Estimating the air exchange rates in naturally ventilated cattle houses using Bayesian-optimized GBDT

Luyu Ding[1,2,3], Lei E[1,4], Yang Lyu[1,2], Chunxia Yao[1,2], Qifeng Li[1,2,3*], Shiwei Huang[4], Weihong Ma[1,2,3], Ligen Yu[1,2,3], Ronghua Gao[1,2,3]

(1. *Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China*;
2. *National Innovation Center of Digital Technology in Animal Husbandry, Beijing 100097, China*;
3. *Beijing Technology Innovation Strategic Alliance for Intelligence Internet of Things Industry in Agriculture, Beijing 100097, China*;
4. *Department of Agricultural Structure and Bioenvironmental Engineering, College of Water Resources and Civil Engineering, China Agricultural University, Beijing 100083, China*)

Abstract: It is challenging to estimate the air exchange rate (AER) dynamically in naturally ventilated livestock buildings such as dairy houses due to the influence of complex and variable outdoor environmental factors, large opening ratios, and the confusion of inflow and outflow at openings. This makes it difficult to efficiently regulate the opening ratio to meet the ventilation requirements in naturally ventilated livestock buildings. In this study, the air exchange rates of naturally ventilated cattle houses (NVCHs) in different seasons and opening ratios were obtained through field measurements and computational fluid dynamics (CFD) simulations. A fast and efficient machine learning framework was proposed and examined to predict AER based on the gradient boosting decision tree (GBDT) combined with Bayesian optimization. Compared with commonly used machine learning models such as multilayer perceptrons (MLPs) and support vector machines (SVMs), the proposed GBDT model has higher prediction accuracy and can avoid falling easily into local optima. Compared with the existing mechanical model based on the Bernoulli equation, the proposed GBDT model showed a slightly higher prediction than the mechanistic model and was much easier to use in AER estimation when inputting easily collected environmental factors in practical applications. Using Bayesian optimization could dramatically reduce the computing time when determining the optimal hyperparameter for establishing the GBDT model, dramatically saving on computing resources. Based on the Bayesian optimized GBDT model, the desirable opening ratio of the side curtain can be determined for automatically regulating the AER of cattle houses in future applications.

Keywords: natural ventilation, Bayesian, GBDT, air exchange rate, cattle house
DOI: 10.25165/j.ijabe.20231601.7309

Citation: Ding L Y, E L, Lyu Y, Yao C X, Li Q F, Huang S W, et al. Estimating the air exchange rates in naturally ventilated cattle houses using Bayesian-optimized GBDT. Int J Agric & Biol Eng, 2023; 16(1): 73–80.

1 Introduction

Natural ventilation has low energy costs and is preferred for cattle housing[1,2]. The estimation of the air exchange rate (AER) is the basis of environmental control and assessing the air pollutant emissions in livestock houses[3]. In a naturally ventilated cattle house (NVCH), side curtains are commonly used to adjust the opening ratio (the ratio of curtain opening height to the height of the sidewall vent) of the sidewall to improve the indoor climate and animal comfort[4]. The AER varies drastically under different opening ratios and outdoor environmental conditions[4]. The orifice equation method, also known as the theoretical modeling method based on the Bernoulli equation, is often applied to estimate the AER of naturally ventilated houses[5,6]. Theoretically, the orifice equation method can be used to control the ventilation system in NVCH. However, the assumption of uniformly distributed pressure and velocity is questionable in a large opening (opening ratio more than 10%) as the condition in cattle houses[2,7]. This relies on some empirical coefficients that vary greatly in different conditions[5,7]. The authors believe that the orifice equation method is more suitable for ventilation design rather than real-time regulation of AERs in NVCH.

