Exploring the Effect of the Variables on Machine Learning Prediction Models

Zainab Ali Mohammed
Department of Computer Engineering
University of Technology, Baghdad, Iraq

Mohammed Najm Abdullah
Department of Computer Engineering
University of Technology, Baghdad, Iraq

Imad Hussain Al-Hussaini
University of Baghdad, Baghdad, Iraq

Abstract- Unexpected events such as traffic incidents, fire, road maintenance, police activities, debris, and weather alarms are still very common, random, and dangerous. They may temporarily reduce roadway capacity to existing traffic conditions, as they the reasons behind the non-recurrent congestion. The consequent reduced speed and queue formation can increase the potential for secondary incidents. The duration of the incident is affected by various factors, knowing these factors and analyze their impact is also a priority in the traffic management process. This paper investigates different machine learning methods to predict the duration of the incident. Support Vector Regression, Random Forest, and Neural Network Multi-Layer Perceptron methods were used to build the incident duration prediction models. Mean squared error, Root mean squared error, and Mean absolute error were used for the evaluation of these models. Furthermore, variable importance technique was used to obtain the variable importance scores, where it determines the relative importance of each input variable used to train the prediction models with respect to the predictability of the response variable. The ML findings showed similar performance for the models. However, SVR model slightly outperforms the RF, and MLP models. MAE for SVR model was 14.23 min. Whereas, RF models slightly outperformed the other two models given RMSE of 18.91 min for the RF model. The variable importance technique results showed that the performance of the model downgraded with fewer variables.

Keywords – Incident Duration, Machine Learning, Prediction, Variable Importance

I. INTRODUCTION

Traffic congestion is now a major aspect of large towns and small cities across the world. Virtually every motorist experiences significant delays at some point in time and wastes a massive amount of time, fuel, and money on the road. An additional vehicle imposes added delays on other drivers, resulting in economically excessive traffic volumes. Non-recurrent congestion caused by traffic incident, which is one of the two main types of the traffic congestion, is difficult to predict but should be dealt with in a timely and efficient manner to minimize its effect on road capacity reduction and enormous travel time loss. The longer an incident scene is in place, the higher is the likelihood for secondary incidents. Total time for an incident to be cleared can be further increased by the occurrence of secondary incidents and the travelers may experience ever growing congestion [1]. Influence factor analysis and reasonable prediction of traffic incident duration are essential in traffic incident management in order to predict incident impacts and aid in the implementation of appropriate traffic operation strategies. Examples of such factors include number of vehicles involved, roadway environment, and weather condition. A well-planned and coordinated traffic incident management process can improve safety, traffic flow, and clearance times [2].

R. Li et al., 2015 [3], H. Park et al., 2016 [4], L. Lin et al., 2016 [5], K. Hamad et al., 2018 [6] developed models to predict the duration of the incident. C. H. Wei and Y. Lee, 2007 [7], investigated an adaptive procedure with two adaptive Artificial Neural Network-based models to subsequent prediction of total incident duration time. Data were collected from Taiwan during the time from November 2004 to April 2005 with a total of 24 incidents. Incident data were split into 75% and 25% to train and validate the models. Mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE) were utilized for the models' evaluation, and particularly to compute the model's accuracy. It was found that this type of prediction from two models gives

better estimation, where the MAE was 409.7 min, the MAPE was 30.3, and the RMSE was 467.3 min. A. T. Hojati et al., 2013 [8], proposed different models of the Parametric Accelerated Failure Time (AFT) which are log-logistic, lognormal, and Weibull with fixed and random parameters, also a Weibull model with gamma heterogeneity, these were applied on one year collected data from Australian freeway network with a total of 3251 incidents. Considering three incident types which are, crashes, hazards, and stationary vehicles on weekdays, with 28 variables investigated. The findings demonstrated that Weibull models with random parameters were more applicable for crashes and hazard incidents, while the stationary vehicle incidents were better dealt with by the Weibull model with gamma heterogeneity. A. T. Hojati et al., 2014 [9], proposed two Hazard-based models, log-logistic accelerated failure time (AFT) and Weibull accelerated failure time (AFT) to predict the incident duration. Data from Southeast Queensland, Australia with twenty-nine variables include information about incident specifics, infrastructure, features of measured traffics, and temporary effects were used to build the models. The developed models showed that the Weibull AFT model with gamma heterogeneity is more suitable for crash incidents, where the log-logistic AFT model with random parameters gives better estimation for hazard and parked vehicle incidents. J. Tang et al., 2020 [10], investigated the gradient boosting machine method (XGBoost) for estimating the clearance time of the incident and to evaluate influencing factors. Washington Incident Tracking System data collected during 2011 were used to develop the models with a total of 2565 incident records. Random forest (RF), support vector regression (SVR), and Adaboost were used for further investigation. MAPE was utilized for the evaluation of the models performance. Findings showed that XGBoost outperforms the other three models with MAPE of 0.348 and 0.221. The purpose of this paper is to investigate different machine learning methods to predict the duration of the incidents, also to identify the most important variables influencing the prediction models, and investigate the effect of reducing the number of the variables on the performance of the attempted models.

