Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 4202735, 11 pages
https://doi.org/10.1155/2019/4202735

*Research Article*

# A Hybrid Method for Traffic Incident Duration Prediction Using BOA-Optimized Random Forest Combined with Neighborhood Components Analysis

**Qiang Shang**,[1] **Derong Tan**,[1] **Song Gao**,[1] **and Linlin Feng**[2]

[1]*School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo, Shandong 255049, China*
[2]*School of Marxism Studies, Shandong University of Technology, Zibo, Shandong 255049, China*

Correspondence should be addressed to Qiang Shang; shangqiangv587@163.com

Received 8 September 2018; Revised 18 December 2018; Accepted 1 January 2019; Published 20 January 2019

Guest Editor: Hamzeh Khazaei



Predicting traffic incident duration is important for effective and real-time traffic incident management (TIM), which helps to minimize traffic congestion, environmental pollution, and secondary incident related to this incident. Traffic incident duration prediction methods often use more input variables to obtain better prediction results. However, the problems that available variables are limited at the beginning of an incident and how to select significant variables are ignored to some extent. In this paper, a novel prediction method named NCA-BOA-RF is proposed using the Neighborhood Components Analysis (NCA) and the Bayesian Optimization Algorithm (BOA)-optimized Random Forest (RF) model. Firstly, the NCA is applied to select feature variables for traffic incident duration. Then, RF model is trained based on the training set constructed using feature variables, and the BOA is employed to optimize the RF parameters. Finally, confusion matrix is introduced to measure the optimized RF model performance and compare with other methods. In addition, the performance is also tested in the absence of some feature variables. The results demonstrate that the proposed method not only has high accuracy, but also exhibits excellent reliability and robustness.

## 1. Introduction

Traffic incidents such as vehicle crashes, fire, road maintenance, debris, police activities, etc. are still very common, random, and dangerous. The occurrence of traffic incidents can reduce road capacity because of lane closures that result in traffic congestion and delays [1]. The National Traffic Accident Management Association estimates that 25% of the congestion on American roads is caused by traffic incidents [2]. Traffic incidents are the main causes of nonrecurrent congestion on urban expressways and urban arterial roads [3]. In addition, traffic congestion and travel delay can increase the occurrence likelihood of secondary incident [4, 5]. For more than two decades, many cities around the world have established traffic management centers and deployed various traffic incident management systems to decrease traffic incidents and alleviate related congestion. Traffic flow management and providing travelers with timely and accurate information during traffic incident clearance periods are two main aspects of efficient traffic incident response [6]. The response strategy during the clearance period depends to a large extent on the duration of an incident. Therefore, accurate prediction of traffic incident duration has attracted the attention of researchers because of its importance.

For decades, researchers have put forward many effective methods for predicting traffic incident duration. The data, variables, and algorithms used in these methods are usually different. From the algorithm point of view, early incident duration prediction methods include probability distribution model [7], the regression prediction method [8], and time series methods [9]. They have an advantage of being easily understood and well-established methodologies, with a long history of application, availability of software, and deep-rooted acceptance. However, these methods often depend on certain assumptions, which limit the generalization of them.

In recent years, the application of hazard/risk-based methods and machine learning methods for traffic incident duration prediction has become increasingly widespread. The

hazard/risk-based method uses hazard function to predict traffic incident duration. The risk function is essentially a conditional probability function, which can be used to analyze the probability that a traffic incident has lasted t minute and ended in the kth minute. In 2013, parametric accelerated failure time (AFT) survival models of traffic incident duration were developed by Hojati et al. [10], including log-logistic, lognormal, and Weibull—considering both fixed and random parameters, in order to better apply to different types of traffic incidents. In 2015, Li et al. [3] developed a competing risks mixture model to investigate the influence of clearance methods and various factors on traffic incident duration and predict traffic incident duration. Three candidate distributions including generalized gamma, Weibull, and log-logistic are tested to determine the most appropriate probability density function of the parametric survival analysis model. Subsequently, Li et al. [11] proposed a sequential prediction method for traffic incident duration based on competing risk mixture model and text analysis. Text analysis was used to process the textual features of the traffic incident to extract time-dependent topics. The empirical results demonstrate that the developed mixture model outperforms the non-mixture model. In 2017, accelerated failure time (AFT) hazard-based models were developed with different underlying probability distributions of the hazard function to predict traffic incident duration [12]. This study indicates that the hazard function—gamma distribution model with a time variable is the best model for the four different duration stages (including preparation, travel, and clearance as well as the total duration of the incident), and different parameters and variables were appropriate for modeling the different duration stages of traffic incidents.

As a typical machine learning method, decision trees have drawn extensive attention due to its excellent performance. Many methods have been proposed for traffic incident duration prediction based on decision trees, such as Bayesian decision trees [13], classification tree [14], M5P tree algorithm [15], and Gradient Boosting decision trees [16].

Besides decision tree-based methods, other machine learning methods for traffic incident duration have also made a number of achievements. In 2014, a model for traffic incident duration prediction was developed using an adaptive Bayesian network, which is more adaptable to the future environment than the traditional Bayesian model [17]. In 2015, Wang et al. [18] proposed a prediction method based on a nonparametric regression model whose core algorithm is K Nearest Neighbor (KNN). In the process of modeling, the distribution of incidents duration is taken into account. It is pointed out that the logarithmic transformation of duration will achieve better prediction results. In 2016, Yu et al. [19] compared the performance of artificial neural network (ANN) and support vector machine (SVM) for predicting traffic incident duration. The results show that both ANN and SVM are able to predict the incident duration within an acceptable range. When predicting the longer duration, the ANN model has better performance. On the whole, the overall performance of the SVM model is better than that of the traditional ANN model. In the same year, Park et al. [20] proposed a continuous time prediction method based on Bayesian neural network and quantified the importance of continuous time input variables by connecting weights to improve the interpretability of the algorithm. In 2017, two nonparametric machine learning methods, including the k-nearest neighbor method and artificial neural network method, were used to develop incident duration prediction models [21]. Based on the performance comparison results, an artificial neural network model can provide good and reasonable prediction for traffic incident duration prediction with mean absolute percentage error values less than 30%, which are better than the prediction results of a k-nearest neighbor model. For more information, a review for traffic incident duration prediction can be referred to [22], mainly including the different phases of incident duration, data resources, and the various methods that are applied in the traffic incident duration influence factor analysis and duration time prediction.

