Traffic Flow Prediction Based on Explainable Machine Learning

Xueting Zhang

School of Digital Media & Design Arts, Beijing University of Posts and Telecommunications, Beijing, China.

1025398230@qq.com

Abstract. Traffic flow prediction is one of the important links to realize an urban intelligent transportation system. Thanks to the in-depth research of artificial intelligence theories, the machine learning method has been widely used in intelligent transportation engineering. However, due to the "black box" as its characteristics, its application and further development are limited. Exploring the explainability of machine learning models in traffic flow prediction is an important issue to make it more reliable in traffic engineering and other practical applications. Apart from selecting the RandomForest model and the CatBoost model as the objects to research the traffic flow prediction against temporal and spatial changes, this paper makes a comprehensive evaluation and comparison with LightGBM and the other two prediction models through different indicators. Meanwhile, aiming at the low explainability of the models, their feature importance is analyzed and compared with reality. The results show that the RandomForest model and CatBoost model make good predictions, whose feature importance is consistent with the actual situation, verifying their explainability.

Keywords: Traffic Flow Prediction; Machine Learning; RandomForest; CatBoost; "Black Box" Model; Explainability.

1. Introduction

With the progress of modern social economy and culture as well as the increasing income of urban households year by year, the number of motor vehicles is on the rise. Accurate dynamic traffic flow prediction not only provides traffic managers with a scientific basis to facilitate their management, but also help drivers to check road congestion and choose more convenient travel plan at any time, thus alleviating traffic congestion. In recent years, how to carry out accurate and reasonable traffic flow prediction has become one of the research hot trends in modern transportation.

Commonly used traffic flow prediction models include the following three types, including that based on nonlinear theory, linear probability statistical analysis, and artificial intelligence. The following table will explain them one by one.

1) Traffic prediction model based on nonlinear theory. A nonlinear theoretical prediction model can find the original characteristic information of complex traffic systems in high-dimensional space by spatial reconstruction to establish a theoretical prediction model. This model has strong advantages when dealing with time series signals, but its deficiency in ignoring the traffic flow spatiality and its relatively low accuracy and real-time performance is gradually out of touch with the era development.

A traffic prediction model based on linear probability statistical analysis. This model adopts the method of mathematical statistics and analyzes the historical data of traffic flow to achieve its processing and research. Then, the future traffic trend is predicted. However, this model requires sufficient historical data and the future traffic characteristics must be completely consistent with the historical data, so as to produce better prediction results with high data requirements. As a static prediction method, this model cannot solve unconventional and sudden traffic conditions.

2) Traffic prediction model based on artificial intelligence. This model adopts the classic "black box" as its learning mode, which can automatically analyze and summarize the data rules through the existing data samples, so as to master the inherent laws of these data. Even if the user does not predict the specific internal mechanism of the problem, a stable input and output mapping model can be automatically established as long as there are reliable input and output samples as well as the internal

automatic adjustment of the "black box". It is in line with the high nonlinearity and uncertainty of traffic flow. The four machine-learning models used in this paper also belong to this category.

With the rapid development of artificial intelligence, traffic flow prediction models based on intelligent theory have been widely used. Although these machine learning methods are superior to human beings in many meaningful tasks, their performance and application are also questioned due to their lacking explainability. For ordinary users, the machine learning model itself is not "transparent" enough. If it is given an input, it will feed back a decision result, which is difficult to give the basis and principle of the prediction result. Therefore, when people have a rational guard against artificial intelligence, the credibility of the result is reduced. Lack of explainability may pose a serious threat to practical tasks or applications. In the past traffic flow prediction tasks, the explainable analysis of most models was ignored.

To sum up, although machine learning has brought great opportunities for the development of traffic flow prediction, its practical application is limited because of its low explainability. Based on two vital machine learning models, this paper forecasts the spatiotemporal traffic flow data. At the same time, the importance of features is analyzed and compared with the actual situation, which verifies the explainability of the models.

2. Model Introduction

In machine learning, the models commonly used in traffic flow detection in different times and spaces include RandomForest (RF), GBDT, Xgboost, LightGBM, CatBoost, etc. RandomForest and CatBoost are of extreme importance, which will be introduced specifically in this paper.

2.1 RandomForest

RandomForest is composed of a series of decision trees, including a classification and regression tree (CART) combined with a bagging algorithm and random subspace method (RSM).