The remainder of this paper is organized as follows: Section II introduces the methodology of the proposed incident duration prediction models, also the variable importance technique. A brief description of the error indexes that are used to evaluate the performance of the prediction models is shown in Section III. Next, the results obtained from the three models, and the results of the variable importance technique are discussed in Section IV. Finally, the conclusions are provided in Section V.

II. METHODOLOGY

Traffic incident logs and traffic sensors data from Eastbound Interstate 70 (I-70) in Missouri, United States were collected over the period from January 2015 to January 2017, with a total of 352 incidents, were used to develop the prediction models to predict the duration of the incidents. Three machine learning (ML) methods, namely, support vector regression (SVR), random forest (RF), and neural network multilayer perceptron (MLP) were used to build the incident duration prediction models. Below is a brief description to each one.

2.1. Support Vector Machine

Support Vector Machines (SVMs) are supervised learning models with related learning techniques which analyze data utilized for the classification and the regression analysis. SVM is one of the most popular models in machine learning in the context of statistical learning theory [11]. SVM is a very robust and flexible ML model, that may execute linear or nonlinear classification, regression, as well as outlier detection. SVMs are particularly suitable for small- or medium-sized datasets [12]. An SVM estimator (*f*) on regression can be expressed as:

$$f(x) = w \cdot \emptyset(x) + b \qquad (1)$$

where $\emptyset$ is a nonlinear transfer function, W is the weight and b is the offset.

2.2. Random Forests

Random forest (RF) is a supervised learning algorithm, it's a flexible, easy to use machine learning method that gives a great result most of the time even without hyper-parameter tuning. And because of its clarity and diversity (it can be used for both classification and regression tasks), It is one of the most used algorithms. Random forests are an efficient tool in prediction. Using the right kind of randomness makes them proper classifiers and regressors. Because of the Law of Large Numbers, they do not overfit, and this is one of the main differences between the Random Forests and the Decision Trees where the latter might suffer from overfitting, random forest prevents this by creating random subsets of the features and uses these subsets to builds smaller trees Afterwards, it combines the subtrees, but this doesn't work every time and it also makes the computation slower, depending on the number of trees built by the random forest. The RF regression prediction can be expressed as:

$$y_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} y_b(x) \qquad (2)$$

Where $y_b(x)$ is the value predicted by the b^{th} tree of the forest, x is the predictor variables, B is the number of trees in the forest, and $y_{rf}^{B}(x)$ is the average of the values predicted by the trees, as this average is the final value predicted by the forest [13] [14].

2.3. Neural Network (Multi-layer Perceptron)

Neural Networks are a collection of algorithms that imitate the processes of a human brain to find relationships between vast amounts of datasets. The Neural Network is a network or circuit of neurons, or an Artificial Neural Network in a modern sense, composed of artificial neurons or nodes. The artificial networks can be utilized for predictive modeling, adaptive control, and applications where they can be trained using the dataset. In networks, self-learning that results from experience may draw conclusions from a complex and unrelated set of information [15].

A Multi-Layer Perceptron (MLP) (Hornik, Stinch combe, & White, 1989) is a class of artificial feedforward neural network (ANN). The MLP composes of over one perceptron, that consist of an input layer to receive the signal, an output layer which makes a decision, or an estimation about the input, and a random number of hidden layers between these two types which represent the MLP 's actual computational engine [16].

