TRAFFIC ACCIDENT RISK CLASSIFICATION USING NEURAL NETWORKS

Z. Purkrábková[*], *J. Růžička*[*], *Z. Bělinová*[*] *V. Korec*[†]

Abstract: The article deals with the current issue of traffic accident risk classification in urban area. In connection with the increase in traffic in the Czech Republic, a higher probability of risks of traffic excesses can be expected in the future. If there is a traffic excess in the city, the aim is to propose a meaningful traffic management solution to minimize the social losses. The main needs are the early identification and classification of the cause of the traffic excess, finding a suitable alternative solution, quick application of that solution, and the rapid ability to resume operations in the area of congestion. Traffic prediction is one of the tools for the early identification of traffic excess. The article describes extensive research focused on the classification and prediction of the output variable of accident risk based on own programmed neural networks. The research outputs will be subsequently used for the creation of a traffic application for a selected urban area in the Czech Republic.

Key words: *neural network, accident prediction, risk probability*

Received: March 24, 2021 **DOI:** 10.14311/NNW.2021.31.019
Revised and accepted: October 30, 2021

1. Introduction

Accidents are nowadays a frequent problem for many cities and we need to find some solution for that. A big trend in terms of the Smart City concept is the collection of traffic data from several possible sources and the use of processed data for designing traffic applications. The article describes extensive research on traffic accident classification and prediction in a selected area (specifically in the Czech town of Uherské Hradiště). The selection was made based on the availability of large data volumes from that location.

Verification of the possibility of using the neural network tool for the purposes of this task was the main motivation of the research team within the project "TAČR – ZETA program TJ01000183 – Prediction of traffic excesses using neural networks". As part of the team's research activities, a real application was created for the classification of the relative risk of an accident. The application was developed jointly by the commercial sector (Eltodo, a.s.).

[*]Zuzana Purkrábková, Jiří Růžička, Zuzana Bělinová; Czech Technical University in Prague, Faculty of Transportation Sciences, Konviktská 20, CZ-110 00 Praha 1, Czech Republic, E-mail: purkrzuz@fd.cvut.cz, ruzicji4@fd.cvut.cz, belinova@k620.fd.cvut.cz

[†]Vojtěch Korec; ELTODO a.s., Novodvorská 1010/14, CZ-142 00 Praha 4, Czech Republic, E-mail: KorecV@eltodo.cz

2. State of the art

The Department of Transport Telematics has been dealing with the possibility of using neural networks for transport problems. One of the first publications on the topic is the following [1]. This is one of the reasons why the aim is to verify the suitability of using this tool for accident risk classification. The use of neural networks is proven for various types of tasks such as prediction, classification, process control, or cluster analysis [10, 18].

From the point of view of the use of the artificial neural network (ANN) for classification, a summary review article can be cited [11]. The ANN with back-propagation is proposed in [12] is used for multispectral image classification. In the paper [13] a comparison of different neural networks are successfully carried out for the classification of wheat grains into three species of Kama, Rosa, or Canadian.

The article states that it seemed the easiest to use ANN for this purpose. A specific example of an article focusing on the estimation of traffic parameters can be named [17], which describes the possibility of predicting traffic intensity based on historical data. The following list of articles focuses exclusively on the use of neural networks for traffic classification and prediction, which is a timely topic. Many of the available resources come from the last few years and also from the other regions of the world. The conclusions of these articles are commented on in more detail below.

The article [2] focuses on predicting the severity of traffic accidents based on a historical data model from the Abu Dhabi region of 2007-13, and the developed neural network model has shown a relatively high accuracy of 74.6 %. The neural network model was compared to statistical models showing accuracy 59.5 %.

The article [3] focuses on the research on traffic accident prediction model based on convolutional neural networks (CNN) in VANET (vehicular ad-hoc networks). In this paper, the authors focus primarily on comparing the accuracy of predictions between CNN and the classical neural network. The paper found that CNN achieves greater accuracy of the prediction.