The bagging algorithm refers to randomly extracted K sample sets {1, 2,..., n} from sample set X, with the same size as X. Each sampling set constructs a decision tree. The idea of feature subspace is to select a subset randomly from the attribute set with the current characteristics of the node when splitting each node of the decision tree, and then an optimal attribute from this subset is selected to divide it.

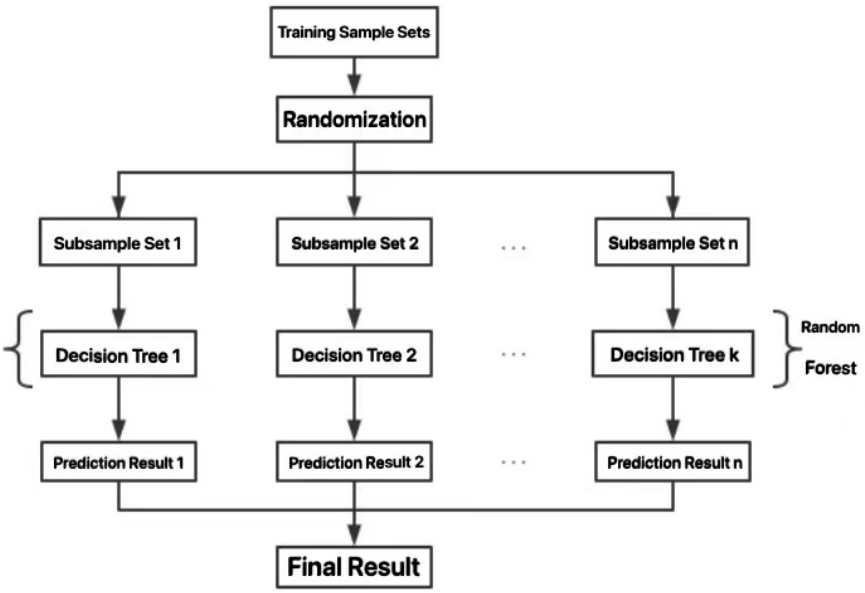

Figure 1. RandomForest Calculation Process

RandomForest can be used for regression and classification. With regard to the traffic flow prediction in this paper, a regression tree is more suitable. The training sample set is 80% of the data set to be divided in the experiment, and the sample set is randomly divided among the model. Meanwhile, the decision tree is constructed to generate the traffic prediction results, which are synthesized into the final results. Then, the trained model is used in the prediction set to obtain the final prediction result of the experiment. The regression calculation flow is shown in Figure 1.

RandomForest with high accuracy in the application can run safely and effectively on large-scale data sets. It also processes input samples with high-order characteristics without dimensionality reduction.

But its disadvantage is that the RandomForest model is difficult to get a better prediction conclusion for some relatively small data sets or some data sets with relatively few feature quantities. As for many others, the "black box" characteristic of RandomForest may greatly reduce the model reliability, because its internal operation cannot be controlled. In order to get the best results, we can only constantly try between multiple parameter trees and random seeds.

2.2 CatBoost Algorithm

2.2.1 GBDT

GBDT is a structure that combines the GradientBoosting algorithm with the decision tree structure.

Gradient lifting is the optimization of traditional one. The traditional concept of gradient lifting trees uses an additive model combined with a forward branching algorithm to realize learning. Optimization is relatively simple only when the loss function is a square error or exponential loss function. However, for the loss function without special rules, it is relatively difficult to optimize every step. In order to solve this problem, GradientBoosting is proposed. Its principle is to train new weak classifiers according to the negative gradient information of the loss function of the current model, then the trained weak classifiers are combined into the existing model in the form of accumulation.

GBDT calculation steps are as follows and the training set is $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

1) Select and initialize the weak learner;

$$f_0(x)=\arg\min_{\sigma}\sum_{i=1}^{n}L(y_i,\sigma) \tag{1}$$

2) Calculate the negative gradient of iteration t times;

$$\gamma_{it}=-\frac{\partial L(y_i,f_{t-1}(x_i))}{\partial f_{t-1}(x_i)} \tag{2}$$

t is the number of iterations, R is the leaf node area corresponding to the fitted regression tree, and J is the number of leaf nodes of the regression tree.

