Short-term prediction of ammonia levels in goose houses via combined feature selector and random forest

Jiande Huang†, Shahbaz Gul Hassan†, Longqin Xu, Shuangyin Liu*

(*College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China*)

Abstract: Ammonia concentration (NH_3) is a dominant source of environmental pollution in geese housing and profoundly affects the healthy growth of geese. Accurately forecasting NH_3 and analyzing its change trends in geese houses is crucial for the survival of geese. A novel forecasting model was proposed by combining feature selector (CFS) and random forest (RF) to improve the prediction accuracy of NH_3 in this study. The developed model integrated two modules. First, combining mutual information (MI) and relief-F, we propose that CFS quantify each feature's importance values and eliminate the low-relation or unrelated features. Second, a random forest model was built using *K*-fold cross-validation grid search algorithm (CVGS) to obtain the RF hyperparameters to predict NH_3. The simulation results show that the prediction accuracy was improved when feature selection after quantification based on the CFS was used. The mean square error (MSE), root mean square error (RMSE), and mean absolute percent error (MAPE) for the proposed model were 0.5072, 0.6583, and 2.88%, respectively. The NH_3 prediction model (CFS-CVGS-RF) based on Combined Feature Selector, cross-validation grid search algorithm (CVGS), and Random Forest (RF) exhibited the best prediction accuracy and generalization performance compared with other parallel forecasting models and is a suitable and useful tool for predicting NH_3 in geese houses. The results of the research can provide a reference for the machine learning method to monitor the dynamic changes of ammonia in goose houses.

Keywords: ammonia concentration prediction, random forest, combined feature selector, goose houses

DOI: 10.25165/j.ijabe.20231606.6378

Citation: Huang J D, Hassan S G, Xu L Q, Liu S Y. Short-term prediction of ammonia levels in goose houses via combined feature selector and random forest. Int J Agric & Biol Eng, 2023; 16(6): 77–84.

1 Introduction

China is the largest geese producer in the world. According to the data published by the China Statistics Bureau in the 2019 National Economic and Social Development Statistical Bulletin, the output of poultry meat and poultry eggs in China was 22.39 million t and 33.09 million t, respectively. The total output value of waterfowl exceeds 160 billion RMB yuan[1]. Due to the increasing consumption of poultry products and an increase in exports, more poultry coops will be built. Historically, outbreaks of poultry disease and even mass deaths are almost inevitable due to the rapid expansion of breeding scale, lack of scientific management, and environmental degradation of the poultry housing[2]. Presently, NH_3 produced in poultry houses poses the most significant concern for poultry health[3]. NH_3 produced through the decomposition of feces and urine by microorganisms is a primary factor in the environmental pollution in poultry coops and can damage the health of the respiratory system, eyes, paranasal sinuses, skin, and other organs[4,5]. The high concentrations of NH_3 directly harm poultry's immune function, health, and growth capability[3,6,7], giving rise to various diseases and leading to economic losses. Currently, research on the influence of NH_3 on poultry has focused on laying and broiler hens. In contrast, the effects of NH_3 on geese have not been widely discussed[8]. However, meat farming and laying geese in China are gradually transforming from outdoor or semi-outdoor to indoor. Therefore, the hazards associated with NH_3 are expected to pose a severe problem for meat geese production. Due to the severe implications of NH_3 on poultry health, there is a need to provide a beneficial and stable environment for poultry, which is suitable for the complex nature of NH_3 for its modeling and prediction.

In recent years, several intelligent algorithms have been proposed to predict NH_3 levels[9-13]. Artificial neural networks (ANN), support vector machine (SVR), and decision tree (DT) are useful tools. They have been widely used for solving complex prediction problems. However, DT often faces the over-fitting issue, so it performs well on the training data set but not on the test set. SVM uses the quadratic programming approach to measure the supporting vector making it difficult in large-scale training sets to implement and make its output heavily dependent on the choice of various hyperparameters[14]. Ensemble learning integrates the prediction of several foundation estimators established with a given learning algorithm to enhance the generalization performance over a single estimator. It has become a hotspot in prediction and has been successfully applied in some fields[15]. Random forest (RF) is a representative ensemble learning method. Compared with the methods mentioned above, RF with fewer hyper-parameters seldom over-fits and is relatively robust to outliers and noise[16-18].

Many studies have indicated that changes in NH_3 are related to temperature, humidity, and other environmental factors. Xie et al.[12] used an adaptive neuron fuzzy inference system (ANFIS) to predict NH_3 using indoor relative humidity, indoor temperature, pig temperature, and other indicators. Zhu et al.[19] predicted NH_3 based on a genetic algorithm (GA) and optimized backpropagation neural

Received date: 2020-12-25 **Accepted date:** 2022-12-29

Biographies: Jiande Huang, Master candidate, research interest: multimodal prediction, Email: jiande.huang@icloud.com; **Shahbaz Gul Hassan**, Lecturer, research interest: nonlinear predictive modeling, Email: mhasan387@zhku.edu.cn; **Longqin Xu**, Professor, research interest: agricultural informatics, Email: xlqlw@126.com.

†The authors contributed equally to this work.

***Corresponding author: Shuangyin Liu**, Professor and Dean of College of Information Sience and Technology, Zhongkai University of Agriculture, research interest: modeling and information systems for agriculture, Email: shuangyinliu@zhku.edu.cn.

network. Temperature, relative humidity, carbon dioxide, total suspended particulates, solar radiation, and atmospheric pressure have been usually used as prediction indicators[20]. While these models select some indicators as inputs to predict NH_3, few studies have considered the correlation between each feature and NH_3, and the methods used in these studies display several shortcomings. For example, using vast data directly in intelligent models without feature selection increases the training time and the risk of over-fitting.

Moreover, the contribution of the algorithm optimization and model combination may be lower than that of screening for good prediction indicators[18]. Thus, feature selection is necessary. Extraction and selection are common operations in feature engineering. Feature selection is better than feature extraction concerning readability and understandability and will not alter the primitive feature data[21]. Recently, mutual information-based algorithms (MI) have played an increasingly significant role in data mining and machine learning. These methods have good non-linear and linear processing capabilities for considering the relation of diverse sets of features[22,23]. However, mutual information algorithms ignore the influence of the proportions of labels on the correlation degree between features and label sets[24]. The Relief algorithm presented by Kira et al.[25] was initially used for two-category problems. Relief-F, an extensively adaptable filter-based feature evaluation technique[26], was presented to cope with multi-label data and regression difficulties and to adapt to many category problems. In the study of Wang et al.[21], the fusion of mutual information and relief-F was proposed to improve feature selection capability and bring a more accurate feature selection.

