# Short-Term Prediction of Ammonia Levels In Geese Houses Based On Combined Feature Selector and Random Forest

**Jiande Huang**
Zhongkai University of Agriculture and Engineering
**Jianjun Guo**
Zhongkai University of Agriculture and Engineering
**Huilin Wu**
National S&T Innovation Center for Modern Agricultural Industry
**Xinglong Zhang**
National S&T Innovation Center for Modern Agricultural Industry
**Shuangyin Liu**
Zhongkai University of Agriculture and Engineering
**Shahbaz Gul Hassan** ( mhasan387@zkhu.edu.cn )
Zhongkai University of Agriculture and Engineering

**Article**

**Keywords:**

**Posted Date:** March 21st, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1449095/v1

# Short-Term Prediction of Ammonia Levels In Geese Houses Based On Combined Feature Selector and Random Forest

**Jiande Huang [1,2,3,4,5,†], Jianjun Guo [1,2,3,4,5,†], Huilin Wu [6], Xinglong Zhang [6], Shuangyin Liu [1,2,3,4,5,*] and Shahbaz Gul Hassan [1,2,3,4,5,*]**

[1]College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[2]Smart Agriculture Engineering Technology Research Center of Guangdong Higher Education Institues, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[3]Guangzhou Key Laboratory of Agricultural Products Quality & Safety Traceability Information Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[4]Academy of Intelligent Agricultural Engineering Innovations, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[5]Guangdong Provincial Agricultural Products Safety Big Data Engineering Technology Research Center, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[6]National S&T Innovation Center for Modern Agricultural Industry, Guangzhou 510520, China
[†]Contributed equally to this work.
*Correspondence: Shuangyin Liu (hdlsyxlq@126.com); Shahbaz Gul Hassan (mhasan387@zkhu.edu.cn)

## 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. To improve the prediction accuracy of $NH_3$, we propose a novel forecasting model by the combination of feature selector (CFS) and random forest (RF). The developed model integrated two modules. First, combining mutual information (MI) and relief-F, we propose CFS quantify the importance values of each feature and eliminate the low-relation or unrelated features. Second, we built a random forest model and used the K-fold cross-validation grid search algorithm (CVGS) to obtain the RF hyper-parameters 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, 2.88%, respectively. The $NH_3$ prediction model based on CFS-GS-RF exhibited 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.

## Introduction

China is the largest geese producer in the world. According to the data published by the China Statistics Bureau in 2019 National Economic and Social Development Statistical Bulletin, the output of poultry meat and poultry eggs in China were 22.39 million tons and 33.09 million tons,

respectively. The total output value of waterfowl exceeds 160 billion 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 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 the health of poultry [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 the immune function, health, and growth capability of poultry [3,6,7], gives rise to a variety of diseases and leads to economic losses. Currently, research on the influence of $NH_3$ on poultry has focused on laying and broilers hens, while the effects of $NH_3$ on geese have not been widely discussed [8]. However, meat farming and laying geese in China are now gradually transforming from the outdoor or semi-outdoor to the indoor type. Therefore, the hazards associated with $NH_3$ are expected to pose a severe problem for meat geese production. Due to the severe implications of NH3 on poultry health, there is a need to provide a beneficial and stable environment for poultry. Which is suitable for the complex nature of the NH3 for its modelling and prediction.

In recent years, several intelligent algorithms have been proposed for the prediction of $NH_3$ levels [9–13]. Artificial neural network (ANN), support vector machine (SVR), and decision tree (DT) are useful tools and have been widely used for solving complex prediction problems. However, DT often faces the over-fitting issue, so that 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. Instead, ANN has a strong generalization ability. Still, it can easily fall into a partial solution [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, and has become a hotspot in the prediction field 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. [12] used an adaptive neuron fuzzy inference system (ANFIS) to predict $NH_3$ using indoor relative humidity, indoor temperature, pig temperature and other indicators. [19] predicted $NH_3$ based on genetic algorithm (GA) and optimized backpropagation neural 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 of the 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 the vast data directly in intelligent models without feature selection not only increases the training time but also increases 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 with respect to readability and understandability and will not alter the primitive feature data [21]. Recently, mutual information-based algorithms (MI) play an increasingly significant role in data mining and machine learning. These methods have good non-linear and linear processing capability 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 [25] was initially used for two-category problems. To adapt to multiple category problems, [26] proposed the relief-F algorithm based on Relief to deal with multilabel data and regression problem, which is a widely adaptable filter-based feature evaluating algorithm [21]. In this study, the fusion of mutual information and relief-F was proposed to improve the capability of feature selection and bring a feature selection that is more accurate.

In this paper, we investigate the $NH_3$ prediction. Our target is to accurately forecast $NH_3$ levels using the data from an intelligent geese house Internet of Things system. To overcome this challenging problem, we design an RF underpinned framework that effectively predicts the $NH_3$ level. Although RF is a promising approach, two challenges must be addressed. The first is that unrelated and redundant features give rise to a high cost of the RF training process and decrease the prediction

accuracy. The second is the tuning of the parameters. There are three hyper-parameters: number of a tree, 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.

To address the challenges mentioned above, this paper combines the feature selector with the RF and designs a new hybrid prediction model for predicting $NH_3$ in geese houses. First, the combined feature selector (CFS) that combines the MI and relief-F evaluates the importance of prediction indicators. Then, the parameter *M* is controlled to eliminate the features that have low feature importance. 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 the selection of the prediction indicators in this study. Comparison with other models horizontally and vertically using empirical results shows that the prediction results of this model can provide technical support for the precise management and control of intensive aquaculture environment of waterfowl.

# Materials and Methods

## 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, we housed animals (stone goose) in the area including poultry house (25 × 16 $m^2$), 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(fig. 1).

Because geese house environment parameters are mainly affected by physical and chemical factors. We developed an IoT system(Fig2) to monitor temperature, humidity, carbon dioxide(CO2), total suspended particulates (TPS), NH3 of the houses and temperature, humidity, atmospheric pressure, and solar radiation of the surrounding environment.