3) Calculate the optimal fitting value of leaf node area J;

$$\sigma_{jt}=\arg\min_{\sigma}\sum_{x_i\in R_{jt}}L(y_i,f_{t-1}(x_i)+\sigma) \tag{3}$$

4) Update strong learners;

$$f_t(x)=f_{t-1}(x)+\sum_{j=1}^{J}\sigma_{jt}I(x\in R_{jt}) \tag{4}$$

5) Repeat the above process to obtain the final strong learner.

2.2.2 CatBoost Algorithm

CatBoost is an open-source machine learning library developed by Yandex Company in April 2017. It is an improved algorithm based on symmetric decision trees under GBDT algorithm architecture. Compared with GBDT, CatBoost uses more reasonable and effective strategies to reduce over-fitting, and participates in training by using the whole data set to use data information more effectively.

1) The observed values are randomly divided into multiple random sequences.

2) In this paper, by replacing categories with the average marked values of the training data set and transforming the classified features into digital features, the direction information in the data set, that is, southeast and northwest classification features, can be transformed into digital features for the next fitting.

$$avgT = \frac{countinClass + prior}{totalCount + 1} \tag{5}$$

countinClass is the number of classification features and totalCount is the number of objects.

CatBoost has the following advantages and innovations in several common machine-learning models:

1) The use of combined category features can effectively utilize the relationship between features and enrich feature dimensions.

2) The ranking lifting algorithm is used to counter the noise points in the training set, so as to avoid the deviation of gradient estimation and solve the prediction offset.

3) Completely symmetric tree is used as the basic model, which has a regular effect and high operation efficiency.

The disadvantage is that the setting of different random numbers has a certain influence on the prediction results of the model. Thus, it is difficult for people to intuitively find the random numbers that make the best effect.

3. Experiment

3.1 Data Description

The data in this paper come from the traffic flow data of different periods and places from April 1991 to September 1991. The time interval of data collection is 20min, with 848,834 traffic flow data in total. The correlation thermodynamic diagram between specific information and characteristics is shown in Figures 1 and 2.

Table 1. Data Set Information1

Parameter	Value
Time Span	1991-04-01——1991-09-30
Time Interval（min）	20
East-West Midpoint Coordinates of the Road	0-2
South-North Midpoint Coordinates of the Road	0-3
Direction of the Road	8 Directions （Northwest, East, etc.）

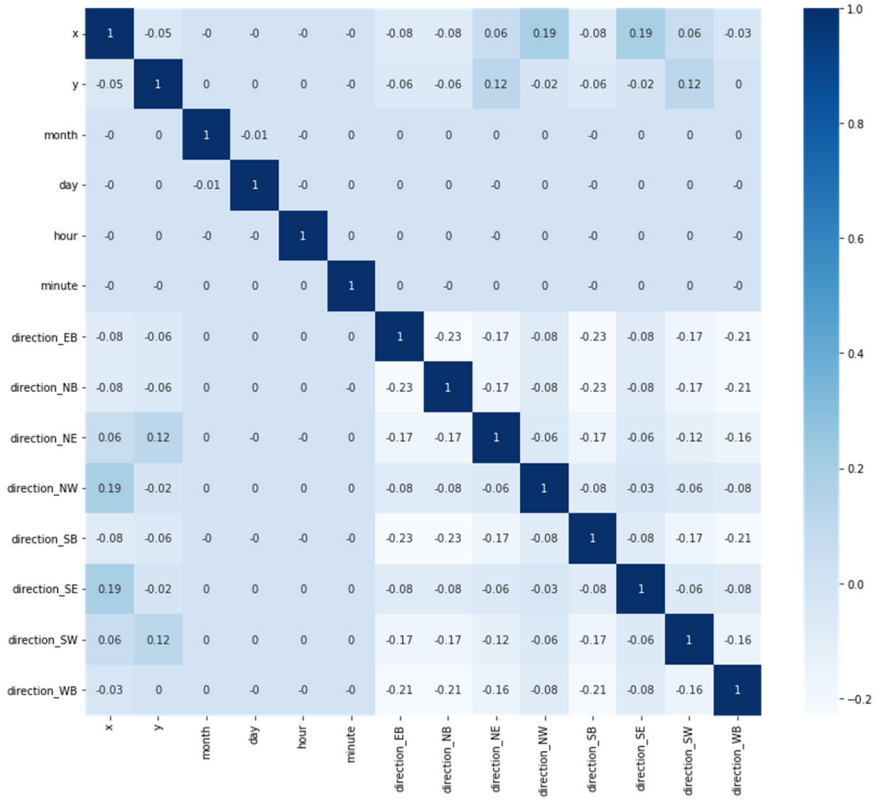

Figure 2. Correlation Thermodynamic Diagram Between Various Characteristics1

Because of the abundant data and the low correlation between each feature, it can be used as a data set for traffic flow prediction.