In summary, both hazard/risk-based methods and machine learning methods have advantages and disadvantages. The forms of results obtained by hazard/risk-based methods often better meet the needs of traffic managers, but it is necessary to assume that traffic incident duration obeys one or more probability distributions. The machine learning methods are not limited by the hypothesis conditions, which can directly extract rules from the data set for traffic incident duration prediction with better applicability. However, the interpretability of such methods is usually poor. With the expansion of data scale and the improvement of computer performance, machine learning methods have a broader prospect for traffic incident prediction. At present, most methods use as many relevant variables as possible in order to achieve better prediction results. However, an important fact is ignored that is lack of relevant information at the beginning of the traffic incident. Therefore, it is very necessary to select relevant variables and explore more robust methods for traffic incident duration prediction using incomplete variables.

To tackle the shortcoming as mentioned above, a novel method named NCA-BOA-RF is proposed for traffic incident duration prediction. As a well-known machine learning algorithm, Random Forest (RF) is chosen as the basic algorithm because of its excellent performance of regression and classification. Firstly, based on the analysis of the influencing factors of traffic incident duration and considering the characteristics of the data sets used, 18 influencing factors were selected as the relevant variables of traffic incident duration. Then, feature weights of the relevant variables are calculated by Neighborhood Components Analysis (NCA), and feature variables of traffic incident duration are determined by the feature weights. Finally, the training set is constructed using the feature variables for training RF, and the Bayesian Optimization Algorithm (BOA) is used to optimize the parameters of RF.

The main contributions of this paper are highlighted in the following aspects. (a) NCA is used to calculate feature weights of relevant variables to determine the feature variables. (b) The cross-validation method is used to optimize the regularization parameter of the NCA to ensure its best performance. (c) BOA is used to optimize both parameters of RF at the same time, rather than a single parameter.

(d) Consider that some feature variables are unavailable when testing performance of the proposed method.

The remainder of this paper is organized as follows. Section 2 elaborates the methodology of the proposed model. Section 3 presents the experimental result and discussion. Finally, the conclusions and future study are summarized in Section 4.

## 2. Methodology

In this section, the basic methods relevant to the proposed model named NCA-BOA-RF are briefly introduced, including Neighborhood Components Analysis (NCA), Bayesian Optimization Algorithm (BOA), and Random Forest (RF). Then, the main steps of NCA-BOA-RF model are given. In the NCA-BOA-RF model, RF is the basic method for traffic incident duration prediction, feature variables were extracted from relevant variables of traffic incident duration, and BOA are utilized to optimize two parameters of RF at the same time.

*2.1. Neighborhood Component Analysis.* Feature selection is considerably important in data mining and machine learning, especially for high dimensional data. Proper feature selection not only reduces the dimensions of features, but also improves algorithms generalization performance and execution speed [23, 24]. For traffic incident duration prediction, there are many factors that are considered to be related to the traffic incident duration. Usually, these factors are transformed into relevant variables as the prediction model input, such as incident type, number of casualties, number of closed lanes, and weather conditions. However, the more variables are used as input to the model, the better prediction results are not necessarily obtained. Moreover, in the initial stage of incident occurrence, the available data are often very limited (one or more variables may be missing), which reduces the performance of the prediction model. Therefore, it is very important and necessary to select feature variables for traffic incident duration prediction. In this study, Neighborhood Component Analysis (NCA), as a feature selection method, is used to select feature variables of traffic incident duration. NCA is an algorithm that learns a Mahalanobis distance metric in the supervised k-nearest neighbor (KNN) algorithm by minimizing the leave-one-out (LOO) classification error on the training data [25]. In 2012, NCA-based feature selection is proposed, which learns a feature weighting vector by maximizing the expected LOO classification accuracy with a regularization term [26]. Principal Component Analysis (PCA) and Sequential Feature Selection (SFS) are classic and popular methods for feature selection. However, PCA may result in loss of information when mapped to lower dimensions, and SFS may not be able to remove features that become useless after adding other features [27]. In contrast, NCA is not subject to data conditions (e.g., dimension and distribution) and has the advantage that no information will be lost during the dimension reduction process.

Let $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_N, y_N)\}$ be a labelled training data set, where $\mathbf{x}_i \in \mathbf{R}^d$ are feature vector, $y_i \in \{1, \ldots, C\}$ are the corresponding labels, and $N$ is the number of observations.

A Mahalanobis distance between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ that are denoted in terms of weighing vector $\mathbf{w}$ is defined as follows [26]:

$$D\mathbf{w}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{l=1}^{d} w_l^2 \left|x_{il} - x_{jl}\right| \tag{1}$$

where $w_l$ is the $l$th feature weight. LOO is considered to maximize classification accuracy on training data set $T$. Randomly select a reference point from $T$ to be labelled accordingly. Given a data point $\mathbf{x}_i$, the probability of drawing $\mathbf{x}_j$ as a reference point of $\mathbf{x}_i$ is defined by [26]

$$p_{ij} = \begin{cases} \dfrac{k\left(D\mathbf{w}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)}{\sum_{k \neq i} k\left(D\mathbf{w}\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)}, & i \neq j \\ 0, & i = j \end{cases} \tag{2}$$

where $k(z) = \exp(-z/\sigma)$ is a kernel function and $\sigma$ is kernel width that influences the probability of each point being selected as the reference point. The average probability of LOO correct classification is the probability $p_i$ that the randomized classifier correctly classifies observation $i$ which can be expressed as

$$p_i = \sum_j y_{ij} p_{ij} \tag{3}$$

where

$$y_{ij} = \begin{cases} 1, & \text{if } y_i = y_j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Then, the approximate LOO classification accuracy can be calculated as follows [26].

$$\xi(\mathbf{w}) = \sum_i p_i = \sum_i \sum_j y_{ij} p_{ij} \tag{5}$$

The goal of NCA is to maximize objective function $\xi(\mathbf{w})$ associated with $\mathbf{w}$ using regularized term $\lambda$.

$$\xi(\mathbf{w}) = \sum_i \sum_j y_{ij} p_{ij} - \lambda \sum_{l=1}^{d} w_l^2 \tag{6}$$

where $\lambda > 0$ is a regularized term that can be optimized using cross-validation, which balances the first term of maximization of NCA probability and the second term of minimization of the Frobenius norm.