In this study, the NH_3 prediction was investigated. The aim of this study was to accurately forecast NH_3 levels using the data from an intelligent goose house Internet of Things system. To overcome this challenging problem, an RF-underpinned framework was designed that effectively predicts the NH_3 level. Although RF is a promising approach, two challenges must be addressed. One is that unrelated and redundant features give rise to a high cost of the RF training process and decrease prediction accuracy. The second one is the tuning of the parameters. There are three hyper-parameters: the number of a tree, the size of sampling subsets, and the minimum number of samples required to split an internal node. These hyper-parameters affect the performance of RF, and currently, there is no consensus regarding how their values should be set.

This study combined the feature selector with the RF and designed a new hybrid prediction model for predicting NH_3 in goose houses to address the abovementioned challenges. First, the combined feature selector (CFS) that combines the MI and relief-F evaluates the importance of prediction indicators. Then, the parameter M was controlled to eliminate the low-importance features. Finally, the parameters of RF are used by CVGS to build a model for predicting NH_3 in geese houses. The hybrid model makes full use of the CFS and is highly suitable for selecting the prediction indicators in this study.

2 Materials and methods

2.1 Study area and data source

The data used in this study were obtained from a waterfowl breeding farm in Haifeng County (23°05'N, 115°19'E) in Shanwei City, China. With an area of approximately 53.3 hm^2, the farm is a multifunctional integrated aquaculture base integrating waterfowl breeding, seeding breeding, and intensive aquaculture. In this experiment, the animals (stone goose) in the area were housed, including a poultry house (25×16 m^2), a playground (25×30 m^2), and a swimming pool (20×3×1 m^3). Fans, control equipment (temperature and light), and various sensors were installed for online monitoring of aquatic environment parameters in the geese houses (Figure 1).

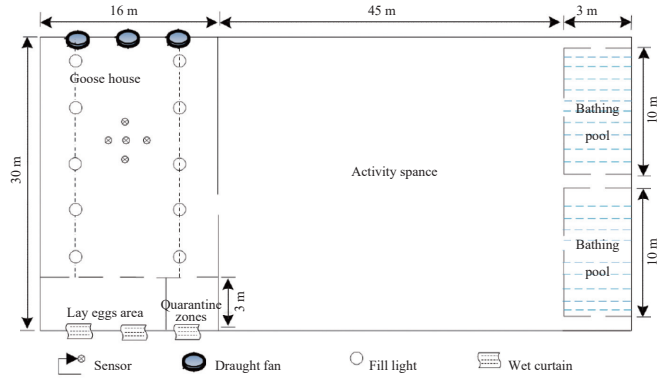


Figure 1 Schematic diagram of the breeding environment monitoring based on the Internet of Things.

Because geese house environment parameters are mainly affected by physical and chemical factors, an IoT system (Figure 2) was developed to monitor temperature, humidity, carbon dioxide (CO_2), total suspended particulates (TPS), NH_3 of the houses and temperature, humidity, atmospheric pressure, and solar radiation of the surrounding environment.

2.2 Combined Feature Selector

To fully evaluate feature importance between each feature and target, MI and relief-F are both used in CFS. These processes compute the feature importance of each feature to decide whether a feature is eliminated or preserved. The parameter M is the threshold. The feature can be eliminated with low feature importance by controlling M. In this section, a module was introduced. Figure 3 shows the processes of CFS.

2.2.1 Relief-F

The relief algorithm is an effective filtered feature selection method proposed by Kira and Rendell[25]. The relief algorithm is initially limited to the classification of two types of data, so the relief-F algorithm, which Kononeill later extends, can solve the multi-class and regression problems[26]. This algorithm is a weighted algorithm that assigns weights to each feature according to the relevance of the target. The larger the feature weight, the higher the contribution of the feature, and vice versa, the lower the feature classification contribution.

The relief-F algorithm estimates feature weight according to the degree of distinguishing samples close to each other based on the value of the feature. The relief-F algorithm randomly selects a simple X_i (X_i has classified p) from training set D, which has $|y|$

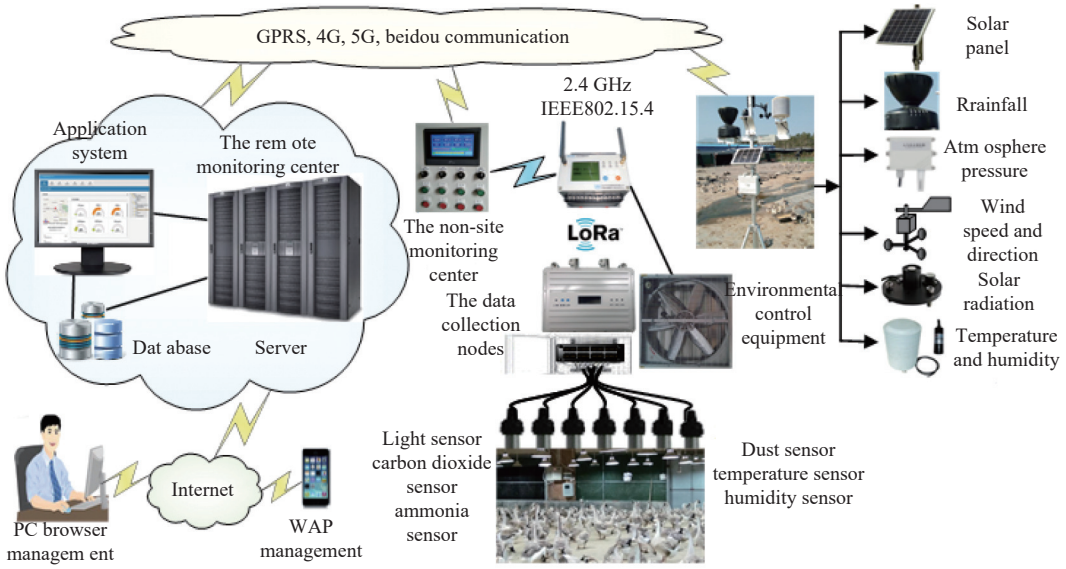
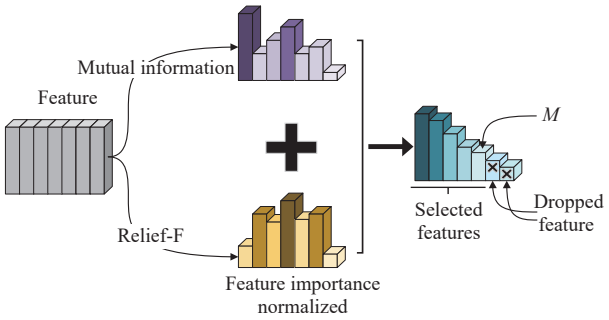