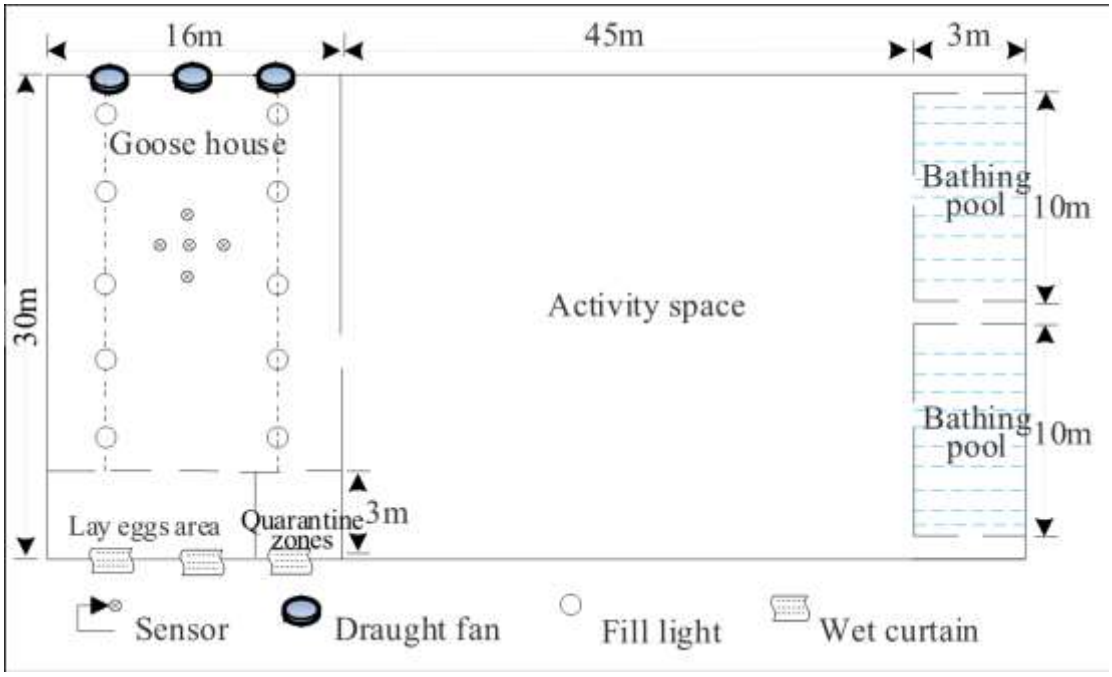



**Fig.1 Schematic diagram of the breeding environment monitoring based on the Internet of Things..**

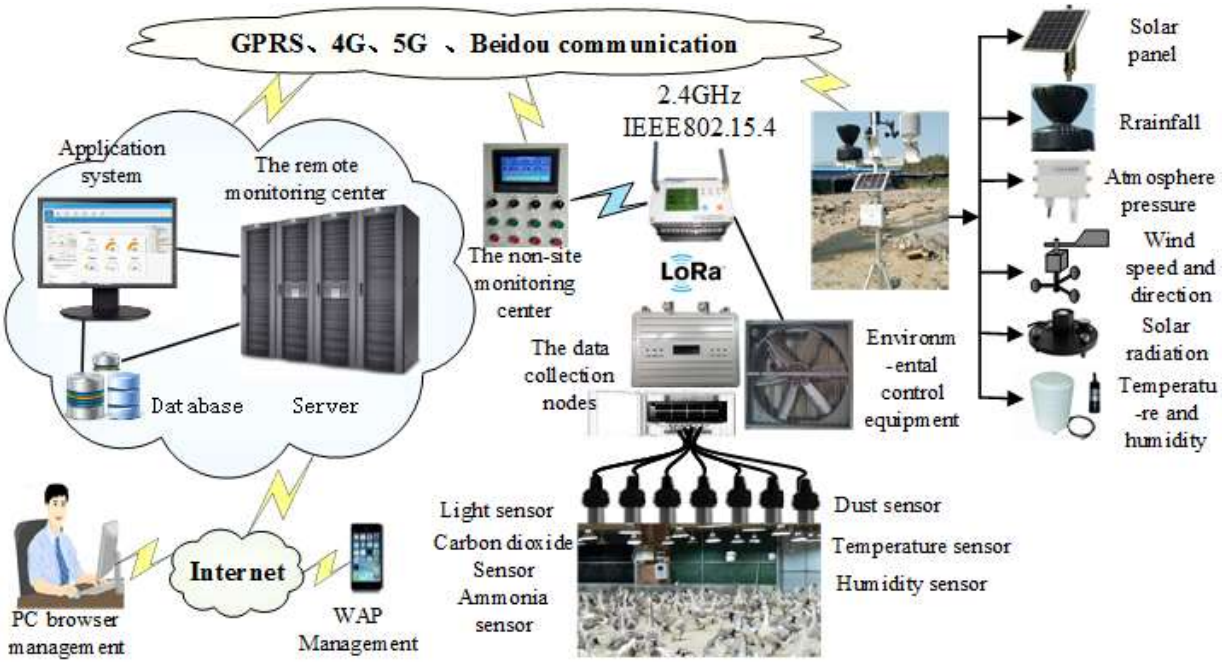



**Fig2. Topology structure diagram of the platform.**

## 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 preserves. The parameter *M* is the threshold. We can eliminate the feature with low feature importance by controlling *M*. In this section, we will introduce this module. Fig. 3 shows the

processes of CFS.

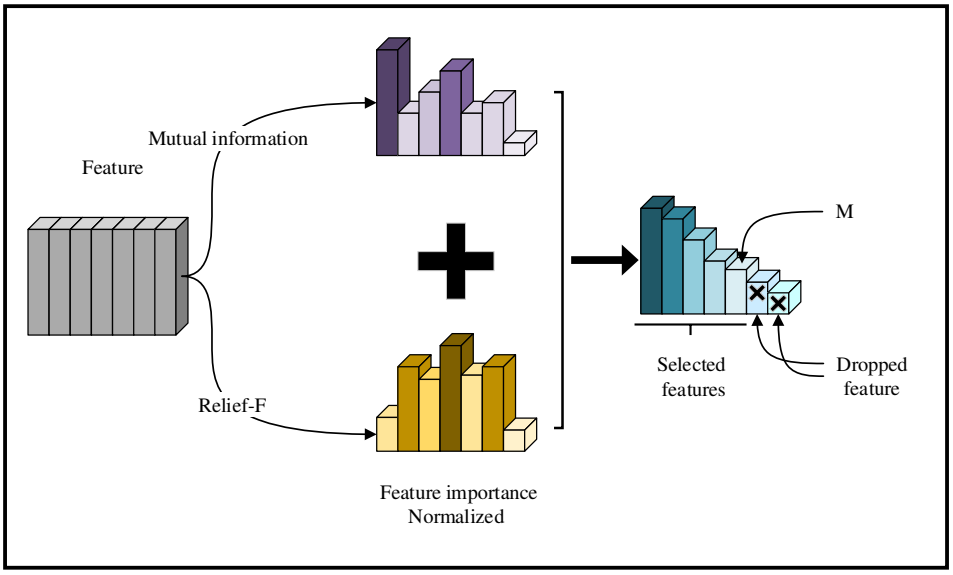



Fig.3. The structure of CFS.