2.4. Variable Importance Technique

It cannot be assumed that all the input factors used in the model are equally important to predict the duration of the incident. Generally, a few can have a significant effect on the incidents duration, whereas, many may be insignificant and therefore can be omitted when modeling the incidents duration [17]. The purpose of this section is to know the relative importance or contribution of every input factor when estimating the target (that is, the duration of the incident). The findings of the study provided in this section will help to specify the most relevant variables influencing the duration of the incident to be collected; On the contrary, other variables might be optionally gathered for this purpose.

Random Forest (RF) method is used to compute the variable importance as presents in Figure (1), that shows the most important variables (from the dataset that are used in this research) affecting the estimation of the incident duration, and these variables that have less effect on model performance. The modeling tool (H2O) is used to perform this technique with the RF method.

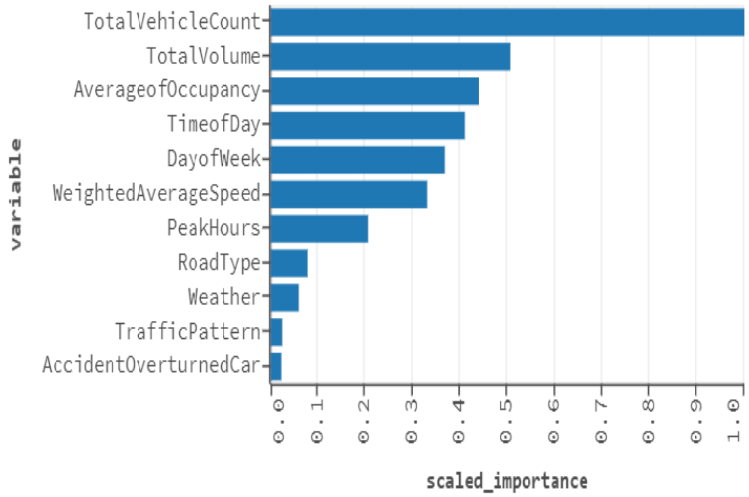

Figure 1. Predictor Variables Importance

From the Figure (1), it can be seen that the most influencing variables are: total vehicle count, total volume, average of occupancy, day time, day of week, and weighted average speed, while the other variables have less impact. These findings are further used to determine the impact of reducing the number of factors on the output of the attempted SVR, RF, and MLP models. Where the performance of the SVR, RF, and MLP models using all variables is compared with the performance of other SVR, RF, MLP models which are developed using the most affecting variables.

III. ERROR INDEXES

The mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) error indexes were used to evaluate the incidents duration prediction models developed in this study.

The mean squared error (MSE) of a predictor measures the average of the squares of the errors—that is, the average squared difference between the predicted values and the real value. MSE is defined in Equation below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \quad (3)$$

Where MSE is the mean squared error, n is the number of incident records, y_i is the observed value, and $\hat{y_i}$ is the predicted value [18].

The root mean squared error (RMSE) is a widely used measurement of the differences between the values that estimated by a model or a predictor and the observed values. RMSE is defined in Equation below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (4)$$

Where, RMSE is the root mean squared error, n is the number of incident records, y_i is the observed value, and $\hat{y_i}$ is the predicted value [19].

The mean absolute error (MAE) is a measure of difference between two continuous variables, which are the measured value and the true value. MAE is a model evaluation metric used with regression models, it is an average of the absolute errors, and it is defined in Equation below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \quad (5)$$

Where MAE is the mean absolute error, n is the number of incident records, y_i is the observed value, and $\hat{y_i}$ is the predicted value [20].

IV. RESULTS

4.1. Machine Learning Prediction Models Results

Incident duration prediction models are developed using SVR, RF, and MLP methods. Table (1) illustrates the MSE, RMSE, and MAE error indexes that have been used to evaluate each model for training, testing, and prediction datasets.

Table 1. SVR, RF, and MLP Models Performance Indexes

Model	Category	MSE (min)	RMSE (min)	MAE (min)
SVM	Training	638.84	25.27	14.39
	Testing	338.93	18.40	13.87
	Prediction	548.75	23.42	14.23
RF	Training	340.87	18.46	14.14
	Testing	425.62	20.63	16.34
	Prediction	357.96	18.91	14.58
MLP	Training	582.29	24.13	15.54
	Testing	350.65	18.72	15.14
	Prediction	512.71	22.64	15.42

Results showed that the SVR, RF, and MLP models, for the training, testing, and prediction datasets, have almost the same error range in terms of RMSE and MAE. Where it can be noticed that the SVR model slightly outperformed the other two models, for the prediction dataset. The SVR model scored MAE of 14.23 min as compared to 14.58 min, and 15.42 min MAE for the RF and MLP models, respectively.