Figure 2 Topology structure diagram of the platform



Note: M is a parameter used to eliminate features with low feature importance.

Figure 3 Structure of combined feature selector (CFS)

class. Then searches for k of its nearest neighbors from the same class, called near-hit $X_{i,nh}$, and also k nearest neighbors from each of the different classes, called near-miss $X_{i,j,nm}$ (j=1, 2, …, $|y|$; $j\neq p$), then the weight of feature L (δ^L) can be computed as follows:

$$\delta^L = \sum_i -\text{diff}\left(X_i^L, X_{i,nh}^L\right)^2 + \sum_{j\neq p}\left(qj\times\text{diff}\left(X_i^L, X_{i,j,nm}^L\right)^2\right) \quad (1)$$

where, qj is the proportion of class j samples in data set D, diff(a^j, b^j) denoted distance between simple a and b in feature j, diff() is defined as,

$$\text{diff}\left(a^j, b^y\right) = \begin{cases} \dfrac{|a^y - b^j|}{\max(j)-\min(j)}, & \text{if } j \text{ is continuous} \\ 0, & \text{if } j \text{ is discrete and } a^j \neq b^j \\ 1, & \text{if } j \text{ is discrete and } a^j = b^j \end{cases} \quad (2)$$

2.2.2 Mutual information estimator

MI was selected based on its information theory background among the estimates of independence between random variables. MI(X, Y) between the two random variables X and Y is defined by the common information found in two variables with a joint probability distribution $P(X, Y)$. MI(X, Y) computes the degree of correlation between vector X and target vector Y and is given by

$$\text{MI}(X,Y) = \int_y\int_x P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)\text{d}x\text{d}y \quad (3)$$

where, $P(x, y)$ is the probability density function of random variable Z=(X, Y), $P(x)$ and $P(y)$ represent the marginal probability density function of X and Y, respectively. In fact, since $P(x, y)$ is usually unkonwn in advance, some methods should be used to estimate MI(X, Y). K-nearest neighbor is a non-parametric method that has been confirmed to be useful in MI estimation[27]. For the K-nearest neighbor, Euclidean distance was used as a distance metric, and the maximum norm for the space Z=(X, Y) is written as,

$$\|z-z'\| = \max\{\|x-x'\|, \|y-y'\|\} \quad (4)$$

Let $\kappa(i)/2$ represent the distance from z_i to its kth neighbor, and $\kappa_x(i)/2$ and $\kappa_y(i)$ represent the distance between the same points projected into the X and Y subspaces. It is clear that:

$$\kappa(i) = \max\left\{\kappa_x(i), \kappa_y(i)\right\} \quad (5)$$

Then, we denote by $\eta_x(i)$ the number of points for which the distance from x_i is strictly less than $\kappa(i)$ and by $\eta y(i)$ the number of points for which the distance from y_i is strictly less than $\kappa(i)$. This study noted that $\kappa(i)$ is not a fixed value, and $\eta_x(i)$ and $\eta_y(i)$ is also not fixed. $\langle\cdots\rangle$ denotes both all $i\in(1, ..., N)$ and all realizations of the random samples.

$$\langle\cdots\rangle = \frac{\sum_{i=1}^{N} E[\cdots(i)]}{N} \quad (6)$$

The estimate for MI by K-nearest neighbor is then:

$$I(X,Y) = \psi(k) - \left\langle\psi(\eta_x+1)+\psi\left(\eta_y+1\right)\right\rangle + \psi(N) \quad (7)$$

where, ψ() is the di-gamma function and satisfies the recursion $\psi(x+1)=\psi(x)+1/x$ and $\psi(1)=-C$, where C=0.577 215 6 is the Euler-Mascheroni constant. If MI is equal to 0, the two random variables are independent and higher MI values mean higher dependency.

2.3 Random Forest.

The RF algorithm is a non-linear ensemble model that establishes and averages a large number of random distribution DT for regression or classification tasks[28]. A DT or classification and regression tree (CART) that constructs the RF is a non-parametric model. According to the complexity of the input data, the tree grows in the learning process. Decision nodes and leaf nodes are the main components of DT. Each input sample is estimated by a test function of decision nodes and passed to different branches according to the features of the sample. Let us denote by X={x_1, x_2, x_3, ..., x_n} the input vector with n features, Y is the output scalar, and D_m is the training set with m observations which can be written

as follows:

$$D_m = \{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3) \ldots, (X_m, Y_m)\}, X \in R^n, Y \in R \quad (8)$$

At each node, the input data are split by a specific algorithm in the process of training to optimize the parameters of the split function to the fit data set D_n. In the first step, the DT must be optimally split among all variables. The splitting procedure begins at the root node, and each node uses its split function for the new input X. This operation is recursive until a leaf node appears. The tree stops growing either when the maximum number of levels is reached or when the observation number of a node is less than a predefined number. The result of the DT learning process is a prediction function $\dot{T}(D_m, X)$ generated over D_m.