**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 is later extended by Kononeill 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.

Relief-F algorithm estimates feature weight according to the degree of distinguishing samples that are close to each other based on the value of the feature. For that purpose, the relief-F algorithm randomly selects a simple $X_i$ ($X_i$ has classified $p$) from training set $D$ which has $|y|$ class. Then searches for $k$ of its nearest neighbours from the same class, called near-hit $X_{i,nh}$, and also $k$ nearest neighbours from each of the different classes, called near-miss $X_{i,j,nm}(j=1,2,\ldots, |y|;j\neq p)$, then the weight of feature $L$ can be computed as follows:

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

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 $j$, $diff()$ defined as:

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

**Mutual Information Estimator**

Among the estimates of independence between random variables, MI is selected based on its information theory background. *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:

$$MI(X,Y) = \int_Y \int_X P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) dxdy \quad (3)$$

where *P(x,y)* is the probability density function of random variable *Z=(X,Y)*, *P(x)* and *P(y)* represents the marginal probability density function of *X* and *Y*, respectively. In fact, since we usually do not know *P(x,y)* in advance, some methods should be used to estimate *MI(X,Y)*. K-nearest neighbour is a non-parametric method that has been confirmed as a useful method in MI estimation [27]. For K-nearest neighbour, we use Euclidean distance as a distance metric, and the maximum norm for the space *Z=(X,Y)* written as:

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

Let $\kappa(i)/2$ express the distance from $z_i$ to its $k^{th}$ neighbour, and $\kappa_x(i)/2$ and $\kappa_y(i)$ express the distance between the same points projected into the $X$ and $Y$ subspaces. It is clear that:

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

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 $\eta y(i)$ is the number of points for which the distance from $yi$ is strictly less than $\kappa(i)$. We note that $\kappa(i)$ not a fixed value, and $\eta_x(i)$ and $\eta_y(i)$ is also not fixed. We denote by $\langle \cdots \rangle$ both all $i \epsilon (1,...,N)$ 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 neighbour is then:

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

In equation [], $\psi(\,)$ is the di-gamma function and satisfies the recursion $\psi(x+1)=\psi(x)+1/x$ and $\psi(1)=-C$, where $C=0.5772156..$ is the Euler-Mascheroni constant. If MI is equal to zero, the two random variables are independent and higher MI values mean higher dependency.

**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 construct the RF is a non-parametric model. According to the complexity of the input data, the tree grows in the process of the learning. 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 we denote by $X=\{x_1,x_2,x_3,...,x_n\}$ the input vector with n features, $Y$ the output scalar and $D_m$ the training set with $m$ observations which can be written as follows:

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

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 of the variables. The splitting procedure begins at the root node, and each node uses its own 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 $\hat{T}(D_m, X)$ generated over $D_m$.

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 subset. An RF is an ensemble of $P$ DT $\hat{T}(D_m^1, X)$, $\hat{T}(D_m^2, X)$,..., $\hat{T}(D_m^p, X)$. Here, $(D_m^1, D_m^2, \ldots, D_m^3)$ are the bootstrap sample obtained by random sampling of $m$ observations with replacement from $D_m$, where each observation has $1/m$ probability of being drawn. This sample process is known as bootstrap resampling. During the splitting of each node, only a small part of $n$ feature are randomly selected instead of all features; this is known as random feature selection. The ensemble learning result $P$ output $\hat{Y}_1 = \hat{T}(D_m^1, X)$, $\hat{Y}_2 = \hat{T}(D_m^2, X)$,..., $\hat{Y}_p = \hat{T}(D_m^p, X)$. Then, the final estimation output $\hat{Y}$ is the average of $P$ output, that is described as follows:

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

where $\hat{Y}_i$ is the output of $i^{th}$ DT, $i = 1,2,\ldots,P$. The framework of RF regression is illustrated in Fig.4, and its training process can be summarized as:

1. Obtain bootstrap samples from the training data set by bootstrap resampling
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.
3. Repeat steps 1 and 2 until P DT tree is generated.
4. Aggregating the output of P tress by an average method to forecast unknown data.

*2.4 Cross-validation grid search*

According to the previous section, we notice that the RF algorithm has two important parameters:

- $P$: the number of decision trees that are the base estimators of RF.
- $mtry$: 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, $mtry$ also a sensitive parameter, and increasing $mtry$ can improves 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 algorithm that is currently the most widely used method for parameter optimization is a highly suitable model with fewer hyper-parameters. GS exhaustively generates candidates from a grid of parameter values specified by the user parameters, and then trains each candidate set of parameters and marks the score of the model, finally obtaining the optimal combination of parameters. GS optimizes all of the parameters of the model 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, because in this case, the model that merely repeats the samples that it has just learned will have a high score. To avoid this, we combine K-fold cross-validation (CV) and grid search (GS). 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:

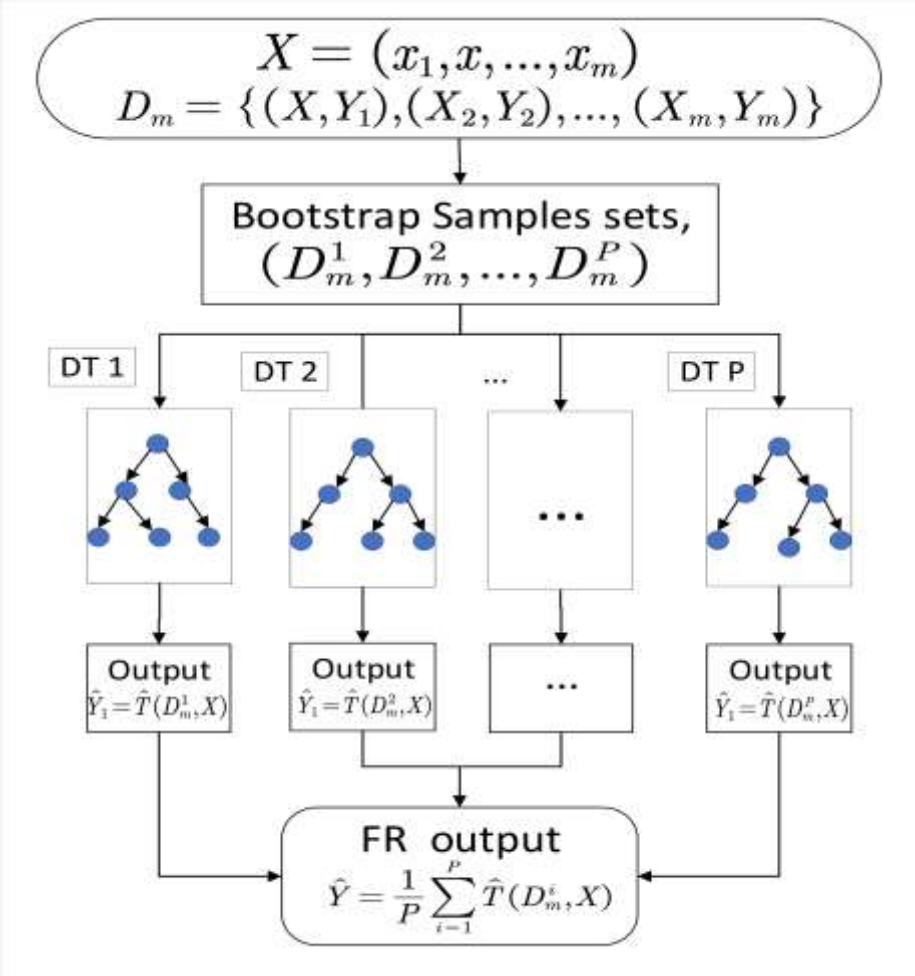



**Fig.4. The framework of random forest regression.**

- 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 the step 4. Then, the parameter combination with the lowest MSE is identified as the optimal.

## Hybrid prediction model based on CFS-CVGS-RF.

This paper proposes a CFS-CVGS-RF model to predict $NH_3$ in a geese house. The methodology for conducting this model is shown in Fig.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 will also decrease the performance of prediction. To address this problem, we use the normalization that can eliminate the difference between the data units and dimensions and facilitates the study of the correlation between environmental factors. The normalized process method is described by:

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

where $y_i'$ are the normalized data, $y_{max}$ and $y_{min}$ are the max and min values of the original data, and $y_i$ and $\overline{y}$ are the original data and its 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. The final feature importance is the sum of two different dimensions features important values after normalization. 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 size of the random feature subset $mtry$. To find the optimal values of these two parameters, we adopt the CVGS method. In the CVGS process, we first establish the grid coordinates of the parameters. In this paper, combining the RF related literature [29] and the experimental parameters wave motion, we 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 [30]. After $k$ parallel operations, each parameters' combinations have $k$ MSE. According to the mean of each calculation results, we select the parameter combination with minimum average MSE as the optimal parameters and establish the RF model using these values.
- Step 4: Result output denormalization. Denormalize the output to obtain the results in the normal dimension. The denormalize process is described by:

$$y_i = y_i'^{*}(y_{\max} - y_{\min}) + \overline{y} \quad (11)$$

where $y_i$ is the denormalized data, $y_{max}$ and $y_{min}$ are max and min value of original data, and $y_i'$ and $\overline{y}$ are the original data and its mean.

In this article, a simulation model was built in order to validate the performance of the proposed method. The $NH_3$ was tested in a high-density geese culture farm from September 10st to September 27st, 2019 in Sanwei city, Guangdong province. Fig.6 shown the feature importance computed by CFS for each factor, and the performances of different threshold M are given in Table 1. The *P* and *mtry* combinational parameters of optimal CVGS-RF model are 500 and 3, respectively, and the MSE values for different parameter combinations are shown in Fig.7. It is observed that after growing 100 trees, the MSE value decreases very slowly and apparently converges after 500 trees, with mtry=3 resulting in the lowest MSE. Fig.8 shows the $NH_3$ series prediction result obtained by the combined model based on CFS-CVGS-RF. The performance estimation statistics of the testing are given in Table 2. Fig.9 shows the residual distribution of the proposed model. The results of $NH_3$ prediction demonstrate the robustness and effectiveness of CFS-CVGS-RF.

