb{R}$ denotes the corresponding target output.

1.1 Feedforward neural network

In this paper, a multi-layer perceptron with one hidden layer was used. The output of the neural network with n inputs, one output, and one hidden layer with *Z* neurons is computed as follows:

$$y_{FNN} = \sum_{j=1}^{Z} \omega_{1,j(2)} f\left(\sum_{i=1}^{n} \omega_{j,i(1)} x_i + b_{j(1)}\right) + b_{1(2)} \tag{1}$$

where: y is the output of the neural network, $\omega_{1,j(2)}$ denotes is the weight from the jth hidden neuron to the output, f is the activation function in the hidden layer, $\omega_{j,i(1)}$ denotes he weight from the ith input to the jth hidden neuron, x_i is the ith input, $b_{j(1)}$ is the bias for the jth hidden neuron, $b_{1(2)}$ is the bias for the output neuron.

The performance of a neural network with a single hidden layer depends on the hyperparameters, which should be carefully chosen. Hyperparameters are the number of hidden neurons that determines the complexity and capacity of the layer, the type of the activation function of these neurons, the algorithm used for training as well as the parameters of the selected algorithm. Optimization strategies for the FNN involve the use of both conventional and metaheuristic approaches [13].

Support vector regression

Support Vector Regression (SVR) is a type of SVM used for regression tasks. Unlike traditional regression methods, SVR aims to find a function that approximates the data within a specified margin of tolerance (ε), while maximizing the margin between parallel hyperplanes [18].

$$y_{SVR} = \sum_{i=1}^{nSVR} \left(\alpha_i - \alpha_i^*\right) K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) + b \tag{2}$$

where: $nSVR$ is the number of support vectors, α_i and α_i^* denote the Lagrange multipliers, $K\left(\boldsymbol{x}_i, \boldsymbol{x}\right)$ is the kernel function that measures the similarity between the support vector $\boldsymbol{x}_i$ and the new sample $\boldsymbol{x}$, b is the bias.

In Support Vector Regression (SVR), various types of kernel functions are employed to establish the relationship between input features and target variables. These include:

Linear kernel:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) = \boldsymbol{x}_i^T \boldsymbol{x} \tag{3}$$

Polynomial kernel:,

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) = \left(\boldsymbol{x}_i^T \boldsymbol{x} + r\right)^d \tag{4}$$

Radial Basis Function (RBF) kernel:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) = \exp\left(\frac{-\left\|\boldsymbol{x} - \boldsymbol{x}_i\right\|}{2\sigma^2}\right) \tag{5}$$

where: r , d and σ represent the parameter of the kernel function.

The choice of kernel function (linear, polynomial, RBF) and its parameters significantly influences the model's ability to capture non-linear relationships and its overall predictive performance. Additionally, fine-tuning the regularization parameter C and the tolerance parameter ε is essential for optimizing the trade-off between model accuracy and its generalization capabilities [1].

1.2 Random forest

Random Forest (RF) is a robust machine learning algorithm that utilizes multiple decision trees to predict outcomes. Each tree in the forest is trained on a random subset of the training data and features. This ensemble technique effectively reduces overfitting by averaging predictions across diverse trees, thereby enhancing accuracy and capturing complex relationships between input variables and the target output. In regression tasks, the Random Forest model computes its prediction by averaging the outputs of all individual trees in the forest.

The overall output of the random forest model for input $\boldsymbol{x}$ is computed by averaging the predictions from all trees:

$$y_{RF}\left(\boldsymbol{x}\right) = \frac{1}{N_t} \sum_{i=1}^{N_y} y_{RF_i}\left(\boldsymbol{x}\right) \tag{6}$$

where: N_t denotes the total number of trees in the random forest ensemble, $y_{RF_i}(\boldsymbol{x})$ represents the prediction of the ith tree for the input $\boldsymbol{x}$.