Alternatively, methods based on the mass balance principle or the energy balance principle, including the heat balance method, H_2O balance method, CO_2 balance method, and tracer gas method, are widely used in field measurements to monitor AERs in naturally ventilated livestock houses[1,8-11]. These indirect methods do not relate AERs to the opening ratio, which instructs the execution of the ventilation control system. Moreover, poor air mixing or large variations in spatial concentration in NVCH makes it difficult to arrange the sensors in reasonable positions[7,12]. These issues make it difficult to use real-time regulation for the ventilation system of NVCHs. More knowledge or methods

Received date: 2021-12-30 **Accepted date:** 2022-08-02
Biographies: Luyu Ding, PhD, Senior Engineer/Associate Research Fellow, research interest: gas emission and environmental control in animal production systems, Email: dingly@nercita.org.cn; **Lei E**, Master, research interest: precision livestock farming, Email: rayer97@163.com; **Yang Lyu**, Bachelor, research interest: precision livestock farming, Email: lvyang12355@163.com; **Chunxia Yao**, Master, Engineer, research interest: ventilation and environmental control in animal production systems, Email: yaocx@nercita.org.cn; **Shiwei Huang**, Master, Associate Professor, research interest: agricultural construction and environmental engineering, Email: hswcau@cau.edu.cn; **Weihong Ma**, PhD, Senior Engineer, research interest: precision livestock farming, Email: mawh@nercita.org.cn; **Ligen Yu**, PhD, Associate Research Fellow, research interest: precision livestock farming, Email: yulg@nercita.org.cn; **Ronghua Gao**, PhD, Research Fellow, research interest: precision livestock farming, Email: gaorh@nercita.org.cn.
***Corresponding author: Qifeng Li**, PhD, Research Fellow, research interest: precision livestock farming. Mailing address: No.11 Shuguang Garden Middle Road, Haidian District, Beijing 100097, China. Tel/Fax: +86-10-51503855, Email: liqf@nercita.org.cn.

relating AER to the opening ratio and environmental factors are required for real-time regulation of the ventilation system in NVCHs[5].

With the development of computer science and artificial intelligence, new tools such as computational fluid dynamics (CFD) and machine learning are increasingly used to achieve ventilation rates under different boundary conditions[13,14]. This method may provide a solution to estimating AERs in NVCH. For example, Shen et al.[15] established a statistical model based on response surface methodology (RSM) to predict the ventilation rate in NVCH using the data obtained by CFD simulation under different openings and wind speeds. Ayata et al. developed an Artificial Neural Network (ANN) model to predict indoor air velocity trained by CFD simulation data, which provided the indoor airflow parameters of urban natural ventilation buildings such as wind directions, AER, and the building's opening conditions[16]. Vogeleer et al. established an ANN model for the fast estimation of AERs in a naturally ventilated test facility with measured 2D or 3D local air velocity[17]. There is no doubt that the studies mentioned above are important in estimating AER in NVCH. The main concern is that the layout and structure of the cattle barn for model development are very different from those in China. Furthermore, traditional ANN models are more often adopted, which require a larger sample size for model training, longer computation time, and easy trapping in a locally optimal solution[18,19]. As mentioned above, this study proposed a Bayesian optimized gradient boosting decision tree (GBDT) method to estimate AERs in NVCH based on the mechanism of natural ventilation. The GBDT model is an integration algorithm based on a decision tree that uses an additive model to accumulate the residuals of multiple decision trees to achieve the minimum errors of classification or regression in the training process[20]. Compared with the traditional ANN model, the GBDT model has the advantage of high prediction accuracy when processing data with low dimensions and small sample sizes. This highly matches the data characteristics when estimating AER in NVCH[21,22].

The objective of this study is to develop a machine learning algorithm framework based on a Bayesian-optimized GBDT model to estimate AERs in NVCHs and to examine the effectiveness and reliability of the proposed algorithm framework to seek a potential method for decision-making in regulating the side curtains in NVCHs. Inputs of the GBDT model were selected based on the opening ratio and the environmental factors driven by natural ventilation, which allows the model to be used in environmental control systems. Comparisons were made with the traditional ANN model, the widely used support vector machine (SVM) in machine learning and the machine model to evaluate the performance of the GBDT model in estimating AER. To reduce the computation time and improve the adaptive ability of the model, the Bayesian algorithm was introduced into the model to automatically obtain the optimized hyperparameters in GBDT. The Bayesian algorithm is a probability-based optimal method in automatic machine learning. It has been successfully applied in environmental monitoring and sensor networks, robotics, and reinforcement learning[23]. To improve the generalization ability of the proposed model, field measurements were conducted to monitor the actual AER in two NVCHs in different seasons, and CFD simulation was adopted to obtain the AER at various opening ratios. The developed Bayesian optimized GBDT model can be exported as a 'joblib' file and deployed to the server of the web framework or adapted and inserted into the environmental regulatory system in future applications.