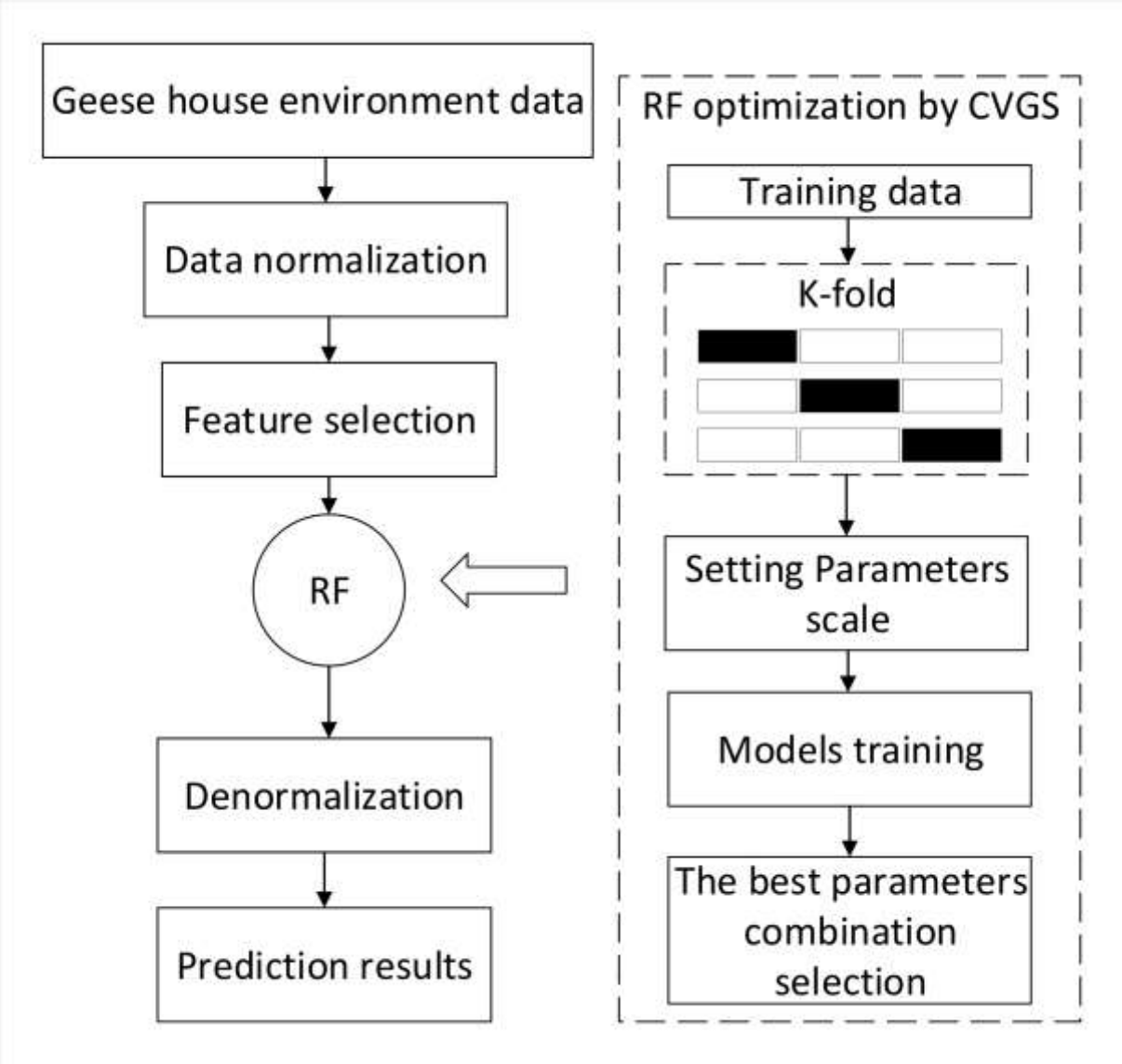



**Fig.5. The schematic of the proposed methodology.**

## Results and discussions

### Experimental environment and settings

The geese house environment data considered in this investigation include the data obtained at the intervals of twenty minutes from September 10 to September 27, 2019. Seventy-two sets of data collected per day yield a total of 1296 observations samples. For a 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.

To confirm the superiority of the framework constructed in this research, we designed a script with Python 2.7vision following the system framework described in Section 3. The script is executed on

the Win10 operation system with Intel Core i5, 4 GB RAM, and 500G hard disk. We use the optimal parameter combination (*P*=500, *mtry*=3) for the prediction model of $NH_3$ by the CVGS algorithm. According to Fig.6 and Table2, 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 paper.

**Performance criteria**

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

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y_i}\right)^2 \quad (12)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y_i}\right)^2} \quad (13)$$

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

**Results and analysis**

Following the method introduced in this article, we use 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. Fig.6 shows the normalized results of relief-F and MI, and it is observed that outdoor humidity has the highest relation and solar radiation has the lowest relation with $NH_3$. To find the optimal input feature combination, the threshold *M* was changed from 0 to 2. The sensitivity test results of *M* were recorded in Table 1, and it is observed that the combination of outdoor humidity, TPS, CO2 and indoor temperature features was the optimal input combination. Fig.7 illuminates the change of 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 apparently converges after 500 trees, and mtry=3 results in the lowest MSE.

To analyze and compare prediction performance, two types of comparison were designed: (1) horizontal comparison between the model with CFS and the model without CFS and (2) vertical comparison between the models used in this paper with other parallel models. The horizontal comparison includes the CVGS-RF model, and vertical comparison consists of the benchmarks used

in this DT, support vector machine and back propagation neural network (BPNN). To verify the performance forecasted by the models in this paper, these models used data sets and predict the $NH_3$ content of the last 72 test sets corresponding to the last 24 hours. Fig.8 and fig.9 show the prediction curves and the error bar plot.

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 (fig.10) 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.

**Table.1. The sumulation results of CFS.**

| M | Accuracy | Running time(second) | 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<br>Indoor humidity |
| 1.6 | 78.5% | 5.194 | TPS<br>Indoor temperature |
| 1.8 | 16.1% | 4.945 | $CO_2$ |
| 2.0 | NaN | NaN | Outdoor humidity |

For a more accurate comparison of the performance of five models, this article computes MSE, RMSE and MaxMAPE of those models for which the details are shown in Table 2. The MSE, RMSE and MAPE of CFS-GS-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. Fig.8 describes the prediction residual distribution condition of five models, and it is observed that CFS-GS-RF has fewer outliers and overall error closer to zero for the residual error compared with other models. This mean that CFS-GS-RF has more stability for forecasting results and is more suitable for the prediction of $NH_3$.

## Conclusions

This paper proposes a novel $NH_3$ prediction hybrid model (CFS-CVGS-RF). The CFS-CVGS-RF

model consists of a combination of four methods: relief-F, mutual information, K-fold cross-validation grid search optimization algorithm and random forest. Using actual experimental geese house environment data from a monitored aquaculture factory farm in Sanwei city, China. Results clearly show 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 prediction performance of the model by screening the unrelated or low-relation features.