The RF model's hyperparameters include the number of candidate variables randomly selected for each split, the sample size determining how many observations are randomly sampled for each tree, whether the sampling is done with or without replacement, the minimum number of observations required in a terminal node, the minimum number of observations required to split a node, the total number of trees in the ensemble, and the criterion used for splitting nodes [15]. These parameters collectively influence the RF model's performance and its ability to generalize effectively to new data.

2 RESULTS AND DISCUSSION

The paper discusses part of the roads of the IB order in the Republic of Serbia. The set contains data on 113 sections of roads 22, 23, 33, 34 and 39, Figure 1. Data on section length, annual average daily traffic volume and the number of traffic accidents are extracted from the Public enterprise roads of Serbia [14] and Road traffic safety agency (Republic of Serbia) [17]. Data on terrain type, curvature, lane width are taken from [8].

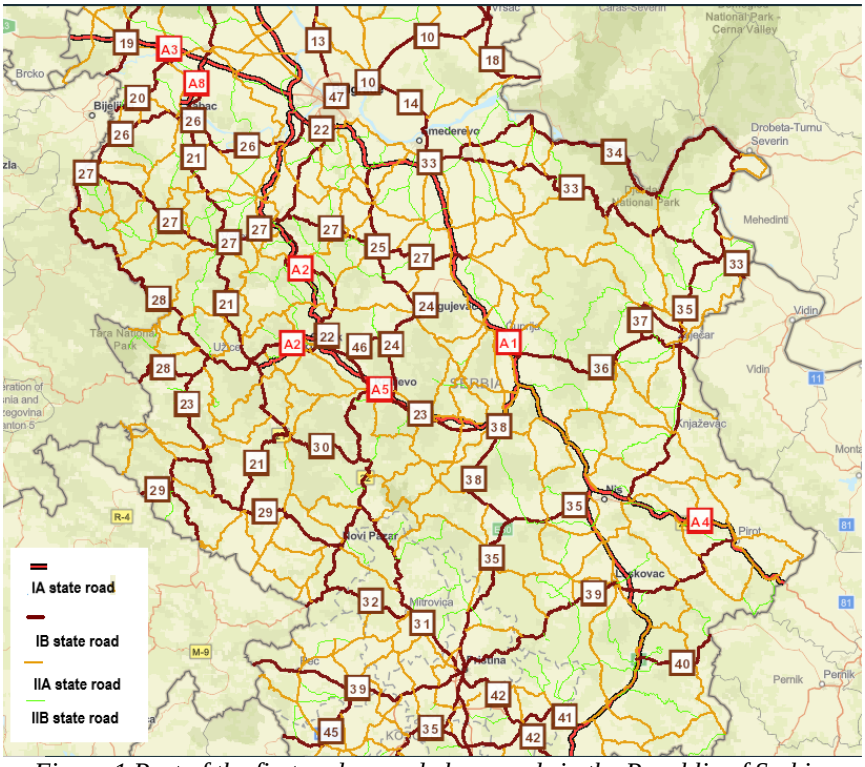

Figure 1 Part of the first and second class roads in the Republic of Serbia.

Section Length (km) - SL, the Annual average daily traffic volume - AADT (veh/day), and the number of traffic accidents are continuous variables, while Terrain type - TT (type 1-level, type 2-rolling, type 3-mountainous), Curvature (curve 1-minimal, curve 2-severe, curve 3-serpentine), and Lane width (5–6 m, >6 m) are categorical variables.The summary statistics of continuous variables are shown in Table 1.

Table 1 The summary statistics of continuous variables

	SL (m)	AADT (veh/day)	TA
Mean	8791.59	6464.72	9.97
SD	7799.58	4970.31	11.81
Min	200	244	0
Max	46900	25581	70
Var	6.08 x 107	2.47 x 107	139.44

As part of data preprocessing, the initial set is prepared and optimized for algorithm performance. Scaling of the numerical attributes (Section Length, Annual Average Daily Traffic Volume, and the number of traffic accidents) to a range of 0 to 1 is performed to achieve greater numerical stability, faster data processing, and increased model stability. Regarding categorical variables (Terrain Type, Curvature, and Lane Width), one-hot encoding is applied, which separates these variables into individual inputs based on their values. This results in obtaining 10 input attributes and one output. Outlier detection is also applied using the Isolation Forest method, which leads to noise reduction in the data and improved model performance. Six instances classified as outliers are detected and removed from the dataset, reducing the final set to 107 instances or rows in the table.