The article [4] focuses on prediction of traffic accidents on a certain section of the expressway in Malaysia. The article focuses on a similar issue as our research because the selected input parameters for the prediction model were the site of the accident, the reported cause of the accident, the condition of the road surface, the lighting, the type of collision and the time of the accident. This paper shows that the prediction model can determine the most frequent accidents in the area for the traffic data sample and can be more accurate than the conventional statistical models.

The paper [5] focuses on new proposed deep learning framework FMCNN (Factorization Machine Combined Neural Network). The experiment result shows that FMCNN outperforms other model DNN (Deep Neural Network).

The article [6] deals with the use of neural networks for prediction of traffic accidents in Jordan. The paper presents the analytical study and develops equations that help to control the behavior of growing traffic accidents.

The paper [7] focuses on similar problematic and similar issues as our research. With the difference that it does not deal with the probability of the risk of a traffic accident, but with the probability of the duration of a traffic accident. Based on the complete accident duration data, this paper describes two models based on multiple linear regression and two models based on artificial neural networks to predict the accident duration. The results of this research show that traffic condition, location type, accident type and police presence affect total duration significantly, while accident type and response time affect clearance time significantly.

State-of-the-art can be broadly divided into two approaches: general research on the development of neural networks and associated algorithms to refine prediction and classification in complex systems [16], and applied research on the applicability of existing proposed neural networks to a specific problem. Within the field of transport, the authors deal mainly with the potential use of neural networks as a tool for a real solution to the transport problem. Our research follows the second approach, where ANN was selected for a specific traffic problem and an available dataset.

3. Basic procedure

The basic scheme of the problem is shown in the Fig. 1. Large amount of traffic engineering, meteorological and other data (input variables) were used. The goal is to predict the only variable, which is the probability of risk of accident, using a multilayer neural network tool.

3.1 Processed data

We had several sources of traffic, meteorological and other related data. In order to successfully use this data as input variables for the neural network prediction model, it was first necessary to sort the data, process them appropriately and then categorize them.

In Tab. I you can see an overview of the input data to the neural network model. It was a total of 12 parameters (traffic intensity, average daily precipitation intensity, temperature, day, daytime, intersection, presence of alcohol at the culprit accident, type of road surface, state of road surface in the time of the accident, weather conditions in the time of the accident, visibility and sight conditions). All

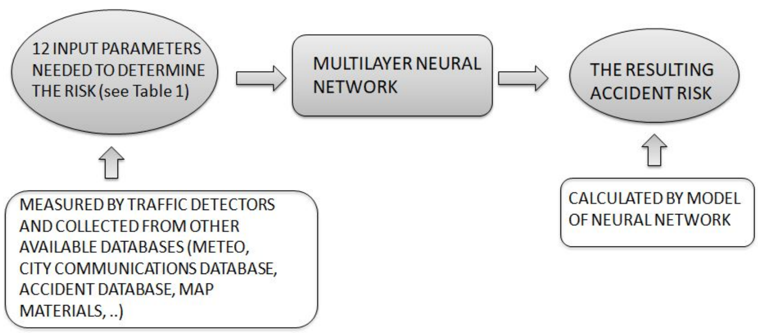

Fig. 1 *The basic scheme [authors].*

parameters take certain values and these values have been appropriately categorized for neural network purposes.

Obviously, the 12 parameters have a certain impact on the relative risk of a traffic accident, which will vary from parameter to parameter. Determining the importance of individual input variables for the model's predictive power is a complex problem. The solution will depend primarily on sufficient historical traffic data. We are interested in those sections that contain a combination of parameters that had a major impact on the occurrence of a traffic accident in the past. Then it is possible with some probability to predict the probability of the risks of similar events in the near future.