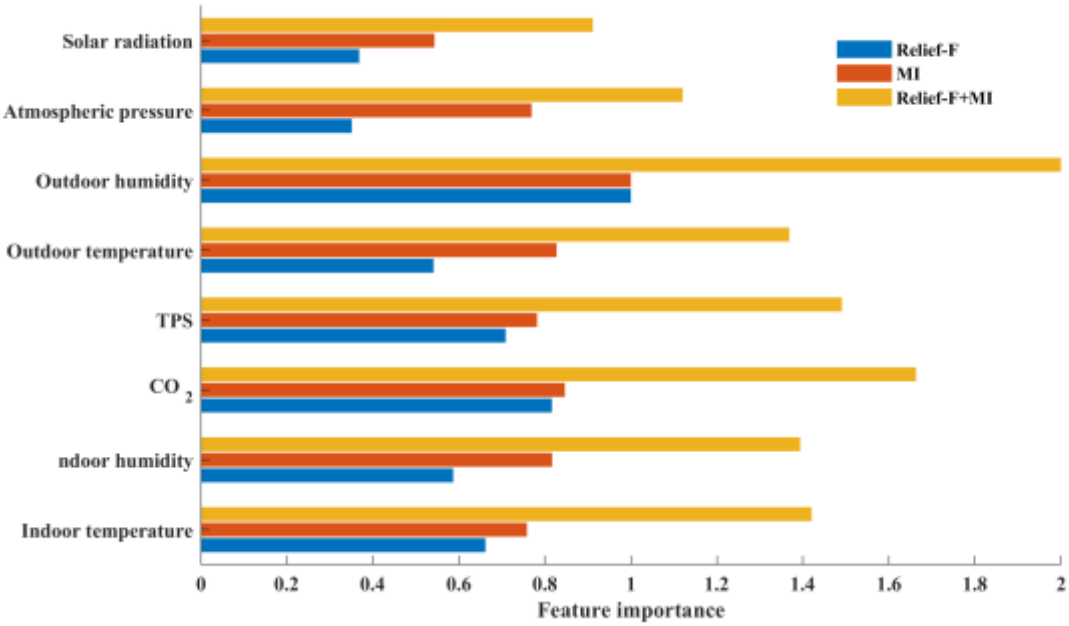



**Fig.6. Feature importance computed by Relief-F and MI.**

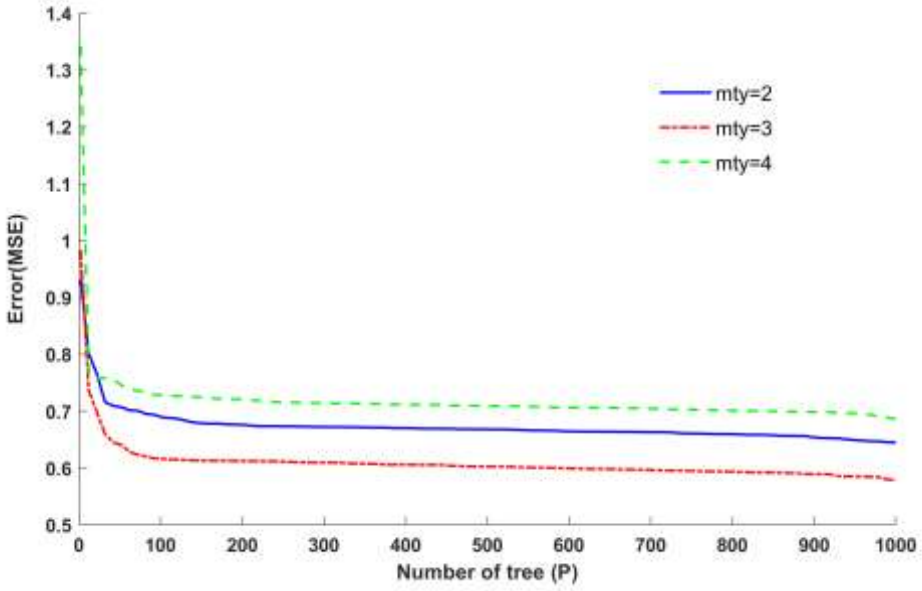



**Fig.7. The results of K-fold cross-validation grid search.**

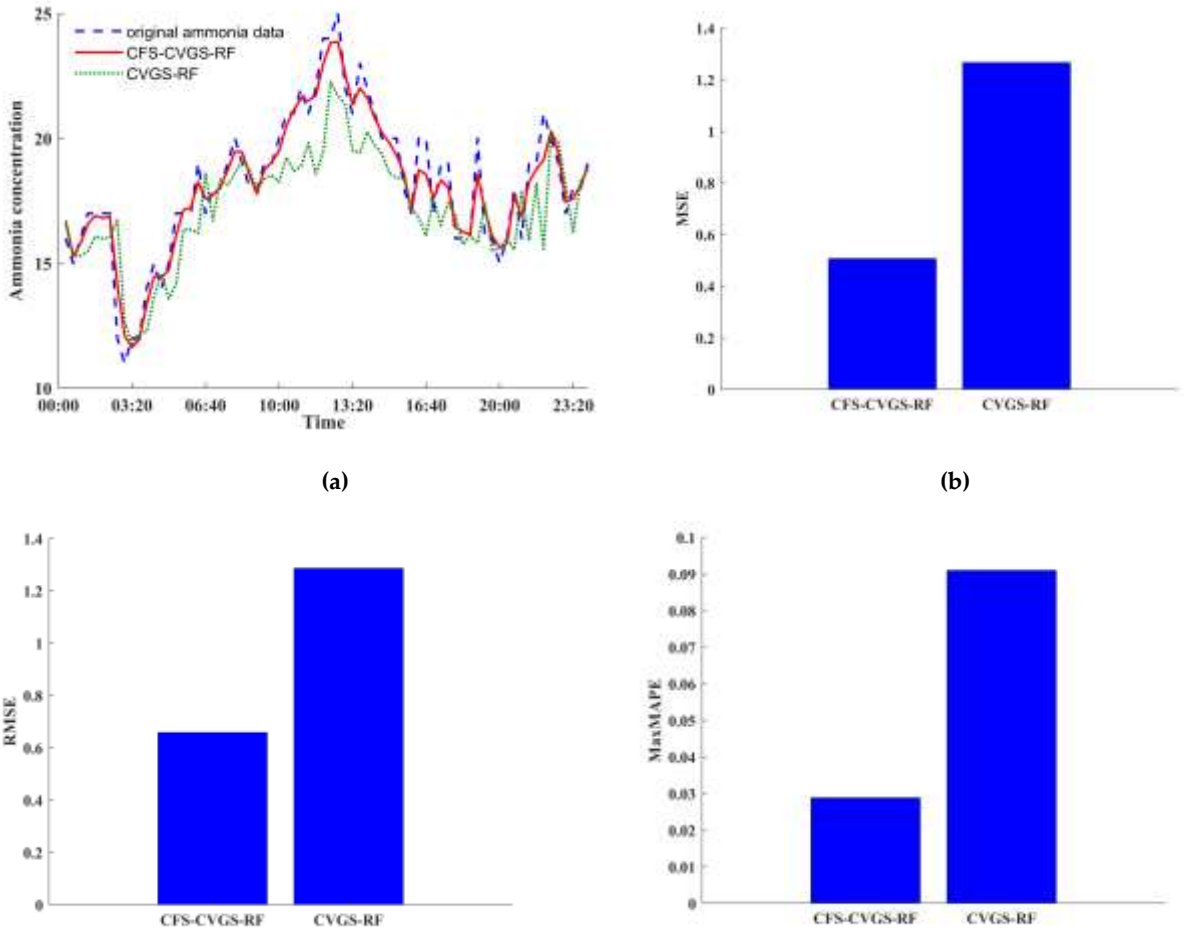
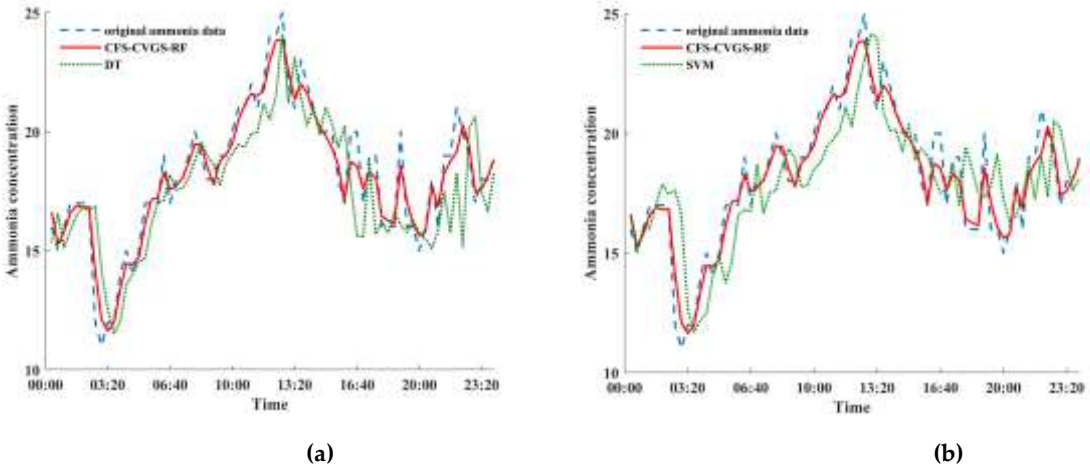