2 Materials and methods

2.1 AER monitoring in field measurements

Field measurements were conducted to monitor AER in two commercial free-stall dairy cattle houses in different seasons in Beijing and Tianjin, China. The cattle house in Beijing (H1) is 92 m long, 28 m wide, 3.5 m eave high, and 6.3 m ridge high. The other one in Tianjin (H2) is 186 m long, 31 m wide, 2.7 m eave high, and 10.8 m ridge high. Both H1 and H2 are naturally ventilated with two sidewall curtains and a central ventilation ridge. Figure 1 shows the layout of the experimental dairy cattle houses.

Field measurements in H1 were conducted in July and August 2019 when two sidewall curtains and ventilation ridges were fully open as the requirement of ventilation in practical production in summer. Indoor carbon dioxide (CO_2) was continuously sampled (Figure 1a, A1-A5) by a six-channel multiplexer (Innova 1409, USA) at a height of 1.5 m and synchronously analyzed by a photoacoustic multigas monitor (Innova 1512i, USA) every 5 min. Background CO_2 was sampled 30 m away from H1 upwind. Indoor air temperature (Ta), relative humidity (RH), and airspeed were recorded at a height of 2 m at nine points (Figure 1a, T1-T9, W1-W9) by a recorder (Apresys 179A-TH, USA) and hot-wire anemometer (XL62 WR1ID4, China). Ta, RH, wind speed (WS) and direction (WD) in background open air were recorded by a weather station (WS1800, China) at a height of 2.5 m and 150.0 m away from H1.

For H2, field measurements were conducted in December 2020. Sidewall curtains to the north were fully closed and those to the south were fully open as the requirement of ventilation in practical production in winter. Similarly, indoor and outdoor CO_2 were sampled (Figure 1b) and analyzed by the six-channel multiplexer and multigas monitor. Indoor Ta and RH were recorded at a height of 2.1 m at six points (Figure 1b, T1-T6). Air velocities at the openings of the side curtains were monitored by ultrasonic 3D anemometers (Wind Master, UK) at a height of 1 m in four locations at the opening (Figure 1b, W1-W4). Outdoor Ta, RH, and air velocity upwind were monitored by a recorder and an ultrasonic 3D anemometer at a height of 5 m.

AERs in the two experimental houses were calculated by the CO_2 balance method, which was recommended by the International Commission of Agricultural Engineering (CIGR) and described in Equation (1)[24]. To improve the reliability of the calculated AER (h^{-1}), ΔC_{co2} values lower than 53.9 mg/m^3 were removed, as suggested by Ding et al.[7]