The RF regression model can offer powerful prediction ability and is an extension of the DT. The main characteristics of RF include bootstrap resampling and random feature subsets. An RF is an ensemble of P DT $\dot{T}(D_m^1, X), \dot{T}(D_m^2, X), ..., \dot{T}(D_m^p, X)$. Here, $(D_m^1, D_m^2, ..., D_m^p)$ are the bootstrap samples obtained by random sampling of m observations with replacement from D_m, where each observation has the probability of being drawn was $1/m$. This sample process is known as bootstrap resampling. During the splitting of each node, only a small part of n features are randomly selected instead of all features; this is known as random feature selection. The ensemble learning result P output $\dot{Y}_1 = \dot{T}(D_m^1, X)$, $\dot{Y}_2 = \dot{T}(D_m^2, X)$, ..., $\dot{Y}_p = \dot{T}(D_m^p, X)$. Then, the final estimation output $\dot{Y}$ is the average of P output, which is described as follows:

$$\dot{Y} = \frac{1}{P}\sum_{i=1}^{P} \dot{Y}_i = \frac{1}{P}\sum_{i=1}^{P} \dot{T}(D_m^1, X) \quad (9)$$

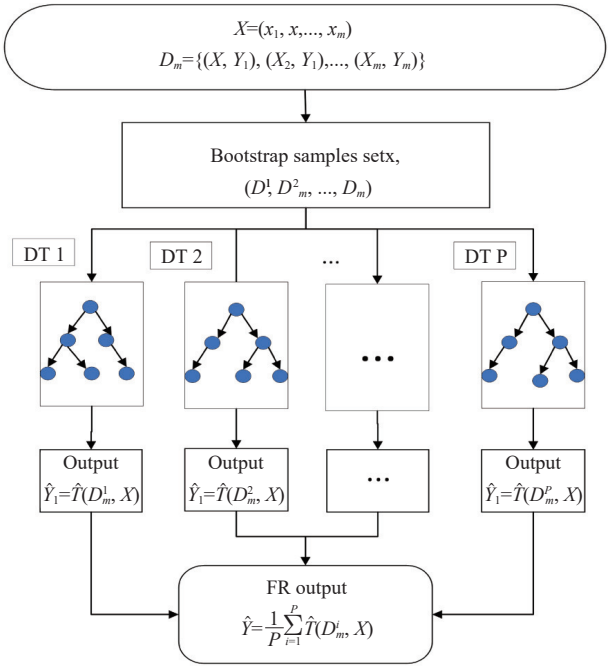
where, $\dot{Y}_i$ is the output of ith DT, i=1, 2, 3, …, P. The framework of RF regression is illustrated in Figure 4, and its training process can be summarized as:

Step 1: Obtain bootstrap samples from the training data set by bootstrap resampling;

Step 2: Generate a regression DT by full use of the bootstrap sample drawn in step 1 with the following modification: at each node, select the optimal split among a random subset sampled in input variables (mtry) instead of all of them;

Step 3: Repeat Steps 1 and 2 until the P DT tree is generated;

Step 4: Aggregating the output of P tress by an average method to forecast unknown data.



Note: DT: Decision tree.

Figure 4 Framework of random forest regression

2.4 Cross-validation grid search

According to the previous section, the RF algorithm was noticed to have two important parameters: 1) P is the number of decision trees that are the base estimators of RF; 2) mtry is the size of the random feature subset.

Generally, a variance of RF decreases as P grows. More accurate predictions are likely to be obtained by choosing a large number of trees, but there is no common setting for P[29]. Additionally, mtry is also a sensitive parameter, and increasing mtry can improve the intensity of each DT but the relation among DT will also be increased. This means that the total strength of RF may be decreasing. Therefore, it is necessary to optimize the parameters of RF and select the optimal RF parameters.

The grid search (GS) algorithm[29], currently the most widely used method for parameter optimization, is a highly suitable model with fewer hyperparameters. GS exhaustively generates candidates from a grid of parameter values specified by the user parameters, then trains each candidate set of parameters and marks the model's score, finally obtaining the optimal combination of parameters. GS optimizes all of the model's parameters to guarantee that the given best parameter combination is the optimal global solution in the pre-setting grid.

At the same time, learning the parameters of a model and testing it on the same data set is a mistake. In this case, the model that merely repeats the samples it learned will have a high score. We combine K-fold cross-validation (CV) and grid search (GS) to avoid this. The score is evaluated based on the mean square error (MSE) average value given from K iterations. Finally, the optimal parameter combination with the lowest MSE is obtained. The specific steps are described as follows:

Step 1: Partitioning the train data into equal-size k sets;

Step 2: Setting the scale of all of the parameters and exhaustively generating candidates from the parameter space;

Step 3: A model with parameters combination is trained using $K-1$ of the folds as the training set. The MSE of the model is computed on the remaining part of the data;

Step 4: Each of the K folds followed the Step 3 procedure;

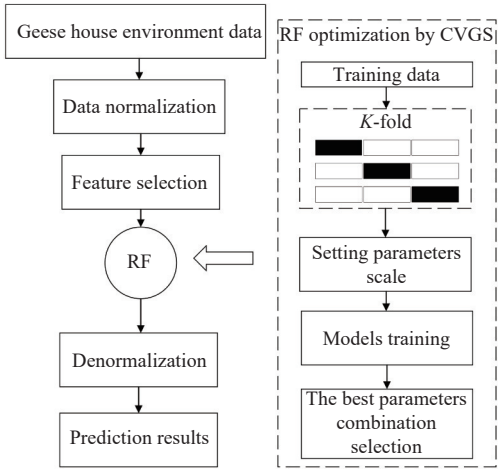
Step 5: The score measured by CVGS is the average of the values computed in Step 4. Then, the parameter combination with the lowest MSE is identified as optimal.

3 Hybrid prediction model based on CFS-CVGS-RF

This study proposed a CFS-CVGS-RF model to predict NH_3 in a geese house. The methodology for conducting this model is shown in Figure 5. The implementation process for NH_3 prediction based on CFS-CVGS-RF can be described as follows:

Step 1: Data normalized processing. Different data of geese house environment has other units and dimensions. Using the original data directly will make the model complex and decrease the prediction performance. To address this problem, we use normalization that can eliminate the difference between the data units and dimensions and facilitate the study of the correlation between environmental factors. The normalized process method is described by

$$y_i' = \frac{y_i - \bar{y}}{y_{\max} - y_{\min}} \quad (10)$$



Note: RF: Random forest; CVGS: *K*-fold cross-validation grid search algorithm.

Figure 5 The schematic of the proposed methodology.

where, y_i' Are the normalized data; y_{max} and y_{min} are the max and min values of the original data; y_i and $\bar{y}$ are the original data and their mean.