Because objective function $\xi(\mathbf{w})$ is differentiable, its derivative with respect to $\mathbf{w}$ can be expressed as

$$\frac{\partial \xi(\mathbf{w})}{\partial w_l} = \sum_i \sum_j y_{ij} \left[ \frac{2}{\sigma} p_{ij} \left( \sum_{k \neq i} p_{ik} \left|x_{il} - x_{kl}\right| \right) - \left|x_{il} - x_{kl}\right| w_l - 2\lambda w_l \right] = \frac{2}{\sigma}$$

$$\cdot \sum_i \left( p_{ij} \sum_{k \neq i} p_{ik} \left| x_{il} - x_{kl} \right| - \sum_j y_{ij} p_{ij} \left| x_{il} - x_{jl} \right| \right) w_l - 2\lambda w_l = 2 \left( \frac{1}{\sigma} \cdot \sum_i \left( p_{ij} \sum_{k \neq i} p_{ik} \left| x_{il} - x_{kl} \right| - \sum_j y_{ij} p_{ij} \left| x_{il} - x_{jl} \right| \right) - \lambda \right) w_l \tag{7}$$

According to the above, the corresponding gradient ascent update equation can be obtained. The problem of maximizing the objective function can be solved via the gradient descent method. More details about NCA for feature selection can be found in [26].

*2.2. Bayesian Optimization Algorithm.* Bayesian Optimization Algorithm (BOA) is one of the most well-known distribution algorithm estimates that combine Bayesian networks with evolutionary algorithms. In the BOA, global statistical information is extracted from optimal solutions searched currently and modeled using Bayesian networks. Therefore, BOA can overcome the disruption of building blocks in genetic algorithms. The BOA has advantages in the optimization of machine learning algorithm hyperparameters, because of its faster search speed and fewer iteration compared to traditional search algorithms [28–30]. In this study, the BOA is employed to optimize the parameters of Random Forest (which is the basic model, see Section 2.3 for details) for traffic incident duration prediction, in order to achieve better prediction results.

The algorithm parameters to be optimized are denoted as $\boldsymbol{\lambda} = \{\lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2, \ldots, \lambda_m \in \Lambda_m\}$. $T_{train}$ is training set and $T_{valid}$ is validation set, and $L(\boldsymbol{\lambda}, T_{train}, T_{valid})$ is the validation accuracy. The goal of optimization is to find a set of parameter values that can maximize $L$.

Firstly, initialization parameters of the machine learning algorithm are $\boldsymbol{\lambda}1, \ldots, \boldsymbol{\lambda}n$. Secondly, evaluate the accuracy of the machine learning algorithm with initial parameters using the validation set and record the accuracy. Thirdly, the Gaussian Process (GP) model $V$ is introduced to fit the recorded accuracy iteratively. Then, update the machine learning algorithm parameters according to the recommendations of the GP model. In this process, select the next operating point by the maximum of the acquisition function. The acquisition function guides the optimization by determining the next point to evaluate. Several acquisition functions have been proposed, such as probability of improvement, expected improvement, and information gain. Here, expected improvement is used as the acquisition function [30], and the best validation accuracy so far is $L^*$.

$$\alpha(\boldsymbol{\lambda}, V) = \int_{-\infty}^{\infty} \max\left(L - L^*, 0\right) p_V\left(L \mid \boldsymbol{\lambda}\right) d_L \tag{8}$$

where $p_V(L|\boldsymbol{\lambda})$ is the probability of $L$ with $\boldsymbol{\lambda}$, which is encoded by the GP model $V$. For more details about acquisition functions, see [31]. More details about BOA for hyperparameters of machine learning can be referred to [28].

*2.3. Random Forest.* Random Forest (RF) [32] is a machine learning algorithm that integrates multiple decision trees by ensemble learning methods. More specifically, RF is an ensemble learning method that can be used for both classification and regression. Ensemble learning is different from traditional statistic reasoning and machine learning that are trying to build an accurate single model. The goal of ensemble learning is to aggregate the results from multiple trained "weak learners" in order to obtain a "strong learner". In general, the "weak learners" are simple and fast models with a poor performance. In the case of RF, "weak learners" are decision trees. Therefore, RF has good resistance to noise and not easy to fall into overfitting. Moreover, there are not any assumptions in the resulting RF model. As a result, it is expected that the RF model has wider applicability and better robustness compared with traditional statistic reasoning and machine learning techniques [33, 34]. Therefore, RF is used as a basic model for traffic incident duration prediction. The feature variables selected by the NCA are used as input to the model, and traffic incident duration is used as the output of the model. For more information on model training, see Section 3.

The detailed steps of RF are shown as follows:

(1) Bootstrap samples are randomly formed from the original data with replacement, which will be the training set for growing the trees. The number of the created samples is equal to the number of the trees. Around one-third of the original data are left out which are called "out-of-bag" (OOB) data, while the remaining data are called in-bag data. RF performs a cross-validation in parallel with the training step by using the OOB samples to measure the prediction error [35].

(2) Each node of decision tree selects $m_{try}$ ($m_{try} < P$) features from the $P$ features as a subset rather than comparing all the input variables (features) and the best split is calculated only within this subset. It is worth noting that $m_{try}$ may affect the stability of the RF model. The sensitivity of other parameters such as the number of trees (denoted as $n_{tree}$) in RF, as well as the size of each tree (i.e., the number of splits in each tree called number of leaves per tree, denoted as $n_{leaf}$), has also been studied [36, 37].