The prediction performances of the machine learning models are calculated using the correlation coefficient (r), the mean absolute error (MAE) and the root mean square error (RMSE):

$$\mathrm{r} = \frac{\sum_{k=1}^{N_s}\left(y_{pk}-\bar{y}_p\right)\left(y_k-\bar{y}\right)}{\sqrt{\sum_{k=1}^{N_s}\left(y_{pk}-\bar{y}_p\right)^2\sum_{k=1}^{N_s}\left(y_k-\bar{y}\right)^2}} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{N_s}\sum_{k=1}^{N_s}\left|y_{pk}(k)-y_k\right| \tag{8}$$

$$\mathrm{RMSE} = \frac{1}{N_s}\sqrt{\sum_{k=1}^{N_s}\left(y_{pk}-y_k\right)^2} \tag{9}$$

where y_{pk} and y_k denote the model output and the measured value respectively; $\bar{y}_p$ and $\bar{y}$ denote their average respectively, and N_s represents the number of observations in the data set.

The forecasting models have been implemented in Python.

A neural network with ten inputs is configured with a single hidden layer where the number of neurons is varied. After evaluating the model's performance with unipolar and bipolar sigmoid activation functions in the hidden layer, the best results are achieved with 16 neurons using the Rectified Linear Unit (ReLU) activation function. The network is trained using the Adam (Adaptive Moment Estimation) optimizer with mean squared error as the loss function over 100 epochs, with a batch size of 8, to predict the number of traffic accidents.

The SVR model achieved the best performance with the RBF kernel ($\sigma = 0.1$), regularization parameter C set to 100, and tolerance parameter ε set to 0.01, determined through the grid search methodology

The parameters of the random forest model are configured as follows: the number of candidate variables randomly selected for each split **(**max_features**)** is set to 'auto'; the sample size determining how many observations are randomly sampled for each tree **(**max_samples**)** is set to None; whether the sampling is done with or without replacement **(**bootstrap**)**is set to True; the minimum number of observations required in a terminal node **(**min_samples_leaf**)** is set to 1 and the criterion used for splitting nodes **(**criterion**)** is set to 'mse'. Through the application of the grid search method, the total number of trees in the ensemble (n_estimators) and the minimum number of observations required to split a node (min_samples_split) were adjusted to 150 and 4, respectively.

Table 2 presents the performance of the trained models and the hyperparameters adjusted for model optimization. Metric results are separately shown for the training and test sets.

Table 2 Performance parameters of the models and hyperparameters

Model	Data set	R	RMSE	MAE	Hyperparameters
Random forest	Training	**0.94**	**4.25**	**2.53**	n_estimators = 150, min_samples_split = 4
	Test	**0.87**	**4.63**	3.62	
SVM	Training	0.82	7.18	4.01	kernel='rbf', $C = 100$, $\varepsilon = 0.01$, $\sigma = 0.1$
	Test	0.79	5.65	3.96	
Neural netork	Training	0.85	6.35	3.98	optimizer='adam', loss='mean_squared_error', epochs=100, batch_size=8, number of neurons in the hidden layer=16, activation function of the hidden neurons: ReLU
	Test	0.84	4.83	**3.52**	

Based on the results from Table 2, it can be concluded that the best predictions for traffic accident frequency were achieved using the Random Forest (RF) algorithm.

Given that the Random Forest model achieved the best results among the three trained in this study, it is taken as representative and used to show the impact of input attributes on the output.