On the other hand, in terms of RMSE indexes for the prediction dataset of these three methods, it has been shown that RF model slightly outperformed the SVR and MLP models, given RMSE of 18.91 min for the RF model as compared to 23.42 min and 22.64 min RMSE for the SVR and MLP models, respectively.

4.2. Variable Importance Technique Results

This section illustrates the implementation results of the attempted ML models based on the variable importance findings to study the effect of reducing the number of factors on the model's performance. MSE, RMSE, and MAE indexes are also utilized to evaluate the performance of the attempted SVR, RF, and MLP models for training, testing, and prediction datasets, as given in Table (2). Figure (2) illustrates the prediction RMSE, and MAE indexes for SVR, RF, and MLP models with the reduced variables. The results show that the performance of the three ML models is downgraded after the input variables have been reduced, where the error indexes slightly increase when only the top six variables have been selected to train and test the models, as shown in Figures (3), (4), and (5).

Table 2. SVR, RF, and MLP Performance Indexes

Model	Category	MSE (min)	RMSE (min)	MAE (min)
SVR	Training	673.54	25.95	15.04
	Testing	357.31	18.90	14.23
	Prediction	578.54	24.05	14.80
RF	Training	343.27	18.52	14.20
	Testing	419.06	20.47	16.28
	Prediction	358.55	18.93	14.62
MLP	Training	613.11	24.76	16.00
	Testing	360.10	18.97	15.11
	Prediction	537.11	23.17	15.74

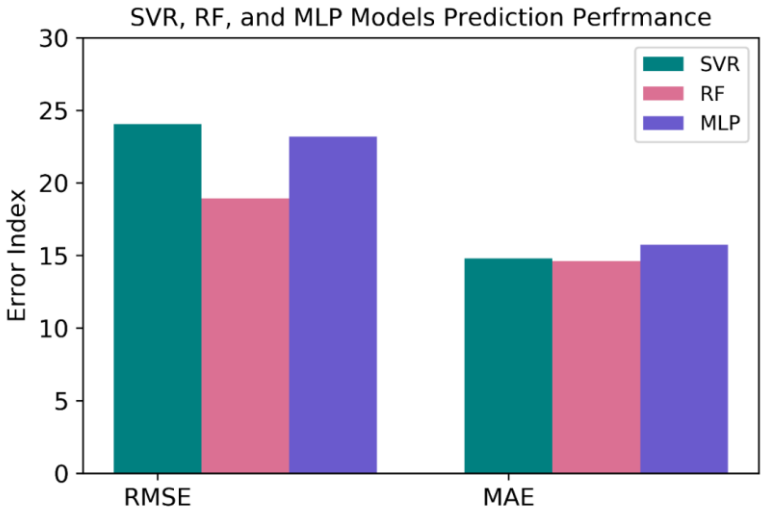

Figure 2. SVR, RF, and MLP Models Prediction Performance

Figure (3) illustrates the prediction RMSE, and MAE error indexes for SVR models performance comparison.

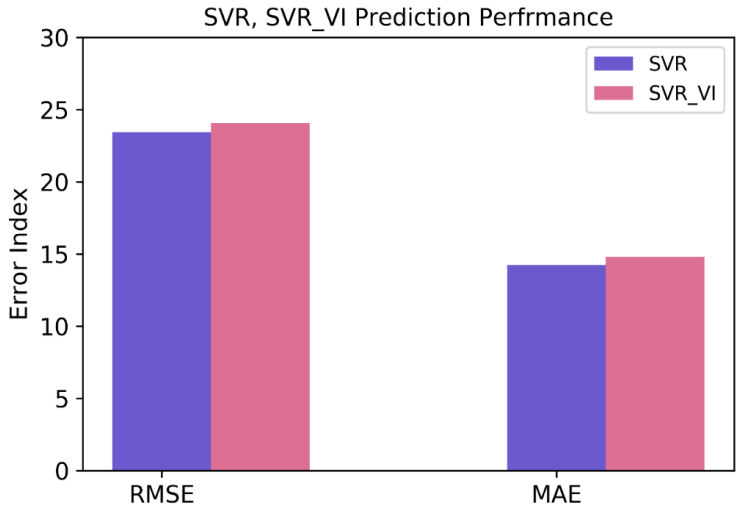

Figure 3. SVR Models Comparison

Figure (4) illustrates the prediction RMSE, and MAE error indexes for RF models performance comparison.