(3) Each tree splits to its maximum size without pruning throughout the growth of the forest [32].

(4) Each decision tree gives a prediction result. For regression problems, the average results of all decision trees are calculated to find the final prediction value; and for classification problems, the majority voting result is taken as the final output prediction value.

*2.4. NCA-BOA-RF Method.* The flowchart of NCA-BOA-RF method for traffic incident duration prediction is shown in Figure 1. As can be seen from Figure 1, the application of the NCA-BOA-RF method includes two stages, and the main steps of the NCA-BOA-RF method are as follows.

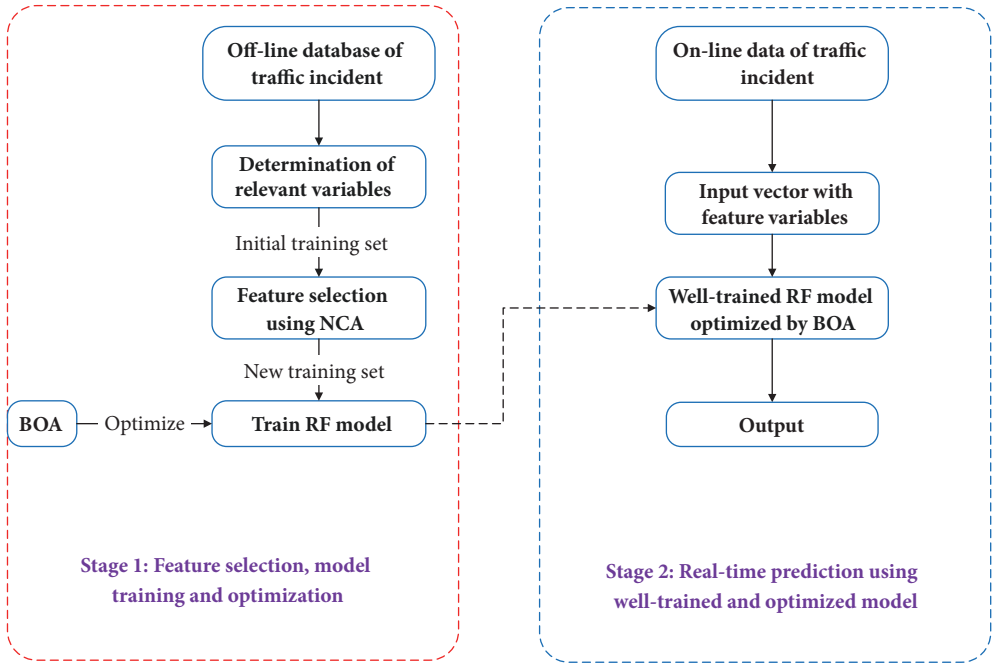


Figure 1: The flowchart of NCA-BOA-RF method.

*Step 1.* Determine relevant variables of traffic incident duration according to the available traffic incident dataset.

*Step 2.* The NCA method is used to select feature variables from the relevant variables for traffic incident duration prediction.

*Step 3.* Construct a training set using the feature variables.

*Step 4.* The RF model is used to learn the training set and the RF parameters are optimized by BOA.

*Step 5.* Input real-time collected feature variables data into the well-trained RF model and then the model will output prediction result of the traffic incident duration.

It is worth noting that more useful information may be obtained during the duration of a traffic incident. Therefore, if new useful information is collected and the traffic incident has not ended, the NCA-BOA-RF method can be used to predict the traffic incident duration again, but it is necessary to consider the time the incident has passed. In addition, with the increase of traffic incident data, offline database should be continuously updated and used to retrain and optimize the RF model.

## 3. Empirical Analysis

*3.1. Data Description.* Data used in this work were obtained from traffic incident dataset on Interstate 880 (well known as I-880 dataset), which were collected in the Freeway Service Patrol (FSP) project. In the two periods of the FSP project, the number of recorded traffic incidents was1210 and 971, respectively. Some incidents whose start or end time was not within the observation period cannot be used for modeling and testing due to their unknown duration. In addition, some traffic incidents were planned and predictable, such as road maintenance and traffic restrictions, which were excluded in this study.

The I-880 dataset not only records the time, type, location, and duration of the incident, but also the relative location and distance between the incident location and the road exit, the number of vehicles involved, the type and color of involved vehicles, the location and number of lanes affected, the weather during the incident, and casualties. In addition, whether or not rescue vehicles such as trailers, ambulances, and fire engines are required, as well as the arrival and departure times of rescue vehicles are also recorded. In summary, the basic information of all available traffic incidents is given in Table 1.

A total of 440 traffic incidents (235 in the before period and 205 in the after period) were used in this study. 308 traffic incidents (70%) data were randomly selected as the training set, and the remaining 132 traffic incidents data were used as test set. According to the actual situation of I-880 dataset, the relevant variables of traffic incident duration selected in this study are given in Table 2. The first 18 variables are the input, and the 19th variable is output.

*3.2. Feature Selection Using NCA.* The NCA algorithm is used to select feature variables for traffic incident duration. Regularization parameter $\lambda$ is an important parameter of NCA. First of all, it is necessary to determine whether the value of $\lambda$ is reasonable. In general, the value of $\lambda$ is $1/n$, where $n$ is the number of input variables of training set ($n = 18$ in this study). In the process of selecting feature variables using NCA, it is necessary to add a set of irrelevant variables as a

Table 1: The basic information of all available traffic incidents.

| Incident type/Location | Before period | | After period | |
|---|---|---|---|---|
| | Number | Mean (min) | Number | mean (min) |
| Accidents | 45 | 27.3 | 38 | 28.9 |
| Breakdowns | 184 | 25.3 | 166 | 23.6 |
| Debris/Pedestrian | 6 | 10.8 | 1 | 32.0 |
| Lane | 21 | 24.7 | 16 | 30.6 |
| Right Shoulder | 195 | 25.0 | 169 | 23.8 |
| Central Divide | 19 | 28.4 | 20 | 26.9 |
| All | 235 | 25.6 | 205 | 24.7 |