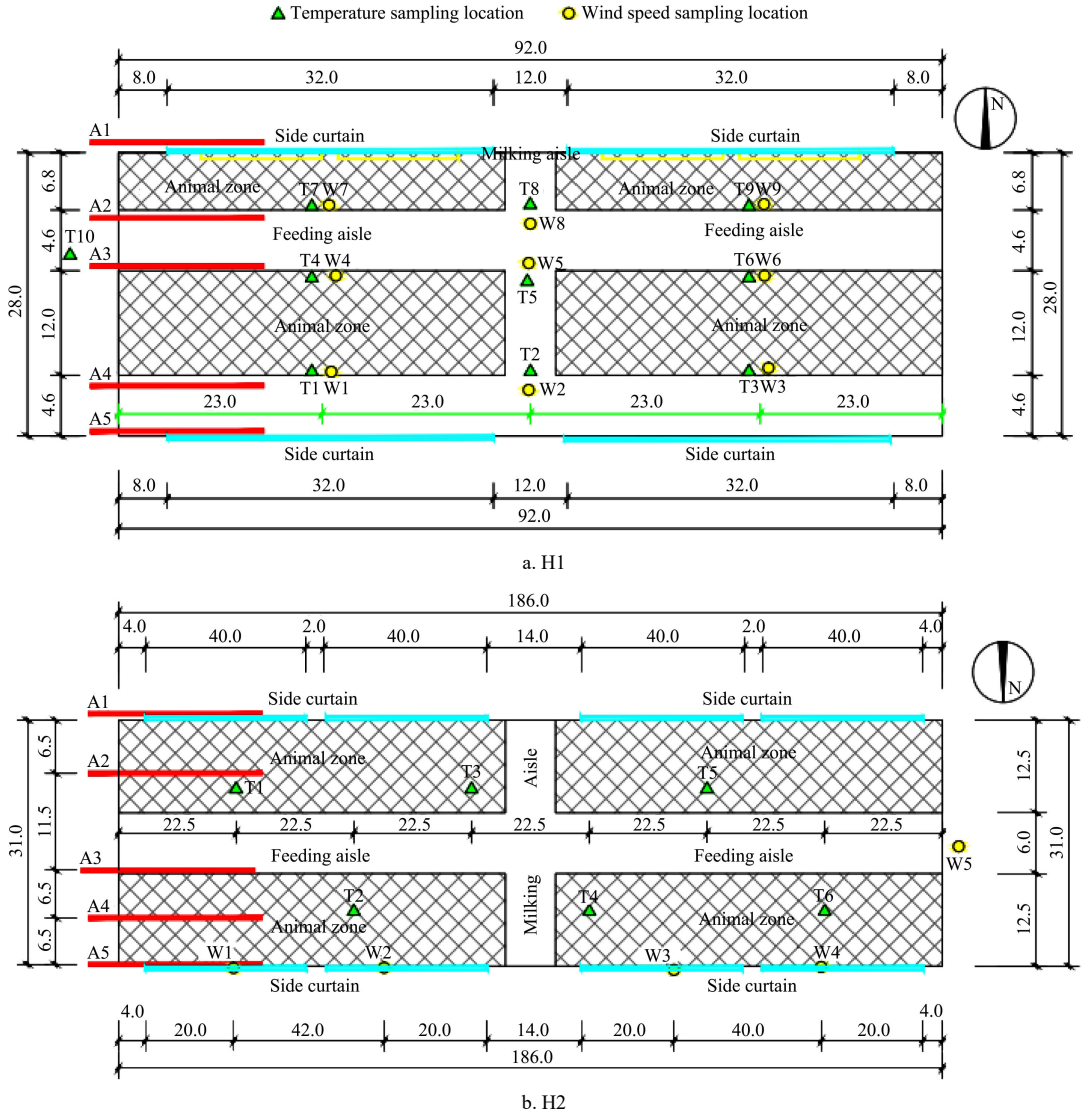
$$\mathrm{AER}=\frac{n\cdot\rho\cdot P_{\mathrm{CO_2}}}{\Delta C_{\mathrm{CO_2}}}\cdot\frac{1}{V} \tag{1}$$

$$P_{\mathrm{CO_2}}=(\mathrm{HPU}_t\cdot 0.185)\cdot\left\{1-a\cdot\sin\left[\frac{\pi}{12}\cdot(h+6-h_{\min})\right]\right\} \tag{2}$$

$$\mathrm{HPU}_t=\frac{n\cdot(5.6\cdot B^{0.75}+22\cdot Y+1.6\times10^{-5}\cdot p^3)}{1000}\cdot[1+0.004\cdot(20-t)] \tag{3}$$