Description of parameters	Parameter designation and values
Traffic intensity	**P1** (0-10 veh/min, 10-20 veh/min, 20-30 veh/min, 30-40 veh/min, 40-50 veh/min)
Average daily precipitation	**P2** (very low, low, moderate, strong, very strong)
Daytime	**P3** (morning, afternoon)
Visibility	**P4** (in day, at night)
Presence of alcohol	**P5** (yes, no, under the influence of drugs, under the influence of drugs and alcohol, not identified)
Temperature	**P6** (-20 to -10, -10 to 0, 0 to 10, 10 to 20, 20 to 30)
Type of road surface	**P7** (pavement, bitumen, concrete, panels, gravel, other)
Outlook conditions	**P8** (good, bad, visibility covered by a stationary vehicle, other bad)
Intersection	**P9** (specific intersection in the city)
Surface state of road surface in the time of the accident	**P10** (dry, wet, mud, snow covered, spilled oil, continuous snow layer, sudden change in road conditions, other)
Day	**P11** (monday, tuesday, wednesday, thursday, friday, saturday, sunday)
Weather condition in the time of the accident	**P12** (unburdened, fog, light rain, rain, snow, frost, gusty wind, other burnered)

Tab. I *Overview of the input data and their descriptions.*

3.2 Categorization of relative accidents rate

The relative accident rate indicator is the most common used criterion for assessing road safety. Its value is predominantly indicative of the likelihood of an accident on a given road section in relation to driving performance. For road sections, the relative accident rate indicator is given by the formula [14]:

$$R = \frac{N_0}{365 * I * L * t} * 10^6$$

The following relationship applies to intersections:

$$R = \frac{N_0}{365 * I * t} * 10^6 \qquad (1)$$

where R is the value of the relative accident rate indicator (number of accidents / mil. vehicle-km and year), N_0 is the total number of accidents in the reporting period. I is the average daily traffic intensity (vehicles / 24 hours), L is the length of the section (km) and t is the reference period (years).

Values of the indicator are relative and usually range from 0.1 to 0.9. Values around 1 indicate minor shortcomings in terms of operational safety. Values higher than 1.6 indicate fundamental shortcomings. The values of the indicator of the relative accident rate can be entered into the map and for every day of week can be assigned its own data. A map of relative accident rates will provide a clear picture of traffic safety in a particular area.

This analysis was made based on a sample of historical data of accidents from the past 10 years. Data was assigned to eight intersections in Uherské Hradiště.

Data were categorized by quantile analysis from available data. The results of quantile analysis are shown below. In Fig. 2 you can see the area of interest in the yellow rectangle. A boxplot was used to illustrate the boundaries of the area of interest.

1. 25 % quantile: 0.08
2. 50 % quantile: 0.14
3. 75 % quantile: 0.51

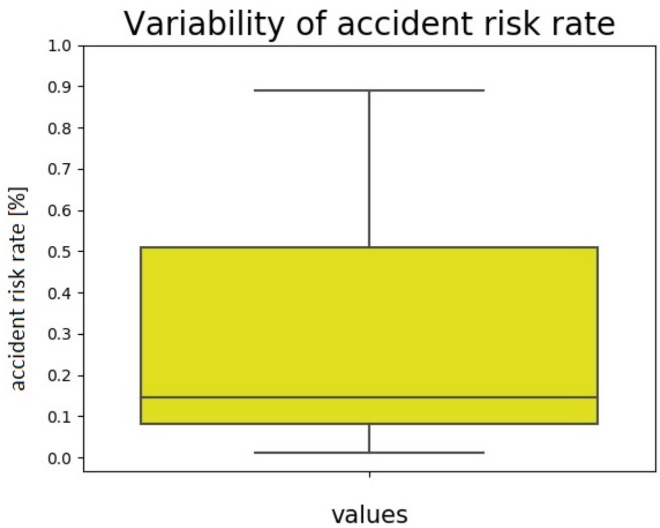

Fig. 2 *Variability of accident risk rate [authors]*

Therefore, the final categorization has the following form:

1. Very low risk of accident: (0.01–0.08)
2. Low risk of accident: (0.08–0.14)
3. Strong accident risk: (0.145–0.51)
4. Very strong accident risk: (0.51–0.89)

3.3 Neural network

The next step was to program our neural network. The neural network was programmed in Python. We used already available libraries for this purpose. Four of them played a key role in our model, namely:
#Pandas – library used to analyze data that can be represent by a 2D table.
#NumPy – a Python programming language library that provides an infrastructure for working with vectors, matrices, and generally multidimensional arrays.
#from sklearn import preprocessing – library for data transformation.
#sklearn.model_selection import train_test_split – library to divide data into training and test sets.
The accident risk prediction algorithm then works based on several basic steps:

- Retrieve data file with historical accident data.
- Shuffle data (a step that should reduce the risk of neural network overfitting).
- Load input variables (reading of variables based on accident risk prediction, variables can be added and removed, the model is adapted).
- Recode the output variable to numbers and then recode the numbers to dummy variables (explained below).
- Divide data into training and test data set (test data set is 15 % of total data set).
- Division of training data set into training and validation data.
- Neural network weights in interval $(-1, 1)$.
- Set the maximum number of learning periods to 400 (empirically proved the best) and initializing a neural network with a learning coefficient of 0.01 (a learning coefficient of 0.01 proved to be most effective in neural network training for our purposes).
- If the number of learning epochs is greater or equal to the maximum number of epochs, the neural network learning is terminated, otherwise, for each learning epoch:
  - Selection and submission of one neuron input vector from the training set.
  - Obtaining a neuron response (backpropagation method, evaluation of classification error based on comparison of actual and predicted output).
  - Correction of neuron weights based on error.
  - Evaluation of classification error across the entire training set. If the error is less than the acceptable limit, learning will end.
  - Evaluation of network activity success on the test set.
- Prediction of the output variable value vector for specific input variable values.

- Calculation of accuracy of predictions predicted by the model.

Generally, a multilayer neural network with an activation function was used for research purposes. We have previously discussed the choice of mathematical tool and comparison with other options. In our case, the transfer function of the backpropagation method is a sigmoid. This is mainly due to the fact that it has predicted the highest accuracy of predictions for our input variables, as shown in the Fig 3.

In addition to sigmoids, ReLU function, Leaky ReLU function and hyberbolic tangent function were tested on neural network. Leaky ReLU accuracy was the worst, the resulting prediction accuracy values were very unstable and varied a lot with the neural network epoch setting. The hyperbolic tangent and ReLU functions were better off, their accuracy increased along with the neural network set epoch. Sigmoid showed the highest accuracy even in a smaller number of set epochs. Since the choice of a suitable transfer function depends on experience, it cannot be certainly said that accuracy based on the selected activation function would be completely transferable. It is advisable to always try more options, compare these options and choose the one that shows the highest accuracy of predictions. Mathematical apparatus of a similar type of neural network was already described in previous publications [8, 9].

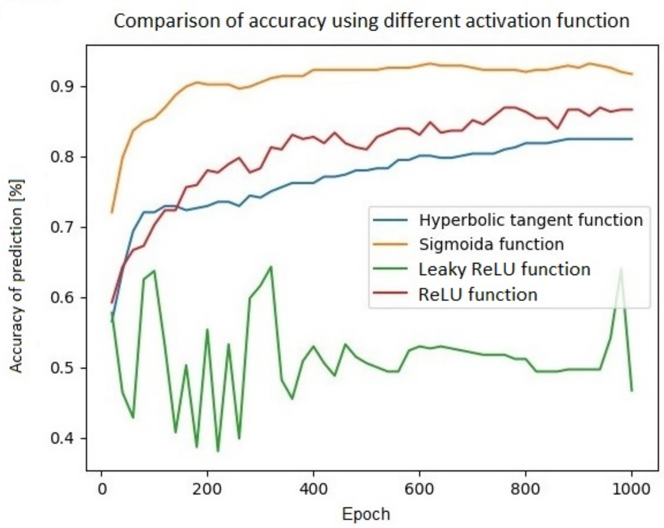

Fig. 3 *Comparison of accuracy of activation function [authors].*

3.4 Description of the working of the neuron network

As an input of the neural network there is a set of input variables:

$$P = (P_1, P_2, \dots P_N) \tag{2}$$

At the output from our neural network there is one output variable, which can have 4 values that represent the probability of the risk of a traffic accident:

$$R_A = (VeryLow, Low, Strong, VeryStrong) \tag{3}$$

In this case, these categorical variables cannot be directly pasted into plan matrices. A neural network predicts so-called *dummy variables*.