Step 2: Feature selection based on CFS. CFS selects the features that are strongly related to NH_3 and eliminates the low-importance factors. The remaining features are used as the input to the regression model. CFS reduces the dimension of input and solves the problem of information redundancy. In the CFS process, we first compute linear and non-linear correlation strengths between each environmental factor and target (NH_3) by using relief-F and MI, respectively. After normalization, the final feature importance is the sum of two dimensions of the important values of features. Then, threshold *M* is set to the screen factor with feature importance lower the *M*.

Step 3: CVGS-RF modelling. RF has two key parameters, namely, the number of DT *P* and the size of the random feature subset mtry. To find the optimal values of these two parameters, the CVGS method was adopted. In the CVGS process, the grid coordinates of the parameters were first established. In this study, combining the RF-related literature[30] and the experimental parameters of wave motion, set *P*=[2, 1000] and mtry=[2, 4]. Then, the data set is divided into *K* subsets, where 10-fold (*k*=10) cross-validation is considered to be better[31]. After *k* parallel operations, each parameter's combinations have *k* MSE. According to the mean of each calculation result, the parameter was selected in combination with minimum average MSE as the optimal parameters and established the RF model using these values.

Step 4: Result output denormalization. Denormalize the output to obtain the results in the normal dimension. The denormalizing process is described by

$$y_i = y'(y_{max} - y_{min}) + \bar{y} \tag{11}$$

where, y_i is the denormalized data, y_{max} and y_{min} are max and min values of original data, and y_i' and $\bar{y}$ are the original data and their mean.

4 Results and discussion

4.1 Data collection

This article used a simulation model to validate the proposed method's performance. The NH_3 was tested in a high-density geese culture farm from September 10th to September 27th, 2019, in Sanwei City, Guangdong province. The details for some data are listed in Table 1. Figure 6 shows the feature importance computed by CFS for each factor, and the performances of different threshold *M* are listed in Table 2.

Table 1 Some of the original experimental data collected on September 10-27, 2019

Time	Indoor temperature /°C	Indoor humidity /(%RH)	CO_2 mg/L	TPS/ ($\mu g \cdot m^{-3}$)	NH_3/ $mg \cdot L^{-1}$	Outdoor temperature /°C	Outdoor humidity /%RH	Atmospheric pressure /Pa	Solar radiation /($W \cdot m^{-2}$)
00:00	25.9	72.1	570	7	18	27.55	93.94	101.11	1.27
00:20	25.7	70.6	578	8	23	27.6	94.07	101.08	1.31
00:40	25.9	69.7	573	7	20	27.34	94.57	101.09	1.3
01:00	25.6	70.6	578	4	21	27.57	94.99	101.11	1.28
…	…	…	…	…	…	…	…	…	…
21:40	26.7	67.7	614	9	19	30.18	68.85	101.05	1.31
22:00	25.9	71.2	550	13	19	29.72	73.88	101.04	1.31
22:40	25.7	68.9	661	15	21	30.06	77.69	101.06	1.31
23:00	26.0	65.6	602	11	20	30.37	82.14	101.08	1.31
23:20	25.8	65.6	574	9	19	30.27	84.44	101.06	1.31
23:40	25.8	65.4	597	12	18	29.8	87.09	101.06	1.31

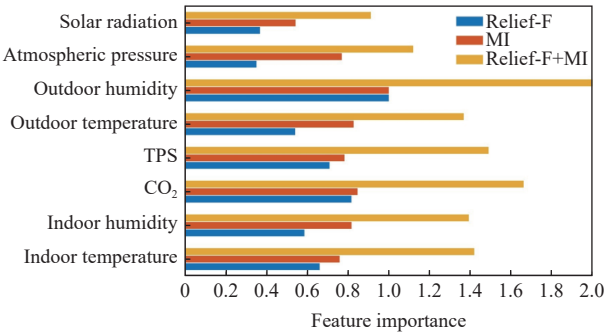


Figure 6 Feature importance computed by Relief-F and MI

Table 2 Simulation results of CFS

M	Accuracy	Running time/s	Eliminated feature
0	82.0%	6.817	
…	…	…	…
1.0	88.1%	6.489	Solar radiation
1.2	94.3%	6.286	Atmospheric pressure
1.4	95.6%	5.850	Outdoor temperature Indoor humidity
1.6	78.5%	5.194	TPS Indoor temperature
1.8	16.1%	4.945	CO_2
2.0	NaN	NaN	Outdoor humidity

The geese house environment data considered in this investigation include the data obtained at intervals of 20 min from September 10 to September 27, 2019. 72 data collected per day yield a total of 1296 observations samples. Some of the original data are listed in Table 1. For model generation, the first 864 sets of the data were used for model training, and the remaining 432 sets were used as the testing data to estimate the prediction performance of the constructed model.

4.2 Performance criteria

For a reasonable evaluation of each prediction model, three commonly used error standards are proposed to measure the model's prediction accuracy, including MSE, RMSE, and MAPE. The relevant calculation formulae are

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (13)$$

$$\text{MaxAPE} = \text{Max}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| \times 100\% \quad (14)$$

4.3 Results and discussion

A script was designed with Python 2.7 vision following the system framework to confirm the superiority of the framework constructed in this work, as described in Section 3. The script was executed on the Win10 operation system with Intel Core i5, 4 GB RAM, and a 500 GB hard disk. The optimal parameter combination (P=500, mtry=3) was used for the prediction model of NH_3 by the CVGS algorithm. According to Figure 6 and Table 2, the used features include outdoor humidity, TPS, CO_2, and indoor temperature as the optimal input for the model to forecast NH_3 in this study.

Following the method introduced in this study, we used relief-F to compute the linear relation strength between each factor and NH_3, and we use MI rather than relief-F to compute the non-linear relation strength.

Figure 6 shows the normalized results of relief-F and MI. It is observed that outdoor humidity has the highest relation and solar radiation has the lowest relation with NH_3. The threshold M was changed from 0 to 2 to find the optimal input feature combination. The sensitivity test results of M were listed in Table 2, and it is observed that the combination of outdoor humidity, TPS, CO_2, and indoor temperature features was the optimal input combination.