where, n is the number of housed cows; ρ is the density of CO_2, 1.977 g/L; P_{CO_2} is the production of CO_2 for each cow, which can be estimated by the heat production through body weight and air temperature as shown in Equations (2) and (3)[24]; ΔC_{CO_2} is the concentration difference of indoor and outdoor CO_2, mg/m^3; V is the volume of measured NVCH, m^3; HPU_t is the heat production unit at a certain air temperature of t, 1000 W^{-1}; a is a dimensionless

constant representing the amplitude of the sine function, which is 0.22 in the case of free-ranging dairy cows[24]; h is the time in the 24-hour system, h; h_{min} is the time corresponding to the minimum activity of dairy cows, which is 2.9 in the case of free-ranging dairy cows[24]; B is the body weight of dairy cows, kg; Y is the daily milk yield of dairy cows, kg/d; p is the number of days of pregnancy, d.



a. H1

b. H2

Note: A1-A5 indicate indoor carbon dioxide sampling locations; T1-T9 indicate temperature sampling locations; W1-W9 indicate wind speed sampling locations.

Figure 1 Layouts of sampling tubes or points in dairy cattle houses H1 and H2 (m)

2.2 Numerical simulations

A full-scale physical model for H1 was created in ICEM CFD (ANSYS 15.0, PA, USA), and numerical simulations were conducted for 144 cases to enrich the dataset when the opening ratios of the side curtains were 0%, 17%, 34%, 51%, 68%, and 85%. Different combinations of outdoor Ta, RH, WS, WD, and the opening ratio were considered as the boundary conditions in the simulation cases and are summarized in Table 1. These boundary conditions were selected according to typical weather that occurred during field measurements by the weather station.

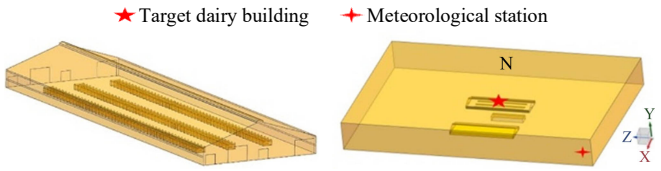
Table 1 Boundary conditions of CFD simulations

Season	Outdoor Ta/°C	Outdoor RH/%	Outdoor WS/m·s^{-1}	Outdoor WD/(°)	Opening ratio	Number of cases
Summer	28.67±1.48 (26.7-30.6)	59.41±6.13 (53.2-70.2)	1.57±0.67 (0.5-2.1)	105.10±44.14 (64.6-115.2)	0-85%	54
Winter	−4.31±3.61 (−9.0-2.0)	76.90±21.13 (37.0-94.2)	2.26±2.52 (0.5-8.6)	159.54±116.84 (0.0-351.3)	0-85%	36*
Transition season	5.36±3.63 (−1.0-8.6)	41.08±18.79 (21.4-73.9)	2.58±2.00 (0.30-7.20)	98.89±120.06 (0.0-320.3)	0-85%	54

Note: Some cases in winter were removed due to a failure of convergence in the simulation. The 36 cases are the remaining effective cases.

The CFD model was verified, and the details can be found in Li et al.[4] This was based on a full-scale dairy building model with the basic geometry shown in Figure 2. Considering the influence of the adjacent buildings around the target dairy house in a real situation, two adjacent buildings with sidewalls next to the target dairy building were included in CFD modeling. Thus, taking the target dairy house and the surrounding buildings as a whole target, this study adopts a domain size of 5H×15H×15H away from the whole target for CFD modeling, where H is the ridge height of the target dairy building.

The geometry was meshed and imported to Fluent for CFD simulation. The minimum size of the mesh was 32 mm, and the surface of the cows was encrypted. A grid independence analysis was conducted to ensure that the resolution of the mesh did not influence the results. A renormalization group (RNG) k-ε turbulence model was used to determine turbulence effects. The finite volume method was used as the discrete method of the governing equation, and the SIMPLEC algorithm was used for the pressure-speed coupling. The kinetic energy and turbulent flow energy were selected in the second-order upwind style.



a. Simulated dairy building b. Entire computational domain

Figure 2 Geometry model of simulated dairy building and entire computational domain in CFD modelling[4]

2.3 Methodology in AER modelling

2.3.1 Gradient boosting decision tree

The prediction of AER in an environmental control system usually requires real-time processing for a large quantity of data. The GBDT model is a classical feature selection integration model based on a decision tree, which is interpretable and fast in data processing, and the issue of overfitting can be effectively avoided by limiting the hyperparameters in the model[25].