$$D = (D_1, D_2, D_3, D_4,) = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

For the values of the input variables according to the formula 2, the following output can be obtained using a neural network in the form of a vector presenting the probabilities of specific states:

$$Y = (0.45; 0.78; 0.65; 0.84)$$

The highest value in vector Y will be recoded to 1, the rest will be recoded to 0. After recoding we get:

$$Y_{DUMMY} = (0, 0, 0, 1)$$

The result can then be presented that for a given set of input variables the relative risk of accident in the area can be called "very strong".

4. Results of testing

From analysis resulted the following optimal neural network settings (this setting shows the highest accuracy of prediction): Multilayer neural network containing two hidden layers, in the first hidden layer there are 100 neurons, in the second hidden layer there are 50 neurons.

During empirical testing, it was found that given the inputs, the accuracy of the prediction slightly increased along with the number of neurons in each hidden layer, up to the stated optimal setting, which shows an accuracy of 85-90 %. With a higher number of neurons than 100/50, the accuracy of prediction no longer increases and on the contrary the complexity of algorithm calculation significantly increases. Also two hidden layers were empirically determined, another number of hidden layers provided lower prediction accuracy.

Accuracy of prediction of our own neural network for traffic accident classification and prediction was compared with accuracy of prediction with other commonly used methods. Specifically, the research involved methods of multiple regression and the use of decision trees. The results of the prediction are clearly shown in Fig. 4. Individual algorithms were evaluated based on the same inputs.

Comparison of the model using multiple regression was possible due to the fact that there no large correlations were demonstrated for individual input parameters, so they can be considered independent. The condition that the predictors are not constant was met. When evaluating multiple regression, calibration was performed based on the calibration data set as in the case of neural networks.

The use of multiple regression showed the lowest accuracy of all the methods used, the accuracy is at values just below 50 %, even though we tried to calibrate the model. This can be explained either by the excessive nonlinearity of the problem. The second possibility is the existence of a significant external correlating variable, which is not included in the model. This can be, in particular, the effects on accidents from the point of view of driver behavior, which the model does not take into account at all.

The use of decision trees shows much greater accuracy between 70-80 %. The accuracy in this case depends on the number of decision trees and the depth of each decision tree. A random decision tree consists of several decision trees and its resulting accuracy is the average value of the accuracy of each tree. It can be stated that, together with the number of decision trees and the growing depth of the decision tree, the resulting prediction accuracy also slightly increases. In the case of neural networks, the accuracy of prediction ranges from 85-90 %.

The neural network was tested using sensitivity analysis, which is further explained in another author's publication [15]. Further use of the neural network for traffic column prediction was also investigated.

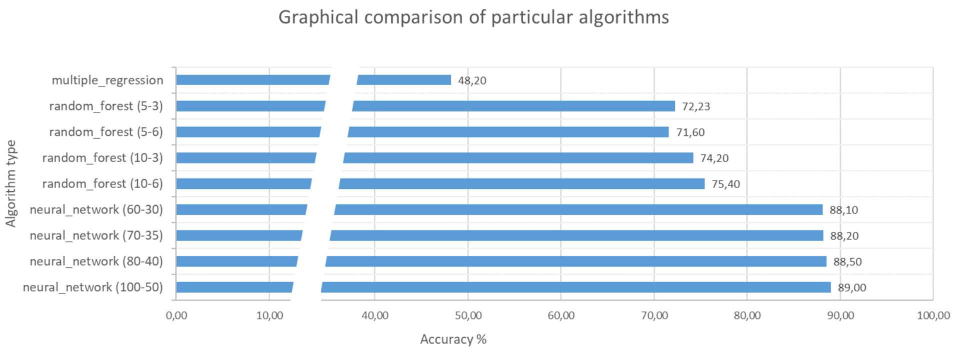

Fig. 4 *Comparison of neural network [authors].*

5. Conclusion and future work

Nowadays there are lots of data collected, but it is important to look for their meaningful interpretations and use. In prediction models, large amounts of data are processed beyond practical targets. The number of cars in cities is constantly increasing, so there will be an effort to eliminate any risk of traffic excesses and to inform road users on time.