Table 2: The relevant variables of traffic incident duration selected in this study.

| No. | Variables | Descriptions |
|---|---|---|
| 1 | Incident type | 1: Accidents, 2: Breakdowns, 3: Debris/Pedestrian |
| 2 | Incident time | 1: A.M., 2: P.M. |
| 3 | Direction | 1: Northbound, 2: Southbound |
| 4 | Incident location | 1: Lane, 2: Right Shoulder, 3: Central Divide |
| 5 | Number of lanes closed | 1: No; 2: One, 3: More than one |
| 6 | Distance from the exit | 1: < 0.5 miles, 2: 0.5~1 miles, 3: >1 miles |
| 7 | Relative position to exit | 1: At the exit, 2: Front of the exit, 3: Behind the exit |
| 8 | Detection type | 1: Call report, 2: Operator detected |
| 9 | Need Police | 1: Yes, 2: No |
| 10 | Need Ambulance | 1: Yes, 2: No |
| 11 | Need Truck Wrecker | 1: Yes, 2: No |
| 12 | Need firefighters | 1: Yes, 2: No |
| 13 | Weather information | 1: Sunny, 2: Cloudy, 3: Light Rain, 4: Heavy Rain |
| 14 | Automobile Count | 1: None; 2: One, 3: Two; 4: More than two |
| 15 | Motorcycle Count | 1: None; 2: One, 3: More than one |
| 16 | Heavy Truck Count | 1: None; 2: One, 3: More than one |
| 17 | Light Truck Count | 1: None; 2: One, 3: More than one |
| 18 | Tractor Trailer Count | 1: None; 2: One, 3: More than one |
| 19 | Duration | 1: <10min, 2: 10~30min, 3: 30~60min, 4: 60~90min; 5: >90min |

control to highlight the importance of the feature variables. In this study, 100 irrelevant variables are randomly generated from a Normal distribution with a mean of 0 and a variance of 20. In the process of selecting feature variables using NCA, it is necessary to add a set of irrelevant variables as a control to highlight the importance of the feature variables. In this study, 100 irrelevant variables are randomly generated from a Normal distribution with a mean of 0 and a variance of 20. The generated 100 irrelevant variables are added to 18 relevant variables (which are considered as the variables relevant to traffic incident duration). The feature weights of all variables are calculated by NCA. If the value of $\lambda$ is appropriate, the feature weight of the relevant variables is large, and the feature weight of the irrelevant variables is small and close to 0. If the value of $\lambda$ is too large, the feature weights of all variables are close to 0; if the value of $\lambda$ is too small, the irrelevant variables also have a large feature weight.

According to the recommendation in [28, 37], the cross-validation method is used to optimize the regularization parameter $\lambda$. In this study, 5-fold cross-validation is employed. That is, the training set is randomly divided into 5 subsets, 4 subsets reconstruct a training set, and the remaining 1 subset is used as the test set, then the process is repeated five times until each subset is used as a test set, and the average value of 5 test results was adopted as the final result. The best $\lambda$ value produces the minimum classification loss. Figure 2 shows the average loss values versus $\lambda$ values, and the best $\lambda = 0.0021$ was obtained that corresponds to the minimum average loss of 0.1250.

The feature weights of all variables are calculated using NCA with the best $\lambda$ value, and the results are shown in Figure 3. As can be seen from Figure 3(a), the feature weights of the irrelevant variables are all close to 0. As can be seen from Figure 3(b), the serial numbers of the variables with a large feature weight are no. 1, no. 5, no. 11, no. 12, no. 14, and no. 16, respectively. And their corresponding variable names are "incident type", "number of lanes closed", "need truck wrecker", "need firefighters", "automobile count", and "heavy truck count".

Under normal circumstances, traffic incidents involve more vehicles, larger vehicles, and more departments involved in rescue. The duration of the incidents tends to

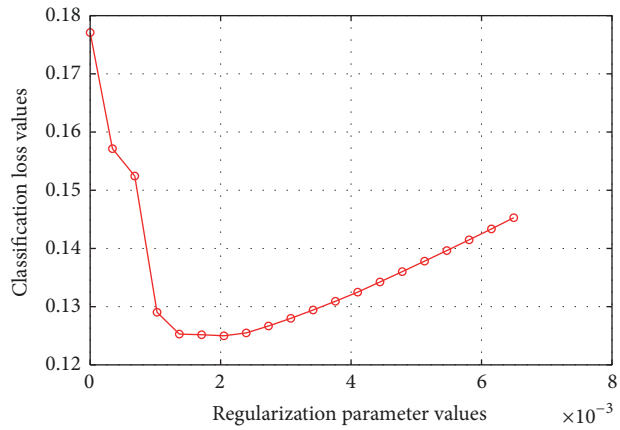


FIGURE 2: The classification loss values versus regularization parameter values.

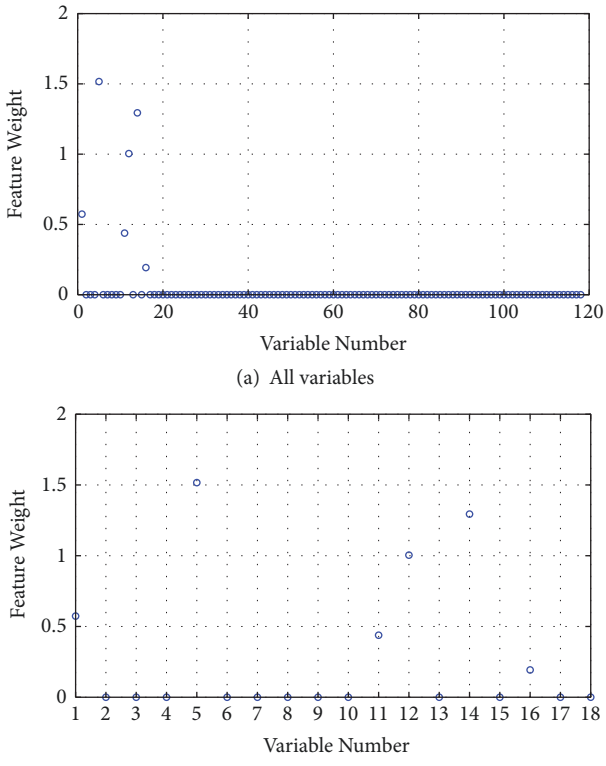


FIGURE 3: The feature weights of variables (after parameter optimization).

be longer, and the calculation results are consistent with the actual situation. Generally, the more vehicles involved, the larger the type of vehicles, and the more the departments involved in rescue, the longer the duration of the incident, and the calculation results are in agreement with the actual situation.