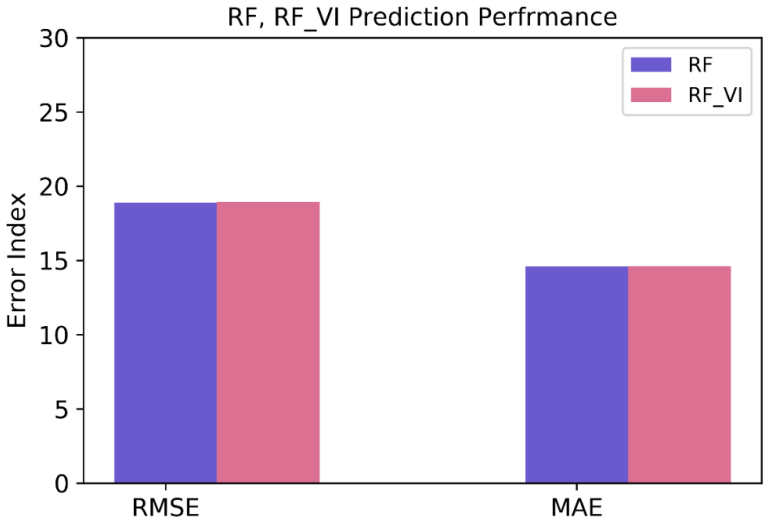

Figure 4. RF Models Comparison

Figure (5) illustrates the prediction RMSE, and MAE error indexes for MLP models performance comparison.

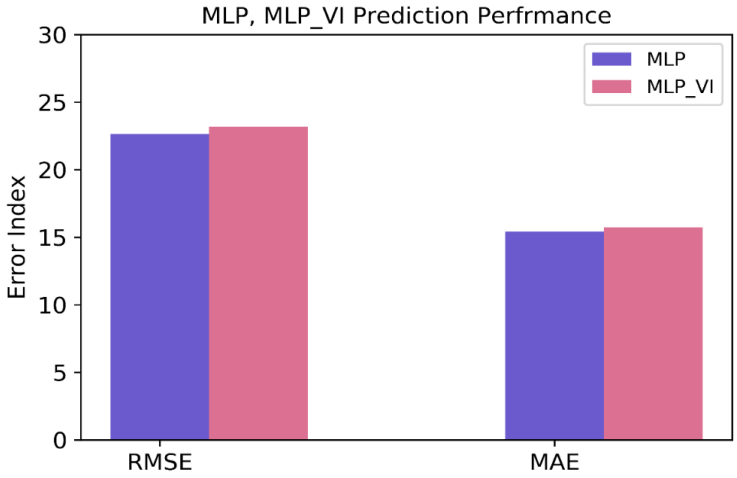

Figure 5. MLP Models Comparison

According to the Figures (3), (4), and (5), it can be noticed that in SVR model with all input factors included, the MSE, RMSE, and MAE indexes (for the prediction dataset) are 548.75 min, 23.42 min, and 14.23 min, respectively. While for SVR model with the reduced factors, the MSE index is 578.54 min, RMSE index is 24.05 min. and the MAE index is 14.80 min, which shows only slight differences in the error indexes. Likewise, for the RF models, the one with all input factors included and the one with the reduced factors, the MSE, RMSE, and MAE indexes are slightly increased by 0.59 min, 0.02 min, and 0.04 min, respectively, in the reduced factors RF model in comparison to the RF model that used all the input factors (for the prediction dataset). Moreover, the MLP model with all input variables included, the MSE index is 512.71 min, RMSE index is 22.64 min, and the MAE index is 15.42 min (for the prediction dataset). While, the MLP model with the reduced variables, the MSE, RMSE, and MAE indexes are 537.11 min, 23.17 min, and 15.74 min, respectively. This also means that the error indexes slightly increase with fewer variables, thus the models' performance downgrades.