Our research dealt with the possibilities of predicting and classification the relative risk of traffic accidents with real input data using neural networks. The neural network was programmed in Python and tested on accident data from Uherské Hradiště. The accuracy of neural network prediction was compared with the precision of multiple regression and decision tree predictions, with neural networks showing the highest accuracy. At the same time, it has been shown that the accuracy of prediction increases slightly with the number of neurons in the individual layers of the neural network

Currently, there is a practical output of our research – an application that predicts and classifies the probability of accident risk based on input data. The data is entered into the system by the user of the application. The user gets a detailed overview of the parameters with the greatest impact on traffic accidents for the area. The application also detects imperfections in the neural network model. The application does not allow the user to enter meaningless combinations of input parameters that can never occur. (Example: temperature 30 degrees and snow on the road). Fig. 5 shows an example of the application, unfortunately only in Czech. In the left drop-down menu or on the right in the map, the user selects the appropriate intersection, enters input parameters for it and the neural network classifies the probability of accident risk based on the entered parameters.

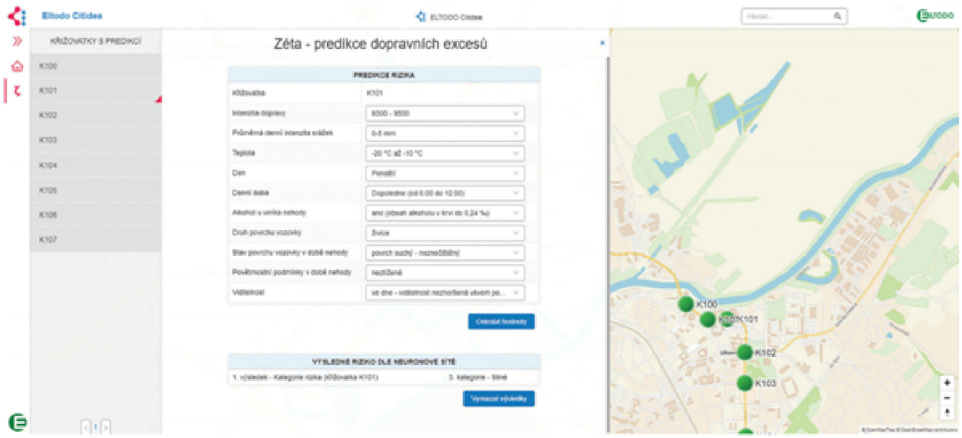

Fig. 5 *Created transport application [authors].*

In the future, it would be possible to provide similar information directly into operation, for example, through the variable message signs. Accurate prediction of probability of risk of traffic accidents and timely warning of drivers will have positive consequences on the safety of traffic flow. The goal of our research, the verification of the possibility of using neural networks for the classification of traffic accidents, was met. Although the proposed application, including the module of classification and prediction of traffic excesses, is not completely self-saving. It can be one of many tools for streamlining the solution of emergency situations in traffic crisis management.

Acknowledgement

The article and research were supported from the TAČR – ZETA program TJ01000183 – Prediction of traffic excesses using neural networks.

References

[1] NOVÁK M., KUFUDAKI O., PELIKÁN E., ŠEBESTA V., MOOS P., MUSÍLEK P. *Umělé neuronové sítě (in Czech) Artificial neural networks.* C.H. Beck, Prague, 1998.

[2] ALKHEDER S., TAAMNEH M., TAAMNEH S. *Severity Prediction of Traffic Accident Using an Artificial Neural Network*. Journal of Forecasting [online], 2017, 36(1), pp. 100–108 [date of access: 22.01.2019], doi: 10.1002/for.2425.

[3] ZHAO H., CHENG H., MAO T., HE Ch. *Research on Traffic Accident Prediction Model Based on Convolutional Neural Networks in VANET*. In: *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, [online]. IEEE, 2019, pp. 79–84 [date of access: 09.10.2019], doi: 10.1109/ICAIBD.2019.8837020.