3.2 Evaluation Index

In order to evaluate the overall performance of the prediction model comprehensively and systematically, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R2 Determination Coefficient (R-Square) are selected to evaluate the accuracy of the prediction model. All of them can well reflect the actual prediction error. The specific calculation formula is as follows.

$$MAE = \frac{1}{\mathrm{m}} \sum_{i=1}^{m} \left| \left(y_i - \hat{y}_i \right) \right| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{m} \left(y_i - \hat{y}_i \right)^2} \tag{7}$$

$$R2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \tag{8}$$

where m is the total number of samples, y_i and $\hat{y}_i$ represent the true value and predicted value of the i-th sample respectively. The smaller MAE and RMSE are, the closer R2 is to 1, and the better the fitting effect is.

3.3 Experimental Content

In order to better evaluate the prediction results of CatBoost, this paper constructs four different models for this data set, which are (1) the XGBoost model, (2) the LightGBM model, (3) the RandomForest model, and (4) the CatBoost model.

Taking 80% of the data as a training set and 20% as a test set, in order to avoid the influence of super-parameters and get a better prediction effect, this experiment adopts the Bayesian optimization method to optimize the model parameters and uses the best parameters of each model to predict.

3.4 Comparative Analysis of Prediction Results

The CatBoost model is compared with the other four models, with the experimental results shown in Figure 3. Figure 4-7 is a comparison of the predicted values and actual values of XGBoost, LightBoost, RF, and CatBoost models (red is the actual value and blue is the predicted value).

Table 2. Evaluation of Prediction Results of Different Models2

model	MAE	RMSE	R2
XGBoost	6.89028	9.70474	0.66752
LightGBM	6.47095	9.32782	0.69285
RandomForest	6.30214	8.99876	0.71445
CatBoost	6.16316	8.83621	0.71811