*3.3. RF Parameters Optimized Using BOA.* Training set is constructed using 6 feature variables selected by NCA. Before RF is trained, the RF parameters need to be determined,

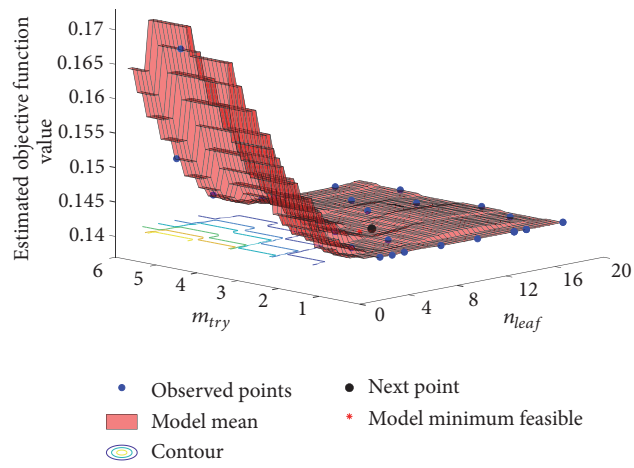


FIGURE 4: The objective function model.

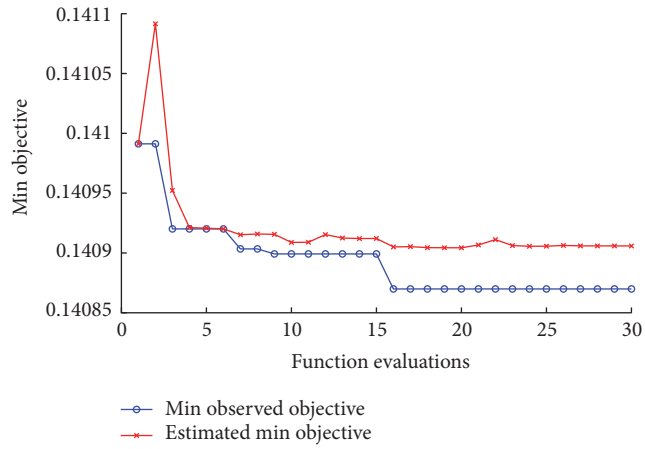


FIGURE 5: The relationship between function evaluations and the minimum objective.

including the number of trees $n_{tree}$, the number of leaves per tree $n_{leaf}$, and the number of random variables used for each node split $m_{try}$. Increasing the number of decision trees can improve the classification accuracy to a certain extent, but will reduce the efficiency of the algorithm. If the minimum classification loss is taken as the goal, the number of decision trees will increase dramatically. Therefore, the number of trees is not optimized in this study, but the other two parameters are optimized for improving the classification accuracy. If $n_{leaf}$ value is too large, it will easily lead to overfitting; if $n_{leaf}$ value is too small, it will easily lead to underfitting. The RF parameters $n_{leaf}$ and $m_{try}$ are tuned using BOA. The RF parameters are set as follows: $n_{tree} = 300$, $n_{leaf} \in [1, 20]$, and $m_{try} \in [1, 5]$. The objective function of BOA is the classification loss of OOB data. Figure 4 shows the objective function model. Figure 5 shows the relationship between function evaluations and the minimum objective. The optimized RF parameters are calculated as $n_{leaf} = 6$ and $m_{try} = 2$, and the observed minimum of the objective function is 0.1409.


Figure 6: The confusion matrices for the results of these three methods.

*3.4. Results and Discussion.* In this section, not only is the test set used to analyze NCA-BOA-RF performance, but also classification and regression tree (CART) [15] and support vector machine (SVM) [20] are introduced for comparison. The parameters of SVM that need to be determined include kernel function parameter and penalty coefficients. The parameters of CART that need to be determined include learning rate and the maximum number of node splits. In order to ensure comparison methods have good performance, BOA is also used to optimize the parameters of SVM and CART.

Figure 6 shows the confusion matrices for the results of these three methods. The confusion matrix can provide detailed classification results of the algorithm [38]. The rows represent the output class (predicted class) and the columns represent the target class (true class). The diagonal cells (green) represent observations that are classified correctly. The off-diagonal cells (red) represent observations which are classified incorrectly. The column on the far right of the plot shows the percentages of all the test samples predicted to belong to each class that are correctly and incorrectly classified. These metrics are often called the precision (or positive predictive value) and false discovery rate, respectively. The row at the bottom of the plot shows the percentages of all the test samples belonging to each class that are correctly and incorrectly classified. These metrics are often called the recall (or true positive rate) and false negative rate, respectively. The cell (blue) in the bottom right of the plot shows the overall classification accuracy.

Therefore, according to confusion matrices obtained as a result of the three methods, the overall classification accuracy

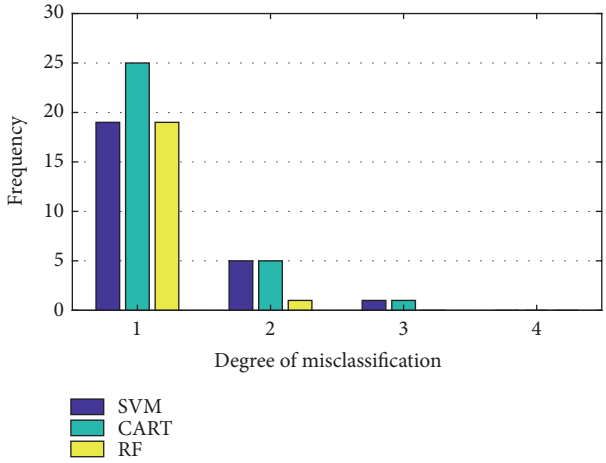


Figure 7: The frequency distribution of the misclassification degree of these three methods.