In this paper, the built-in Gini importance method is used to evaluate feature importance, known for its simplicity and efficiency in assessing the contribution of each feature in the model. After training the model, the importance of the features is extracted using the feature_importances_ attribute, which is part of the trained Random Forest model. The importance of the features is sorted in a created DataFrame in descending order and then visualized in a horizontal bar chart shown in Figure 2.

Figure 2 illustrates that both section length and annual average daily traffic volume exert the greatest influence on the frequency of traffic accidents. Following these factors, terrain type emerges as a significant parameter, particularly type 2 (rolling) and type 3 (mountainous). Additionally, the curvature attributes, particularly curve 2 (severe), curve 3 (serpentine), and curve 1 (minimal), also contribute significantly to accident frequency. At the very end is the

Lane width attribute, with 5–6 m followed by >6 m, between which TT type 1-level is positioned.

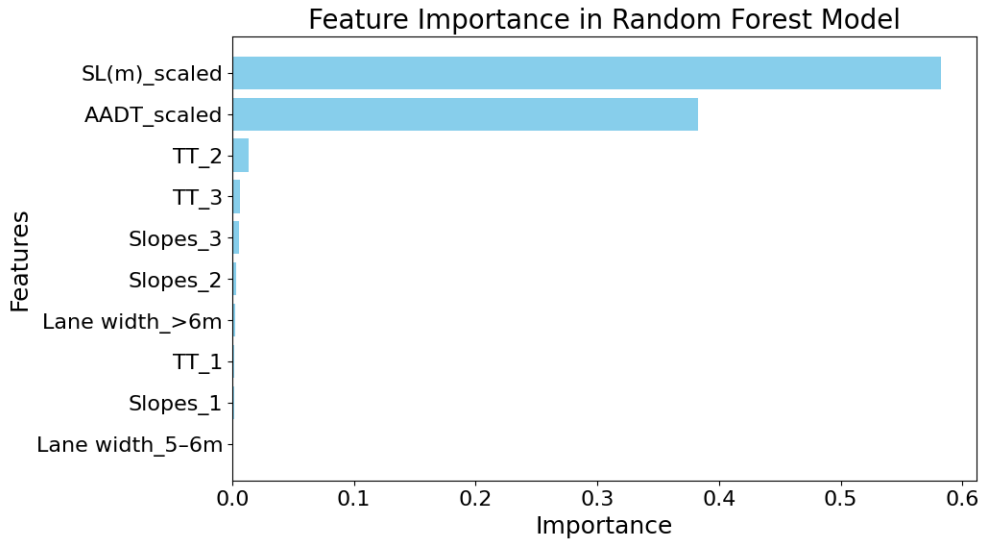

Figure 2 Feature importance in Random Forest Model

3 CONCLUSIONS

This study confirm the effectiveness of machine learning methods, especially support vector machines, neural networks and random forests, in predicting traffic accident occurrences. Evaluation of these models using established statistical metrics consistently demonstrates the superior predictive performance of random forest. For the training set, R, MAE, and RMSE are 0.94, 2.53, and 4.25 respectively, while for the test set, these values are 0.87, 3.62, and 4.63. Furthermore, the use of random forest for feature importance analysis provided detailed insight into the significant predictors affecting accident frequency.

These models are crucial for optimizing traffic management strategies by enabling precise prediction of risks and potential accidents on roads. Their application allows targeted allocation of safety resources, including more efficient deployment of police, emergency services, and other resources at locations where the likelihood of accidents is higher. Additionally, these models facilitate planning of preventive measures such as road improvements, additional signage, changes in traffic organization, or speed adjustments in specific road sections, all aimed at reducing accidents and enhancing the safety of all road users. Furthermore, they can significantly aid in the design and planning of future road infrastructure, ensuring safer and more effective road networks.

ACKNOWLEDGMENTS

This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract number 451-03-65/2024-03/200107 (Faculty of Engineering, University of Kragujevac).

REFERENCES

[1] Açıkkar, M., Altunkol, Y.: "A novel hybrid PSO- and GS-based hyperparameter optimization algorithm for support vector regression", Neural Computing and Applications, Vol. 35, 2023, 19961–19977.