**Fig.8. Horizontal comparison results.**



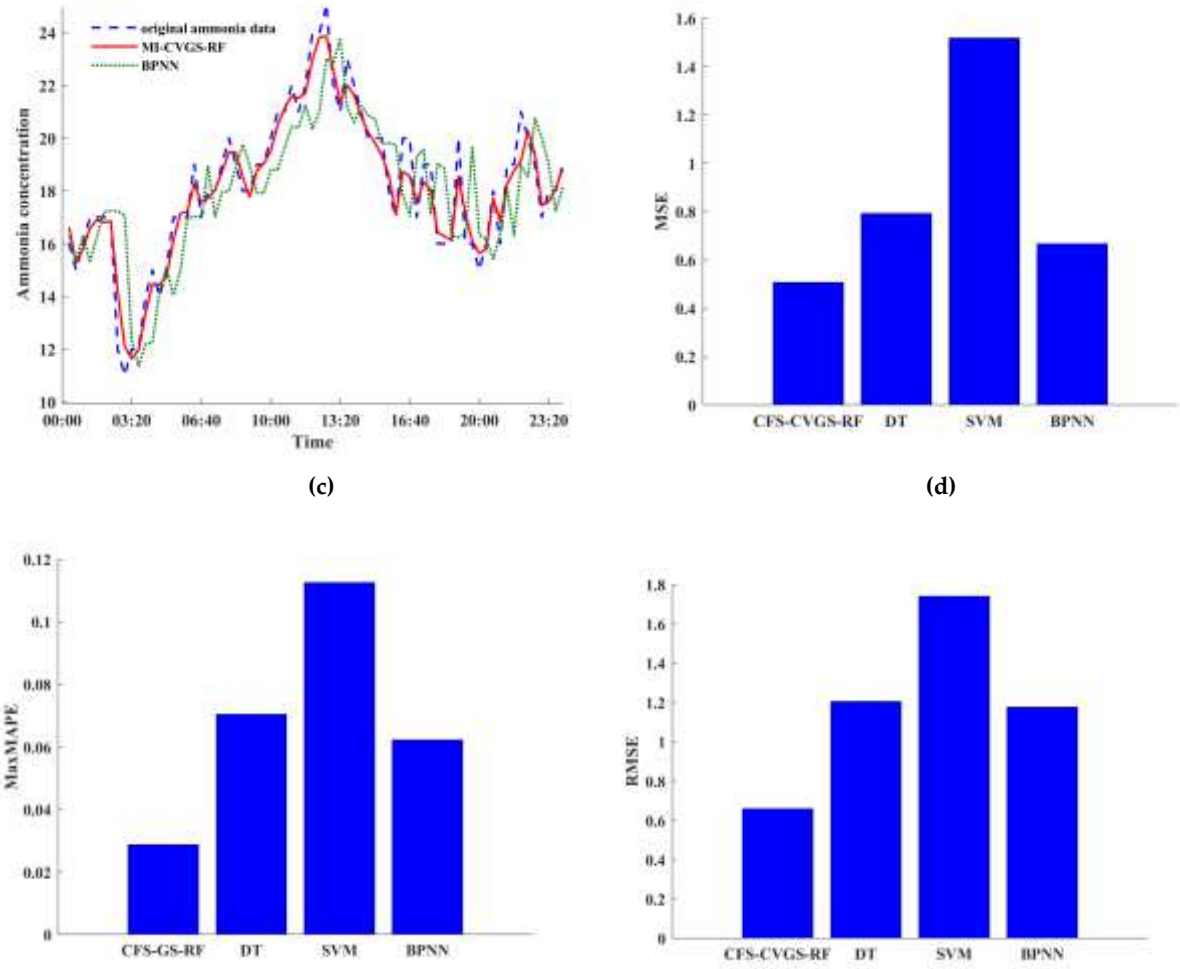



**Fig.9. The vertical comparison results.**

**Table.2. 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 |

Our 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 the use of other ensemble strategies instead of a simple averaging method to improve the RF, such as weighted averaging, weighted voting.

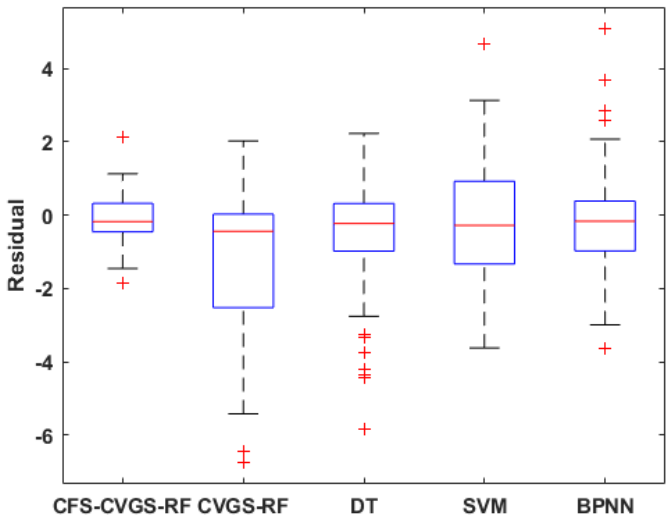



**Fig.10. Boxplot of Residual error in different models.**

## Data availability

Data or code presented in this study are available on request from the corresponding author.