Figure 7 shows the change in the fitness value, with the three convergence curves showing the best fit for RF. Three fitness curves show that after growing 100 trees, the MSE decreases very slowly and converges after 500 trees, and mtry=3 results in the lowest MSE. The P and mtry combinational parameters of the optimal CVGS-RF model are 500 and 3, respectively. The MSE values for different parameter combinations are shown in Figure 7. It is observed that after growing 100 trees, the MSE value decreases very slowly and converges after 500 trees, with mtry=3 resulting in the lowest MSE.

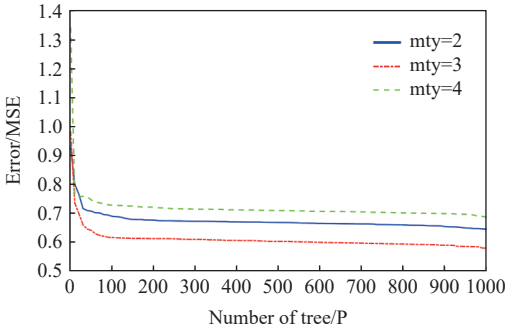
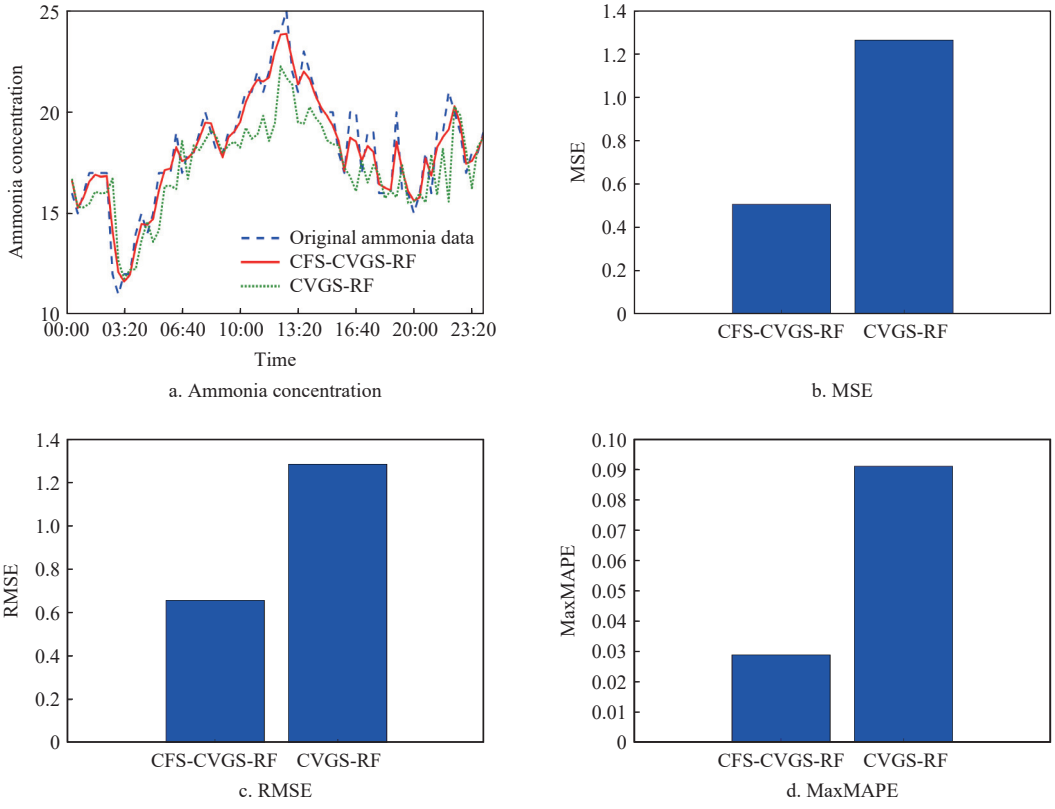


Figure 7 Results of K-fold cross-validation grid search

Two types of comparison were designed to analyze and compare prediction performance: 1) horizontal comparison between the model with CFS and the model without CFS; 2) vertical comparison between the models used in this paper with other parallel models. The horizontal comparison includes the CVGS-RF model. The vertical comparison consists of the benchmarks used in this DT, support vector machine, and back propagation neural network (BPNN). These models used data sets to verify the performance forecasted by the models in this paper. They predicted the NH_3 content of the last 72 test sets corresponding to the last 24 h. Figure 8 and Figure 9 show the prediction curves and the error bar plot. Figure 8 shows the NH_3 series prediction result of the combined model based on CFS-CVGS-RF. The performance estimation statistics of the testing are listed in Table 3.