[4] SAMEEN M., PRADHAN B. *Severity Prediction of Traffic Accidents with Recurrent Neural Networks*. Applied Sciences [online], 7(6), 2017 [date of access: 22.01.2019], doi: 10.3390/app7060476.

[5] ZHAO H., MAO T., DUAN J., WANG Y., ZHU H. *FMCNN: A Factorization Machine Combined Neural Network for Driving Safety Prediction in Vehicular Communication*. IEEE Access [online]. 2019, 7, pp. 11698–11706 [cit. 2019-10-09], doi: 10.1109/ACCESS.2019.2891619.

[6] JABAR Y., RABABAA M. *Neural Technique for Predicting Traffic Accidents in Jordan*. Journal of American Science, 9, pp. 347–358, 2013.

[7] ZHANG J., WANG J., FANG S. *Prediction of urban expressway total traffic accident duration based on multiple linear regression and artificial neural network*. In: *2019 5th International Conference on Transportation Information and Safety (ICTIS) [online]* [online], IEEE, 2019, pp. 503–510, doi: 10.1109/ICTIS.2019.8883690.

[8] BISKUP R. *Možnosti neuronových sítí (in Czech) Neural network options*. Czech University of Life Sciences, Prague, 2009. PhD thesis

[9] CATERINI, Anthony: *A Novel Mathematical Framework for the Analysis of Neural Networks* [online]. Canada, 2017 [viewed 29 october 2019]. Available from: https://uwspace.uwaterloo.ca/bitstream/handle/10012/12173/caterini_anthony.pdf?sequence=3&isAllowed=y.

[10] ARBIB M. *The Handbook of Brain Theory and Neural Network*. Massachusetts Institute of Technology, 2003.

[11] BALA R., KUMAR D.: *Classification Using ANN: A Review International Journal of Computational Intelligence Research* [online], pp. 1811–1820, 13 (7), ISSN 0973-1873. Research India Publications, 2017 [viewed 23 February 2021]

[12] HEERMANN P.D., KHAYENIE N. *Classification of multispectral remote sensing data using a back-propagation neural network*, pp. 81-–88, 30(1), IEEE Trans. Geosci. Remote Sens., 1992.

[13] KAYABASI A., TOKTAS A., SABANCI K., YIGIT E. *Automatic classification of agricultural grains: comparison of neural Neural Network World* [online] pp. 213–224, 28(3), 2018, ISSN 12100552, doi: 10.14311/NNW.2018.28.013.

[14] ANDRES J., MIKULÍK J. *Metodika identifikace a řešení míst častých dopravních nehod (in Czech) Methodology of identification and solution of frequent traffic accident sites* [online], Brno: Centrum dopravního výzkumu, ISBN 80-902141-9-3, 2001 [viewed 23 February 2021].

[15] RUZICKA J., PURKRABKOVA Z., KOREC V. *Smart application for traffic excess prediction* In: 2020 Smart City Symposium Prague (SCSP) [online], IEEE, ISBN 978-1-7281-6821-0, 2020 [viewed 23 February 2021], doi: 10.1109/SCSP49987.2020.9133935.

[16] PELIKÁN E. *Forecasting of processes in complex systems for real-world problems: tutorial*. Neural Network World [online], 2014, 24(6), pp. 567–589, ISSN 12100552 [viewed 23 February 2021], doi: 10.14311/NNW.2014.24.032.

[17] SIDDIQUE M.S.A., HOQUE S. *Predicting the daily traffic volume from hourly traffic data using artificial neural network*. Neural Network World [online], 2017, 27(3), pp. 283–294, ISSN 12100552 [viewed 23 February 2021], doi: 10.14311/NNW.2017.27.015.

[18] BISHOP C.M. *Neural networks for Pattern Recognition* [online]. Oxford: Oxford University Press, 1995 [viewed 23 February 2021]. Available from: https://uwspace.uwaterloo.ca/bitstream/handle/10012/12173/caterini_anthony.pdf?sequence=3&isAllowed=y.