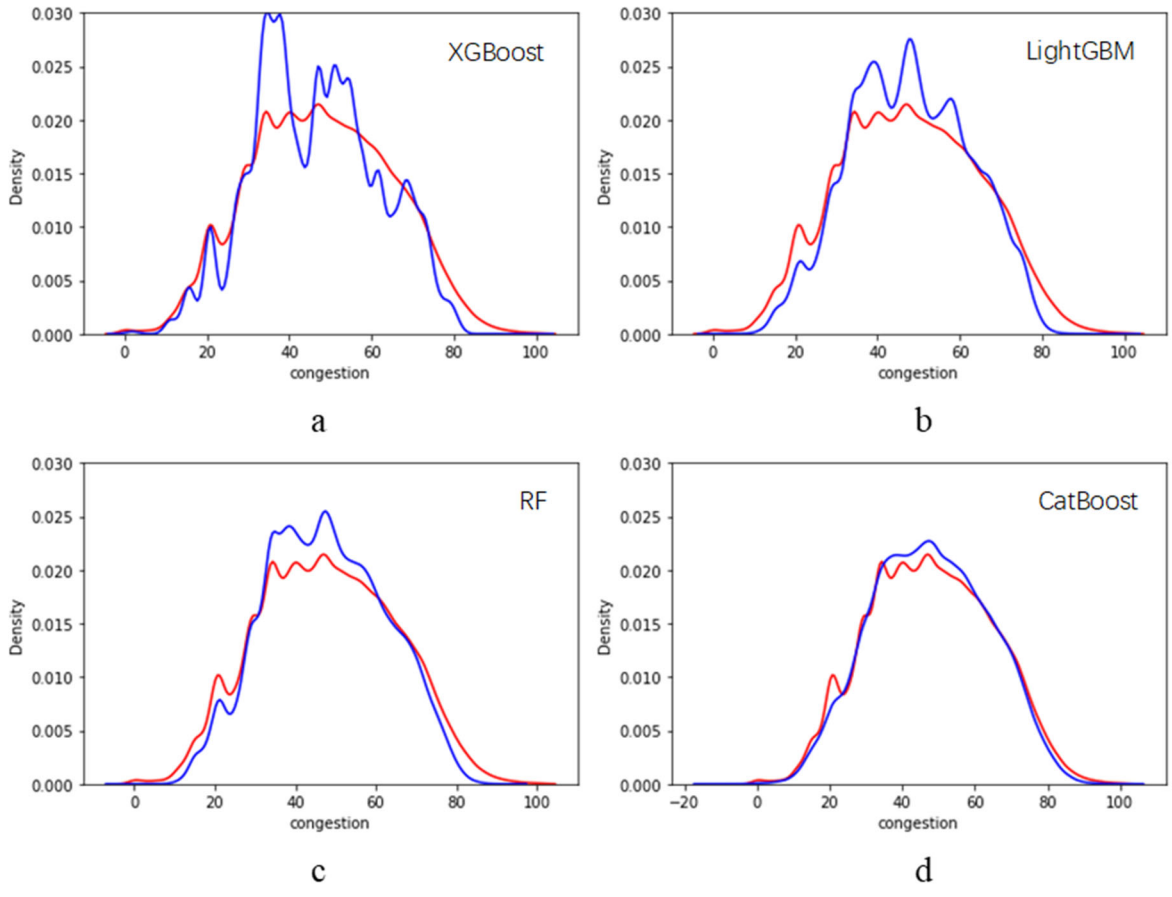

Figure 3. Comparison of Predicted Values and Actual Values of Four Models2

From the performance of prediction results, the error between the prediction results of the XGBoost model and the original traffic flow is relatively large. Although the prediction trend of LightGBM is consistent with the real one, there are large fluctuations and errors at the peak and inflection points. RandomForest has fewer errors than the first two. As can be seen from the values and pictures, CatBoost is better than others.

Therefore, CatBoost has a better effect when using multivariate traffic data sets with temporal and spatial characteristics. The main reason is that Catboost uses combined category features, which can make use of the relationship between features and their respective advantages, greatly enriching feature dimensions and improving the accuracy of the feature allocation ratio. Meanwhile, the model adopts the method of ranking promotion to fight against the noise points in the training set, which avoids the deviation of gradient estimation and solves the prediction offset. Obviously, it is more suitable for solving the problems of traffic flow prediction under temporal and spatial changes.

3.5 Feature Importance Analysis

Feature importance can reflect the mapping relationship between the predicted value of the model and the input features of the model, indicating the importance of different features in prediction. Taking two models with relatively small fitting as examples, with the feature importance of RandomForest and CatBoost models shown in Figure 8, the weighted feature importance diagram of the two models is shown in Figure 9.

Table 3. Feature Importance of RandomForest and CatBoost3

Features	CatBoost	RandomForest
x	16.180	13.646
y	22.315	21.523
month	7.168	5.105
day	9.337	9.632
hour	13.269	12.025
minute	3.235	2.437
direction_EB	3.400	3.292
direction_NB	4.998	6.103
direction_NE	4.866	3.774
direction_NW	3.319	5.698
direction_SB	6.408	7.639
direction_SE	1.857	3.548
direction_SW	1.789	4.108
direction_WB	1.858	1.470

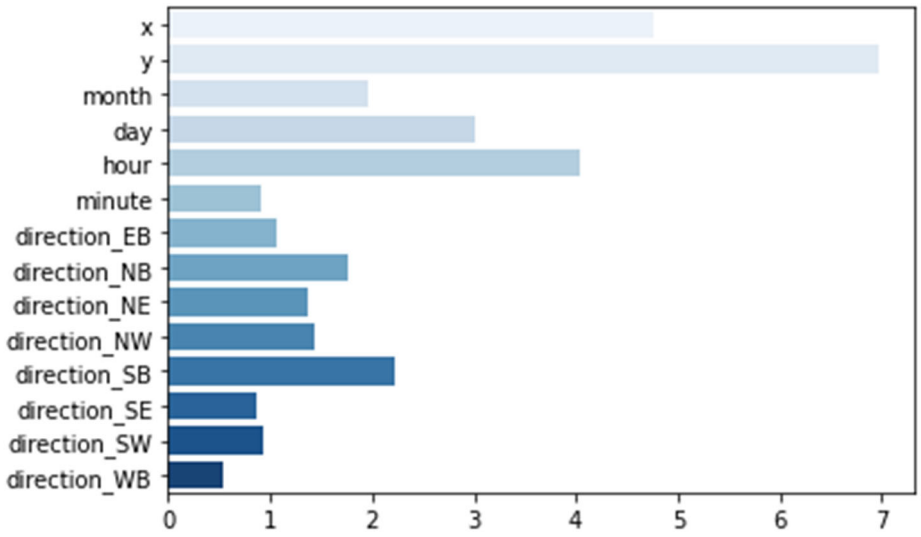

Figure 4. Weighted Feature Importance3

As can be seen from the figure:

1) All the features have a certain impact on the prediction model, which shows that the model makes full use of all conditions in the data set to improve the prediction accuracy.