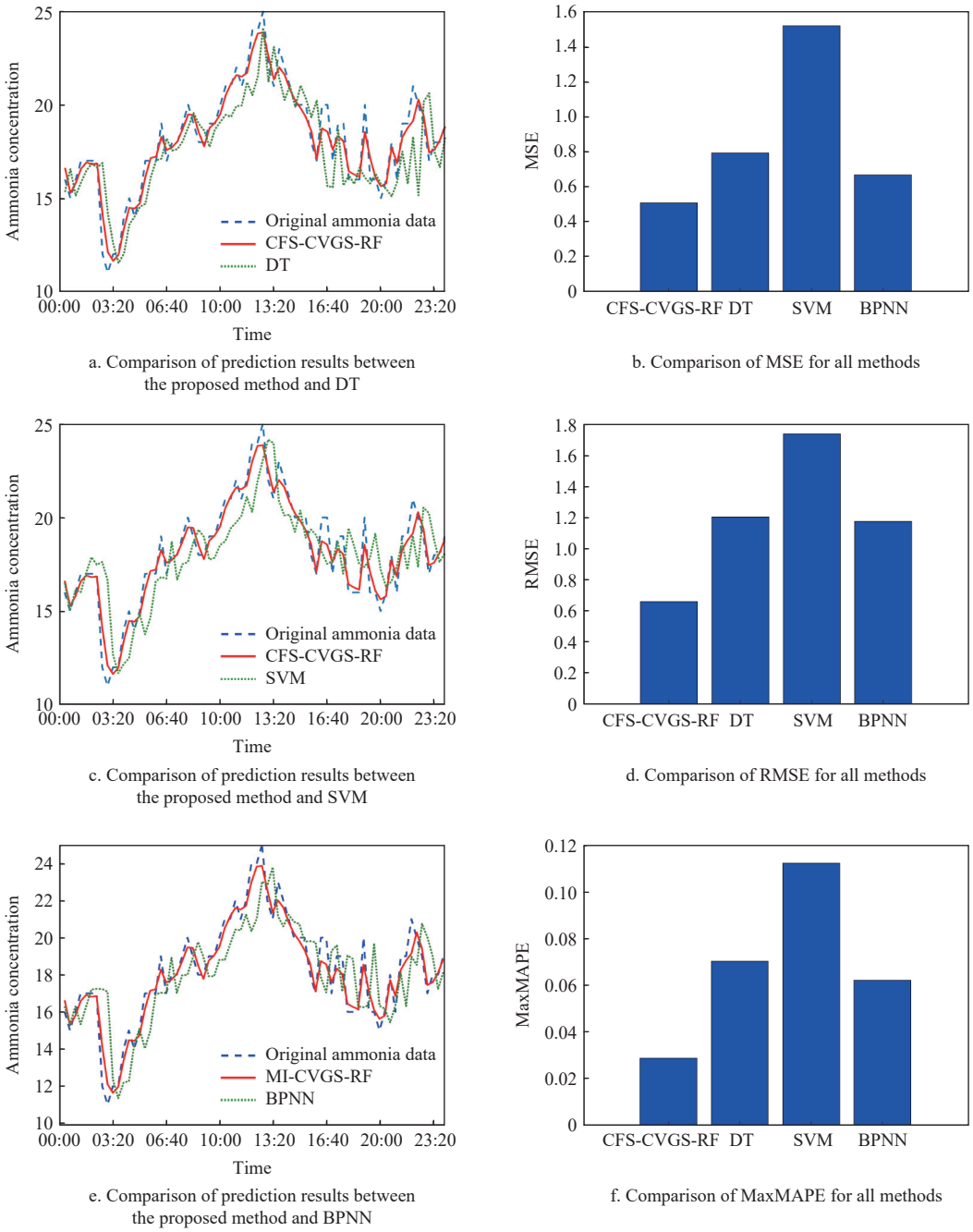
For a more accurate comparison of the performance of five models, this article computes the MSE, RMSE, and MaxMAPE of



a. Ammonia concentration b. MSE

c. RMSE d. MaxMAPE

Figure 8 Horizontal comparison results



a. Comparison of prediction results between the proposed method and DT

b. Comparison of MSE for all methods

c. Comparison of prediction results between the proposed method and SVM

d. Comparison of RMSE for all methods

e. Comparison of prediction results between the proposed method and BPNN

f. Comparison of MaxMAPE for all methods

Figure 9 Vertical comparison results

Table 3 Comparison of NH_3 prediction results

Model	CFS-CVGS-RF	CVGS-RF	DT	SVM	BPNN
MSE	0.5072	1.2658	0.7922	1.5179	0.6667
RMSE	0.6583	1.2851	1.2047	1.7400	1.1764
MaxMAPE/%	2.88	9.10	7.04	11.25	6.23

those models for which the details are shown in Table 3. The MSE, RMSE, and MAPE of CFS-CVGS-RF and GS-RF were 0.5072, 0.6583, 2.88%, and 1.2658, 1.2851, 9.10%, respectively. These values are the best estimation indexes among the five models. Figure 10 shows the prediction residual distribution condition of five models. It is observed that CFS-CVGS-RF has fewer outliers and an overall error closer to zero for the residual error compared with other models. This means that CFS-CVGS-RF has more stability for forecasting results and is more suitable for predicting NH_3.

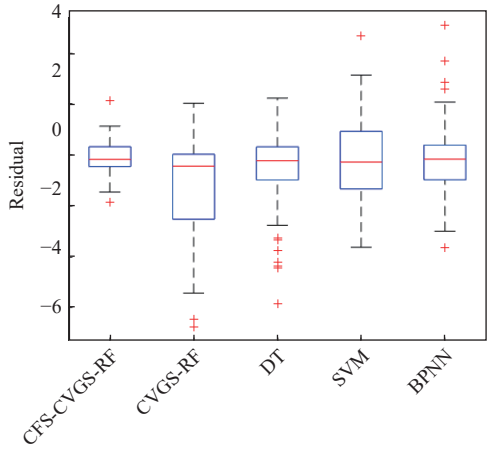


Figure 10 Boxplot of residual error in different models

It is observed from the figures that the prediction curve of the CFS-CVGS-RF model is closer to the original value than the prediction curves of the other four models and has better accuracy than the other four models It can be seen from the prediction error box plot that the CFS-CVGS-RF model has smaller error

fluctuations than the other model, and the CVGS-RF model using feature combination after feature selection by CFS has smaller forecast error fluctuations than CVGS-RF without feature selection, indicating that CFS is effective.

4.4 Conclusions and future work

This study proposed a novel NH_3 prediction hybrid model (CFS-CVGS-RF). The CFS-CVGS-RF model combines four methods: relief-F, mutual information, K-fold cross-validation grid search optimization algorithm, and random forest. The experimental data was collected from a geese house environment in a monitored aquaculture factory farm in Sanwei, China. Results showed that the proposed hybrid method CFS-CVGS-RF has better forecasting performance than CVGS-RF, DT, SVM, and BPNN, as measured by MSE, RMSE, and MaxMAPE. Furthermore, CFS-CVGS-RF can effectively consider the linear and non-linear relations between the input features and the target, reduce redundant information, and improve the model's prediction performance by screening the unrelated or low-relation features.

This study has several limitations that require further research. First, predicting NH_3 is a very complex issue that is influenced by many factors. However, due to equipment limitations, we cannot monitor more related factors that may strongly influence NH_3. Second, concerning experimental time, in the future, we plan to collect data in other months to verify whether the proposed model is useful for a different season. Finally, we plan to investigate using other ensemble strategies instead of a simple averaging method to improve the RF, such as weighted averaging, and weighted voting.