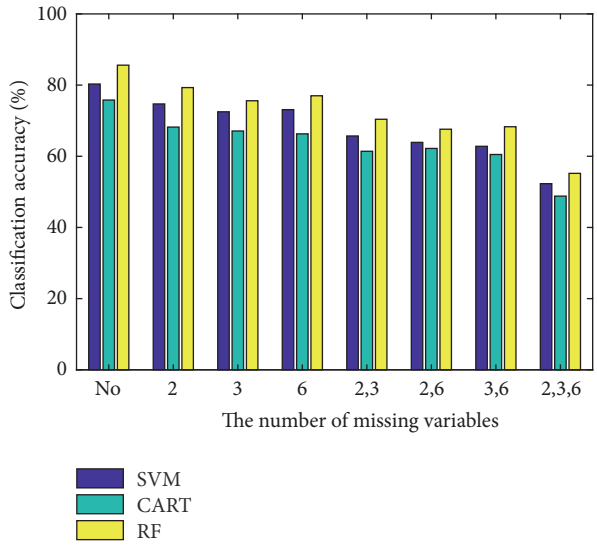


Figure 8: The classification accuracy of the three methods in the absence of variables.

of RF is the highest one. For the "5" class samples, the classification accuracy of the three methods is equal (=33.3%), indicating that the three methods have a poor classification performance on the "5" class samples. This is due to the fact that the number of the "5" class samples is too small. For the "3" class samples, the classification accuracy of RF and SVM is equal (=84.0%). In addition, the classification accuracy of RF is higher than that of SVM and CART on the samples of the remaining classes.

From the confusion matrix, the degree of misclassification can be achieved. For example, if the "1" class is classified into the "2" class, the degree of misclassification is 1; if the "1" class is classified into the "3" class, the degree of misclassification is 2. Figure 7 shows the frequency distribution of the misclassification degree for these three methods. It can be seen from Figure 7 that the misclassification degree of RF is mainly 1, a small amount is 2, and there are no 3 and 4, while the other two methods have a misclassification degree of 3, and the number of misclassification degree of 2 is more than that of RF significantly.

At the beginning of the traffic incident, the available information for incident duration prediction is often limited. Therefore, we need to analyze the performance of the method in the absence of some variables. The six feature variables are 1 "incident type", 2 "number of lanes closed", 3 "need truck wrecker", 4 "need firefighters", 5 "automobile count", and 6 "heavy truck count". When an incident is first identified, 1 "incident type", 4 "need firefighters", and 5 "automobile count" are often the first to be known, so the absence of these three variables is not considered.

In the absence of variables, the classification accuracy of the three methods is shown in Figure 8. As can be seen from Figure 8, when a single variable is missing, the classification accuracy of the three methods reduces to a certain extent; when two variables are missing, the classification accuracy of the three methods continues to decrease; when three variables are missing, the classification accuracy of the three methods is greatly reduced to about 50%. However, from the comparison of the three methods without some variables, the classification accuracy of RF is still higher than that of SVM and CART.

## 4. Conclusions

In this study, a hybrid method named NCA-BOA-RF is proposed to integrate the NCA and the BOA-optimized RF model for traffic incident duration prediction. Firstly, 18 influencing factors are selected as the relevant variables of traffic incident duration, considering influencing factors of traffic incident duration and the data set that we used. Secondly, feature weights of the relevant variables are calculated by Neighborhood Components Analysis (NCA), and feature variables of traffic incident duration are determined by the feature weights. Secondly, the NCA is applied to select the most powerful features (called feature variables) for traffic incident duration prediction. In NCA, regularization parameter is optimized using cross-validation to ensure better classification accuracy (less classification loss). Then, the training set is constructed using the feature variables to train RF model, and the BOA is used to optimize the RF parameters. Finally, we conduct experiments to test performance of the proposed method and two comparison methods. Confusion matrix was introduced to better illustrate the experimental results. Not only has the proposed method the highest classification accuracy on the whole data, but also the classification accuracy of the proposed method is greater than or equal to that of the other two methods on each class of the data. In addition, the performance of the proposed method is tested with some feature variables unavailable. The results demonstrate that the performance is still better than comparison methods, although the performance of all methods is reduced. Based on the comparison and analysis of the experimental results, the conclusions can be drawn that NCA-BOA-RF is a better method for traffic incident

duration prediction, because of its high accuracy rate and good generalization ability.

For future research, more and more comprehensive data sets should be used to further test the method performance for drawing a more general conclusion. From an incident being detected to the incident being cleared up, more useful information is expected to be available. Multistage updates of information should be considered in future study. In addition, predicting different stages of event duration (such as response time) separately is also a direction for future research.

## Data Availability

The traffic incident data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This research is supported by the National Natural Science Foundation of Shandong Province (Grant Nos. ZR2018BF024 and ZR2016EL19), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 18YJC190003), the National Natural Science Foundation of China (Grant No. 61573009), and the Dr. Scientific Research Start Funding Projects of Shandong University of Technology (Grant Nos. 4041/417006, 4033/718003).

## References

[1] Y. Chung, H. Kim, and M. Park, "Quantifying non-recurrent traffic congestion caused by freeway work zones using archived work zone and ITS traffic data," *Transportmetrica*, vol. 8, no. 4, pp. 307–320, 2012.

[2] W. Kim and G.-L. Chang, "Development of a hybrid prediction model for freeway incident duration: a case study in Maryland," *International Journal of Intelligent Transportation Systems Research*, vol. 10, no. 1, pp. 22–33, 2012.

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

[4] A. Khattak, X. Wang, and H. Zhang, "Incident management integration tool: Dynamically predicting incident durations, secondary incident occurrence and incident delays," *IET Intelligent Transport Systems*, vol. 6, no. 2, pp. 204–214, 2012.