## Reference

1. CSB, C. S. B. Statistical Communique of the People's Republic of China on the 2019 National Economic and Social Development. (2020).
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. *Biol. Conserv.* **221**, 1–9 (2018).
3. Wei, F. X. *et al.* Ammonia concentration and relative humidity in poultry houses affect the immune response of broilers. *Genet Mol Res* **14**, 3160–3169 (2015).
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. *J. Agromedicine* **19**, 162–170 (2014).
5. Nemer, M. *et al.* Airway inflammation and ammonia exposure among female

437 Palestinian hairdressers: a cross-sectional study. *Occup. Environ. Med.* **72**, 428–434
438 (2015).
439 6. Xiong, Y., Tang, X., Meng, Q. & Zhang, H. Differential expression analysis of the
440 broiler tracheal proteins responsible for the immune response and muscle contraction
441 induced by high concentration of ammonia using iTRAQ-coupled 2D LC-MS/MS. *Sci.*
442 *China Life Sci.* **59**, 1166–1176 (2016).
443 7. Soliman, E. S., Moawed, S. A. & Hassan, R. A. Influence of microclimatic ammonia
444 levels on productive performance of different broilers' breeds estimated with
445 univariate and multivariate approaches. *Vet. world* **10**, 880 (2017).
446 8. Tao, Z. *et al.* Effects of ammonia on intestinal microflora and productive performance
447 of laying ducks. *Poult. Sci.* **98**, 1947–1959 (2019).
448 9. Bai, Y., Li, Y., Wang, X., Xie, J. & Li, C. Air pollutants concentrations forecasting using
449 back propagation neural network based on wavelet decomposition with
450 meteorological conditions. *Atmos. Pollut. Res.* **7**, 557–566 (2016).
451 10. Lim, Y., Moon, Y.-S. & Kim, T.-W. Artificial neural network approach for prediction of
452 ammonia emission from field-applied manure and relative significance assessment of
453 ammonia emission factors. *Eur. J. Agron.* **26**, 425–434 (2007).
454 11. Qiao, J., Quan, L. & Yang, C. Design of modeling error PDF based fuzzy neural
455 network for effluent ammonia nitrogen prediction. *Appl. Soft Comput.* 106239 (2020).
456 12. Xie, Q., Ni, J. & Su, Z. A prediction model of ammonia emission from a fattening pig
457 room based on the indoor concentration using adaptive neuro fuzzy inference system.
458 *J. Hazard. Mater.* **325**, 301–309 (2017).
459 13. Stamenkovic, L. J., Antanasijevic, D. Z., Ristic, M. D. J., Peric-Grujic, A. A. & Pocajt, V.

460 V. Modeling of methane emissions using artificial neural network approach. *J. Serbian*
461 *Chem. Soc.* **80**, 421–433 (2015).
462 14. Yu, H., Chen, Y., Hassan, S. G. & Li, D. Prediction of the temperature in a Chinese
463 solar greenhouse based on LSSVM optimized by improved PSO. *Comput. Electron.*
464 *Agric.* **122**, 94–102 (2016).
465 15. Barzegar, R., Fijani, E., Moghaddam, A. A. & Tziritis, E. Forecasting of groundwater
466 level fluctuations using ensemble hybrid multi-wavelet neural network-based models.
467 *Sci. Total Environ.* **599**, 20–31 (2017).
468 16. Qiu, X., Zhang, L., Nagaratnam Suganthan, P. & Amaratunga, G. A. J. Oblique random
469 forest ensemble via Least Square Estimation for time series forecasting. *Inf. Sci. (Ny).*
470 (2017) doi:10.1016/j.ins.2017.08.060.
471 17. Rubal & Kumar, D. Evolving Differential evolution method with random forest for
472 prediction of Air Pollution. in *Procedia Computer Science* (2018).
473 doi:10.1016/j.procs.2018.05.094.
474 18. Wen, L. & Yuan, X. Forecasting CO2 emissions in Chinas commercial department,
475 through BP neural network based on random forest and PSO. *Sci. Total Environ.* **718**,
476 137194 (2020).
477 19. Zhu, K., Wu, S. & Li, Q. Prediction Model for Piggery Ammonia Concentration Based
478 on Genetic Algorithm and Optimized BP Neural Network. *Metall. Min. Ind.* (2015).
479 20. Rong, L., Nielsen, P. V, Bjerg, B. & Zhang, G. Summary of best guidelines and
480 validation of CFD modeling in livestock buildings to ensure prediction quality.
481 *Comput. Electron. Agric.* **121**, 180–190 (2016).
482 21. Wang, K., Xu, C., Zhang, Y., Guo, S. & Zomaya, A. Y. Robust big data analytics for

electricity price forecasting in the smart grid. *IEEE Trans. Big Data* **5**, 34–45 (2017).

22. Sun, L., Wang, L., Ding, W., Qian, Y. & Xu, J. Feature Selection Using Fuzzy Neighborhood Entropy-Based Uncertainty Measures for Fuzzy Neighborhood Multigranulation Rough Sets. *IEEE Trans. Fuzzy Syst.* (2020) doi:10.1109/tfuzz.2020.2989098.

23. Gonzalez-Lopez, J., Ventura, S. & Cano, A. Distributed multi-label feature selection using individual mutual information measures. *Knowledge-Based Syst.* (2020) doi:10.1016/j.knosys.2019.105052.

24. Sun, L., Yin, T., Ding, W., Qian, Y. & Xu, J. Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems. *Inf. Sci. (Ny).* (2020) doi:10.1016/j.ins.2020.05.102.

25. Kira, K. & Rendell, L. A. Feature selection problem: traditional methods and a new algorithm. in *Proceedings Tenth National Conference on Artificial Intelligence* (1992).

26. Kononenko, I. Estimating attributes: Analysis and extensions of RELIEF. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (1994). doi:10.1007/3-540-57868-4_57.

27. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E* **69**, 66138 (2004).

28. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

29. Probst, P. & Boulesteix, A.-L. To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.* **18**, 6673–6690 (2017).

30. Witten & Frank, I. H. Data mining. *Pract. Mach. Learn. Tools Tech. with Java Implementations* **13**, 1 (2005).

## Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 61871475,61471-131,61571444, in part by the special project of laboratory construction of Guangzhou Innovation Platform Construction Plan under Grant 201905010006, Guangzhou Innovation Plantform Construction Plan under Grant 2017B0101260016, foundation for High-level Talents in Higher Education of Guangdong Province under Grant 2017GCZX00014, 2016K-ZDXM0013,2017KTSCX094,2018LM2168,and Beijing Natural Science Foundation under Grant 4182023.

## Contributions

All authors contributed extensively to this manuscript. J.H. and S.H. contributed to conceptualization, methodology, writing original draft preparation, software, visualization and supervision. J.G. performed the formal analysis, H.W. and X.Z. contributed to the investigation, S.L. provided the farm and data resources for this research. All authors have read and agreed to the published version of the manuscript.

## Competing interests

The authors declare no competing interests.