GBDT is an iterative decision tree algorithm that can be regarded as an additive model (Equation (4)) composed of M trees[26].

$$F(x,w)=\sum_{m=0}^{M}\alpha_m h_m(x,w_m)=\sum_{m=0}^{M}f_m(x,w_m) \tag{4}$$

where, x is the input; w is the parameter in the model; h is the regression tree; α is the weight of each tree. Given a training dataset $T=\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where, $x_i \in \chi \subseteq R^n$ and χ is the input space, $y_i \in Y \subseteq R$ and Y is the output space. Denoting the loss function as $L(y, f(x))$, the goal of this study was to obtain the final regression tree F_M. Initialize the first weak learner as follows:

$$F_0(x)=\arg\min_{c}\sum_{i=1}^{N}L(y_i,c) \tag{5}$$

To establish M-classified regression trees, m is the number of trees (m=1, 2, 3, …, M), and i is the number of samples (i=1, 2, 3, …, N). Then, calculate the negative gradient of the loss function corresponding to the mth tree.

$$r_{m,i}=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)} \tag{6}$$

For i=1, 2, 3, …, N, the classification and regression tree (CART) is used to fit the data (x_i, $r_{m,r}$) to obtain the mth regression tree. The corresponding leaf node area is $R_{m,j}$, where j=1, 2, 3, …, J_m, and J_m is the number of leaf nodes of the mth regression tree.

For the J_m leaf node regions (j=1, 2, 3, …, J_m), calculate the best- fit value using the following equations.

$$c_{m,j}=\arg\min_{c}\sum_{x_i\in R_{m,j}}L(y_i,F_{m-1}(x_i)+c) \tag{7}$$

Update the strong learner $F_m(x)$:

$$F_m(x)=F_{m-1}(x)+\sum_{j=1}^{J_m}c_{m,j}I(x\in R_{m,j}) \tag{8}$$

Obtain the expression of strong learner $F_M(x)$:

$$F_M(x)=F_0(x)+\sum_{m=1}^{M}\sum_{j=1}^{J_m}c_{m,j}I(x\in R_{m,j}) \tag{9}$$

2.3.2 Bayesian optimization

When applying the GBDT model to predict AER, it is necessary to determine its optimal combination of hyperparameters. Considering that machine learning easily falls into a local optimal solution, this paper used the Bayesian optimization algorithm, which is a global parameter optimization algorithm, to optimize the model parameters. Based on the Bayesian theorem, the Bayesian optimization algorithm obtains the next-most-potential hyperparameter value X_i by maximizing the acquisition function, calculates the objective function value $f(X_i)$, adds the newly obtained (X_i, $f(X_i)$) to the known evaluation point set D, and updates dataset D to obtain the optimal solution in a cycle[23]. The sequential model-based optimization steps of Bayesian optimization are as follows:

Step 1 Input f, χ, S, M;
Step 2 $D \leftarrow$ InitSamples (f, χ);
Step 3 For $i \leftarrow |D|$ to T do;
Step 4 $p(y|X, D) \leftarrow$ FiteModel(M(GP), D);
Step 5 $X_i \leftarrow$ argmax $S(X, p(y|X, D))$;
Step 6 $y_i \leftarrow f(X_i)$;
Step 7 $D \leftarrow D \cup (X_i, y_i)$

where, χ is the hyperparametric search space; $X=\{X_1, X_2, X_3, \ldots, X_n\}$ represents a group of super parameter combinations; X_i is a set of super parameters selected by the acquisition function; T is the total number of function evaluations; f represents the function learning model, $D=(X_1, y_1), \ldots, (X_n, y_n)$ represents a dataset composed of several pairs of data; M is the probabilistic regression model; S is the acquisition function.