[2] Ahmed, S., Hossain, M.A., Ray, S.K., Bhuiyan, M.M.I., Sabuj, S.R.: "A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance", Transportation Research Interdisciplinary Perspectives, Vol.19, 2023, 100814.

[3] Ali, Y., Hussain, F., Haque, M.M.: "Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review", Accident Analysis & Prevention, Vol. 194, 2024, 107378.

[4] AlMamlook, R.E., Kwayu, K.M., Alkasisbeh, M.R., Frefer, A.A.: "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity", 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, 272–276.

[5] Dong, C., Shao, C., Li, J., Xiong, Z.," An Improved Deep Learning Model for Traffic Crash Prediction", Journal of Advanced Transportation, Vol. 2018, Article ID 3869106, 2018, 13 pages.

[6] Fisa, R., Musukuma, M., Sampa, M., Musonda, P., Young, T.: "Effects of interventions for preventing road traffic crashes: an overview of systematic reviews", BMC Public Health, Vol. 22, 2022,1–18.

[7] García de Soto, B., Bumbacher, A., Deublein, M., Adey, B.T.: "Predicting road traffic accidents using artificial neural network models", Infrastructure Asset Management, Vol.5, No.4, 2018, 132–144.

[8] Gatarić D, Ruškić N, Aleksić B, Đurić T, Pezo L, Lončar B, Pezo M. "Predicting Road Traffic Accidents—Artificial Neural Network Approach. Algorithms", Vol 16. No. 5, 2023, 257.

[9] Gutierrez-Osorio, C., Pedraza, C.: "Modern data sources and techniques for analysis and forecast of road accidents: A review", Journal of Traffic and Transportation Engineering (English Edition), Vol. 7, No. 4, 2020, 432–446.

[10] Gorzelanczyk, P.: "Application of neural networks to forecast the number of road accidents in provinces in Poland", Heliyon, Vol. 9, No.1, 2023, e12767.

[11] Lord, D., Mannering, F.: "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives", Transportation Research Part A: Policy and Practice, Vol. 44, No. 5, 2010, Pages 291–305.

[12] Official Gazette: "Strategy for Traffic Safety of the Republic of Serbia for the Period 2023-2030 with the Action Plan for the Period 2023-2025" Official Gazette No. 84/2023, Belgrade. (In Serbian)

[13] Ojha, V.K., Abraham, A., Snášel, V.: "Metaheuristic design of feedforward neural networks: A review of two decades of research", Engineering Applications of Artificial Intelligence, Vol. 60, 2017, 97–116.

[14] P.E. Roads of Serbia. Traffic Counting. Available online: https://www.putevi-srbije.rs/index.php/en/traffic-counting (accessed on 20 June 2024).

[15] Probst, P, Wright, M.N, Boulesteix, A-L.: "Hyperparameters and tuning strategies for random forest", WIREs Data Mining and Knowledge Discovery, Vol. 9, No. 3, 2019, e1301.

[16] Raja, K., Kaliyaperumal, K., Velmurugan, L., Thanappan, S.: "Forecasting road traffic accident using deep artificial neural network approach in case of Oromia Special Zone", Soft Computing, Vol. 27, 2023, 16179–16199.

[17] Road Traffic Safety Agency of the Republic of Serbia. Integrated Database of Traffic Safety Features. 2022. Available online: http://195.222.99.60/ibbsPublic/ (accessed on 20 June 2024).

[18] Sánchez, A.V.D.: " Advanced support vector machines and kernel methods", Neurocomputing, Vol. 55, No. 1–2, 2003, 5–20.

[19] Singh, G., Pal, M., Yadav, Y., Singla T.: "Deep neural network-based predictive modeling of road accidents", Neural Computing and Applications, Vol. 32, 2020, 12417–12426.

[20] World Health Organization: "Global status report on road safety 2023", 2023.