[5] H. Zhang and A. Khattak, "What is the role of multiple secondary incidents in traffic operations?" *Journal of Transportation Engineering*, vol. 136, no. 11, pp. 986–997, 2010.

[6] Y. Lou, Y. Yin, and S. Lawphongpanich, "Freeway service patrol deployment planning for incident management and congestion mitigation," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 2, pp. 283–295, 2011.

[7] D. Nam and F. Mannering, "An exploratory hazard-based analysis of highway incident duration," *Transportation Research Part A: Policy and Practice*, vol. 34, no. 2, pp. 85–102, 2000.

[8] W. M. Liu, L. P. Guan, and X. Y. Yin, "Prediction of incident duration based on multiple regression analysis," *Journal of Highway & Transportation Research & Development*, vol. 22, no. 11, pp. 126–129, 2005.

[9] A. J. Khattak, J. L. Schofer, and M. Wang, "A simple time sequential procedure for predicting freeway incident duration," *Journal of Intelligent Transportation Systems*, vol. 2, no. 2, pp. 113–138, 1995.

[10] A. Tavassoli Hojati, L. Ferreira, S. Washington, and P. Charles, "Hazard based models for freeway traffic incident duration," *Accident Analysis & Prevention*, vol. 52, pp. 171–181, 2013.

[11] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Competing risk mixture model and text analysis for sequential incident duration prediction," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 74–85, 2015.

[12] R. Li, M. Guo, and H. Lu, "Analysis of the Different Duration Stages of Accidents with Hazard-Based Model," *International Journal of Intelligent Transportation Systems Research*, vol. 15, no. 1, pp. 7–16, 2017.

[13] B. Jiyang, X. Zhang, and L. Sun, "Traffic incident duration prediction grounded on Bayesian decision method-based tree algorithm," *Tongji Daxue Xuebao/Journal of Tongji University*, vol. 36, no. 3, pp. 319–324, 2008.

[14] H. L. Chang and T. P. Chang, "Prediction of freeway incident duration based on classification tree analysis," *Journal of the Eastern Asia Society for Transportation Studies*, vol. 10, pp. 1964–1977, 2013.

[15] C. Zhan, A. Gan, and M. Hadi, "Prediction of lane clearance time of freeway incidents using the M5P tree algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1549–1557, 2011.

[16] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing Influential Factors for Freeway Incident Clearance Time Prediction Using the Gradient Boosting Decision Trees Method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2303–2310, 2017.

[17] S. Demiroluk and K. Ozbay, "Adaptive learning in bayesian networks for incident duration prediction," *Transportation Research Record*, vol. 2460, no. 1, pp. 77–85, 2014.

[18] S. Wang, R. Li, and M. Guo, "Application of nonparametric regression in predicting traffic incident duration," *Transport*, vol. 2015, pp. 1–10, 2015.

[19] B. Yu, Y. T. Wang, J. B. Yao, and J. Y. Wang, "A comparison of the performance of ANN and SVM for the prediction of traffic accident duration," *Neural Network World*, vol. 26, no. 3, pp. 271–287, 2016.

[20] H. Park, A. Haghani, and X. Zhang, "Interpretation of Bayesian neural networks for predicting the duration of detected incidents," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 20, no. 4, pp. 385–400, 2016.

[21] Y. Lee, C.-H. Wei, and K.-C. Chao, "Non-parametric machine learning methods for evaluating the effects of traffic accident duration on freeways," *Archives of Transport*, vol. 43, no. 3, pp. 91–104, 2017.

[22] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Overview of traffic incident duration analysis and prediction," *European Transport Research Review*, vol. 10, no. 2, p. 22, 2018.

[23] Y. Lee and C.-H. Wei, "A computerized feature selection method using genetic algorithms to forecast freeway accident duration times," *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 2, pp. 132–148, 2010.

[24] Q. Shang, Z. Yang, S. Gao, and D. Tan, "An Imputation Method for Missing Traffic Data Based on FCM Optimized by PSO-SVR," *Journal of Advanced Transportation*, vol. 2018, Article ID 2935248, 21 pages, 2018.

[25] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, vol. 17, pp. 513–520, 2005.

[26] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *Journal of Computers*, vol. 7, no. 1, pp. 162–168, 2012.

[27] M. Jin and W. Deng, "Predication of different stages of Alzheimer's disease using neighborhood component analysis and ensemble decision tree," *Journal of Neuroscience Methods*, vol. 302, pp. 35–41, 2018.

[28] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012*, pp. 2951–2959, USA, December 2012.

[29] A. Klein, S. Falkner, S. Bartels, P. Hennig, and F. Hutter, "Fast Bayesian hyperparameter optimization on large datasets," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 4945–4968, 2017.

[30] L. Wang, M. Feng, B. Zhou, B. Xiang, and S. Mahadevan, "Efficient hyper-parameter optimization for NLP applications," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pp. 2112–2117, Portugal, September 2015.

[31] E. Brochu, V. M. Cora, and N. De Freitas, *A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning*, 2010, https://arxiv.org/abs/1012.2599.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] Z. Bei, Z. Yu, N. Luo, C. Jiang, C. Xu, and S. Feng, "Configuring in-memory cluster computing using random forest," *Future Generation Computer Systems*, vol. 79, pp. 1–15, 2018.

[34] F. Ouallouche, M. Lazri, and S. Ameur, "Improvement of rainfall estimation from MSG data using Random Forests classification and regression," *Atmospheric Research*, vol. 211, pp. 62–72, 2018.

[35] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[36] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests," *Psychological Methods*, vol. 14, no. 4, pp. 323–348, 2009.

[37] U. Grömping, "Variable importance assessment in regression: linear regression versus random forest," *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.

[38] E. Vigneau, P. Courcoux, R. Symoneaux, L. Guérin, and A. Villière, "Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception," *Food Quality and Preference*, vol. 68, pp. 135–145, 2018.