V. CONCLUSIONS

This paper proposes different prediction models to predict the duration of the incidents. The models are developed and validated on the basis of Support Vector Regression (SVR), Random Forest (RF), and Neural Network Multi-Layer Perceptron (MLP) methods. The ML results show that the performance of these models is comparable. The effect of the input variables on the models is tested using variable importance technique. Where it is used to find the variables that have the most impact on the model's performance. Total vehicle count, total volume, average of occupancy, day time, day of week, and the weighted average speed, are the most important variables that have the most influence on the prediction models. The selected variables are further used to determine the impact of reducing the number of input variables on the performance of the best performing SVR, RF, and MLP models. The results show that the performance of the three ML models is degraded after the input variables have been reduced, where the error indexes are slightly increased when only the top six variables have been utilized for the training and validation of the models.

REFERENCES

[1] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Overview of traffic incident duration analysis and prediction", Eur. Transp. Res. Rev., vol. 10, no. 2, pp. 1–13, 2018.

[2] K. Hamad, R. Al-Ruzouq, W. Zeiada, S. Abu Dabous, and M. A. Khalil, "Predicting incident duration using random forests", Transp. A: Transp. Sci., vol. 16, no. 3, pp. 1269–1293, 2020.

[3] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Competing risks mixture model for traffic incident duration prediction", Accid. Anal. Prev., vol. 75, pp. 192–201, 2015.

[4] H. Park, A. Haghani, and X. Zhang, "Interpretation of Bayesian neural networks for predicting the duration of detected incidents", J. of Intelligent Transp. Sys., Vol. 20, No. 4, pp. 385-400, 2016, https://doi.org/10.1080/15472450.2015.1082428.

[5] L. Lin, Q. Wang, and A. W. Sadek, "A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations", Accid. Anal. Prev., vol. 91, pp. 114–126, 2016.

[6] K. Hamad, R. Al-Ruzouq, W. Zeiada, S. Abu Dabous, and M. A. Khalil, "Predicting incident duration using random forests", TRB 2018 Annual Meeting, 2018.

[7] C. H. Wei and Y. Lee, "Sequential forecast of incident duration using Artificial Neural Network models", Accid. Anal. Prev., vol. 39, no. 5, pp. 944–954, 2007.

[8] A. T. Hojati, L. Ferreira, S. Washington, and P. Charles, "Hazard based models for freeway traffic incident duration", Accid. Anal. Prev., vol. 52, pp. 171–181, 2013.

[9] A. T. Hojati, L. Ferreira, S. Washington, P. Charles, and A. Shobeirinejad, "Modelling total duration of traffic incidents including incident detection and recovery time", Accid. Anal. Prev., vol. 71, pp. 296–305, 2014.

[10] J. Tang, L. Zheng, C. Han, F. Liu, and J. Cai, "Traffic Incident Clearance Time Prediction and Influencing Factor Analysis Using Extreme Gradient Boosting Model," J. Adv. Transp., vol. 2020, 2020.

[11] C. Cortes and V. Vapnik, "Support-Vector Networks", Springer, vol. 20, pp. 273-297, September 1995.

[12] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow", First Edition, O'Reilly Media, USA: 2017.

[13] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests", IEEE, vol. 20, no. 8, pp. 832–844, 1998.

[14] S. Hartshorn, "Machine Learning with Random Forests and Decision Trees", Kindle, 2016.

[15] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", PNAS, vol. 79, no. 8, pp. 7–19, 2018.

[16] C. Gallo, "Artificial neural networks Tutorial", IGI Global, pp. 1–426, 2015, DOI: 10.4018/978-1-4666-5888-2.ch626.

[17] M. Kuhn, K. Johnson, "Applied Predictive Modeling", Springer, New York: 2013.

[18] Z. Wang and A. C. Bovik, "Mean Squared Error: Love It or Leave It?", IEEE Signal Process. Mag., vol. 26, pp. 98–117, January, 2009.

[19] R. G. Pontius, O. Thontteh, and H. Chen, "Components of information for multiple resolution comparison between maps that share a real variable", Environ. Ecol. Stat., vol. 15, no. 2, pp. 111–142, 2008.

[20] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", Inter-Res. Sci., vol. 30, no. 1, pp. 79–82, 2005.