2) Among the influences of various characteristics on the prediction model, spatial location of traffic flow, that is, (x, y) coordinates have the greatest influence, followed by the time series characteristics in hours (minutes have relatively little influence on the prediction results), while the road orientation at the same position has the least impact. This fully shows the importance of time characteristics and spatial correlation of time series to the prediction target accuracy in the process of traffic flow prediction.

3) Some characteristic information, such as minutes and some road directions, have little influence on the prediction results of the model, which has certain enlightenment for future traffic data collection and traffic flow prediction experiments. When collecting data, it can reduce the data collection in this aspect and reduce the hardware cost. Meanwhile, the model parameters can be adjusted directionally, so that the model can focus on more important features, effectively lighten the model, and improve the fitting efficiency of the model.

4. Conclusion

With the gradual integration of artificial intelligence technology into our daily life, it is inevitable to solve many practical problems through machine learning. At the same time, explainability will be an unavoidable and important issue of artificial intelligence in the future. It is also a crucial research direction. In addition to introducing the concept and basic algorithm flow of the RandomForest model and CatBoost model in machine learning, this paper constructs these two models, explores their prediction of the traffic flow by machine learning, compares them with the other two models, and verifies them by analyzing the feature importance for explainability. The results show that the MAPE, MAE, and R2 indexes of the RandomForest model and CatBoost model are superior to others in predicting the traffic flow data transformed simultaneously in time and space, that is, the prediction accuracy is the highest. At the same time, the spatial position of traffic flow, i.e. (x, y) coordinates, has the greatest influence on the feature importance, followed by the time series features in hours, which is consistent with the results recognized by people. The experimental results show that the RandomForest model and CatBoost model can be used to simulate the dynamic changes of traffic flow at the same time and region accurately, which is suitable for the prediction of urban traffic flow with certain explainability.

References

[1] Yin, X., et al. A Comprehensive Survey on Traffic Prediction. (2020).

[2] Sun, S. & Xin, X. Variational Inference for Infinite Mixtures of Gaussian Processes With Applications to Traffic Flow Prediction. IEEE Transactions on Intelligent Transportation Systems. 12.2(2011): 466-475.

[3] Wu, Y., et al. A Hybrid Deep Learning Based Traffic Flow Prediction Method and its Understanding. Transportation Research. 90.5(2018): 166-180.

[4] Hongyuan, et al. Sensus Ensemble System for Traffic Flow Prediction. IEEE Transactions on Intelligent Transportation Systems. 19.12 (2018): 3903-3914.

[5] Wang, L., et al. Cross-City Transfer Learning for Deep Spatio-Temporal Prediction. (2018).

[6] Liaw, A. & Wiener, M. Classification and Regression by RandomForest. R News. 23.23 (2002).

[7] Prokhorenkova, L., et al. CatBoost: Unbiased Boosting with Classified Features. (2017).

[8] Ma, Z., Luo, G. & Huang, D. Short-Term Traffic Flow Prediction Based on On-line Sequential Extreme Learning Machine. 2016 Eight International Conference on Advanced Computational Intelligence (ICACI) IEEE. (2016).

[9] Shafiq, M., Yu, X. & Laghari, A. A. WeChat Text Messages Service Flow Traffic Classification Using Machine Learning Technique. 2016 6th International Conference on IT Convergence and Security (ICITCS) IEEE Computer Society. (2016).

[10] Gunning, D. & Aha, D. W. DARPA's Explainable Artistic Intelligence (XAI) Program. Ai Magazine 40.2 (2019): 44-58.

[11] Stoica, I., Song, D., et al. A Berkeley View of Systems Challenges for AI [R]. Berkeley: Electronic Engineering and Computer Sciences, UC Berkeley. (2017):1-13.

[12] Arrieta, A. B., et al. Explainable Artistic Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion. (2019).

[13] Ribeiro, M. T., Singh, S. & Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. ACM. (2016).