Acknowledgements

This work was financially supported in part by the National Natural Science Foundation of China (Grants No. 61871475; No. 61471-131; No. 61571444), in part by the special project of laboratory construction of Guangzhou Innovation Platform Construction Plan (Grant No. 201905010006), Guangzhou Innovation Platform Construction Plan (Grant No. 2017B010126 0016), the foundation for High-level Talents in Higher Education of Guangdong Province (Grant No. 2017GCZX00014; No. 2016K-ZDXM0013; No. 2017KTSCX094; No. 2018LM2168), and Beijing Natural Science Foundation under (Grant No. 4182023).

[References]

[1] National Bureau of Statistics, China. Statistical Communique of the People's Republic of China on the 2019 National Economic and Social Development.

[2] Zhao Q, Boomer G S, Kendall W L. The non-linear, interactive effects of population density and climate drive the geographical patterns of waterfowl survival. Biological Conservation, 2018; 221: 1–9.

[3] Wei F X, Hu X F, Xu B, Zhang M H, Li S Y, Sun Q Y, et al. Ammonia concentration and relative humidity in poultry houses affect the immune response of broilers. Genetics and Molecular Research, 2015; 14(2): 3160–3169.

[4] Kearney G D, Shaw R, Prentice M, Tutor-Marcom R. Evaluation of respiratory symptoms and respiratory protection behavior among poultry workers in small farming operations. Journal of Agromedicine, 2014; 19(2): 162–170.

[5] Nemer M, Sikkeland L I B, Kasem M, Kristensen P, Nijem K, Bjertness E, et al. Airway inflammation and ammonia exposure among female Palestinian hairdressers: A cross-sectional study. Occupational & Environmental Medicine, 2015; 72(6): 428–434.

[6] Xiong Y, Tang X F, Meng Q S, Zhang H F. Differential expression analysis of the broiler tracheal proteins responsible for the immune response and muscle contraction induced by high concentration of ammonia using iTRAQ-coupled 2D LC-MS/MS. Science China Life Sciences, 2016; 59: 1166–1176.

[7] Soliman E S, Moawed S A, Hassan R A. Influence of microclimatic ammonia levels on productive performance of different broilers' breeds estimated with univariate and multivariate approaches. Veterinary World, 2017; 10(8): 880–887.

[8] Tao Z Y, Xu W J, Zhu C H, Zhang S J, Shi Z H, Song W T, et al. Effects of ammonia on intestinal microflora and productive performance of laying ducks. Poultry Science, 2019; 98: 1947–1959.

[9] Bai Y, Li Y, Wang X X, Xie J J, Li C. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmospheric Pollution Research, 2016; 7(3): 557–566.

[10] Lim Y, Moon Y-S, Kim T-W. Artificial neural network approach for prediction of ammonia emission from field-applied manure and relative significance assessment of ammonia emission factors. European Journal of Agronomy, 2007; 26(4): 425–434.

[11] Qiao J F, Quan L M, Yang C L. Design of modeling error PDF based fuzzy neural network for effluent ammonia nitrogen prediction. Applied Soft Computing, 2020; 91: 106239.

[12] Xie Q, Ni J, Su Z. A prediction model of ammonia emission from a fattening pig room based on the indoor concentration using adaptive neuro fuzzy inference system. Journal of Hazardous Materials, 2017; 325: 301–309.

[13] Stamenkovic L J, Antanasijevic D Z, Ristic M D J, Peric-Grujic A A, Pocajt V. Modeling of methane emissions using artificial neural network approach. Journal of the Serbian Chemical Society, 2015; 80(3): 421–433.

[14] Yu H H, Chen Y Y, Hassan S G, Li D L. Prediction of the temperature in a Chinese solar greenhouse based on LSSVM optimized by improved PSO. Computers and Electronics in Agriculture, 2016; 122: 94–102.

[15] Barzegar R, Fijani E, Moghaddam A A, Tziritis E. Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models. Science of The Total Environment, 2017; 599-600: 20–31.

[16] Qiu X H, Zhang L, Nagaratnam Suganthan P, Amaratunga G A J. Oblique random forest ensemble via Least Square Estimation for time series forecasting. Information Sciences, 2017; 420: 249–262.

[17] Rubal, Kumar D. Evolving differential evolution method with random forest for prediction of air pollution. Procedia Computer Science, 2018; 132: 824–833.

[18] Wen L, Yuan X Y. Forecasting CO_2 emissions in China's commercial department, through BP neural network based on random forest and PSO. Science of The Total Environment, 2020; 718: 137194.

[19] Zhu K H, Wu S Y, Li Q. Prediction model for piggery ammonia concentration based on genetic algorithm and optimized BP neural network. Metallurgical and Mining Industry, 2015; 11: 6–12.

[20] Li R, Nielsen P V, Bjerg B, Zhang G Q. Summary of best guidelines and validation of CFD modeling in livestock buildings to ensure prediction quality. Computers and Electronics in Agriculture, 2016; 121: 180–190.

[21] Wang J L, Xu C Q, Zhang J, Zhong R. Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems, 2022; 62: 738–752.

[22] Sun L, Wang L Y, Ding W P, Qian Y H, Xu J C. Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets. IEEE Transactions on Fuzzy Systems, 2021; 29(1): 19–33.

[23] Qian W B, Huang J T, Xu F K, Shu W H, Ding W P. A survey on multi-label feature selection from perspectives of label fusion. Information Fusion, 100: 101948. doi: 10.1016/j.inffus.2023.101948.

[24] Sun L, Yin T Y, Ding W P, Qian Y H, Xu J C. Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multi-label neighborhood decision systems. Information Sciences, 2020; 537: 401–424.

[25] Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the tenth national conference on Artificial Intelligence, 1992; pp.129–134. doi: 10.5555/1867135.1867155.

[26] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: Machine Learning: ECML-94. ECML 1994 Lecture Notes in Computer Science, 1994; 784: 171–182. doi: 10.1007/3-540-57868-4_57.

[27] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical Review E, 2004; 69: 066138.

[28] Breiman L. Random forests. Machine Learning, 2001; 45: 5–32.

[29] Zhang J W, Song W L, Jiang B, Li M B. Measurement of lumber moisture content based on PCA and GS-SVM. Journal of Forestry Research, 2018; 29: 557–574.

[30] Probst P, Boulesteix A-L. To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research, 2017; 18(1): 6673–6690.

[31] Ortin F, Facundo G, Garcia M. Analyzing syntactic constructs of Java programs with machine learning. Expert Systems with Applications, 2023; 215: 119398.