As the Bayesian optimization algorithm can make full use of historical information, it has a significantly higher efficiency compared with the other optimization methods. The probabilistic regression model and acquisition function are two core parts of Bayesian optimization. The Gaussian process (GP) was adopted as the probabilistic regression model in this study. It is the most widely used nonparametric model for probabilistic regression, has a higher expansibility, and can usually obtain satisfactory prediction results[27]. The acquisition function refers to the function mapped from input, observation, and hyperparametric space to real number space[28]. It is necessary to balance the relationship between utilization and exploration and weigh the distribution of evaluation points. To reduce the training error caused by sampling randomness, 10-fold cross-validation was applied in the Bayesian optimization search[29].

2.3.3 Modelization

Preprocessing, such as outlier discarding and normalization, was performed on the data from field measurements and CFD simulations to obtain datasets for AER modelling. There are twelve factors in the dataset: indoor Ta, indoor RH, outdoor Ta, outdoor RH, indoor and outdoor temperature difference (ΔTa), indoor and outdoor humidity difference (ΔRH), wind speed (V_{out}), wind direction (V_{dir}), north opening ratio (Nopening), south opening ratio (Sopening), AER and season. There is an obvious stratification after visualization of the collected environmental data due to the seasonal variation in measurements. Moreover, ΔTa and ΔRH are the driving force and natural ventilation and its mass transfer, respectively. Thus, these three factors (ΔTa, ΔRH, and season) were included in the dataset. 80% of the data in the dataset were randomly selected and used for model training. The remaining 20% of the data was used for model validation.

Figure 3 shows the methodological framework to predict AER in this study. The GBDT model was trained by the 10-fold cross-validation method to predict AER. Hyperparameters in the models were first obtained by the grid search method based on the mean square error. To evaluate the effectiveness of GBDT, the model performance was compared with the commonly used SVM model and the multilayer perceptron (MLP), a feedforward ANN model. Then, Bayesian optimization was applied to the GBDT model to optimize the hyperparameters in the model. When Bayesian optimization was used to optimize the GBDT model, the

different hyperparameter combinations of GBDT were taken as the independent variables, and the mean square error obtained by the cross-validation evaluation was taken as the output of the Bayesian framework. Iterations were conducted until the loss function was minimized.

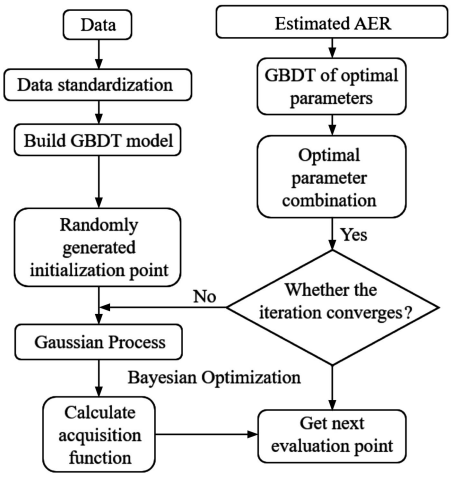


Figure 3 Overview of a methodological framework to predict AER by Bayesian optimized GBDT

2.4 Model evaluation

The mean absolute error (MAE), mean absolute percentage error (MAPE), the goodness of fit (R^2), and mean square error (MSE) were used to evaluate the models in predicting AER.

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\hat{y}_i\right| \tag{10}$$

$$\mathrm{MAPE}=\frac{1}{m}\sum_{i=1}^{m}\frac{\left|y_i-\hat{y}_i\right|}{y_i} \tag{11}$$

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

$$R^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\overline{y})^2} \tag{13}$$

where, m is the number of samples; $\hat{y}_i$ is the predicted results of the model; y_i is the measured or CFD simulated results (true values); $\overline{y}$ is the average of true values.

3 Results and discussion

3.1 Evaluation of GBDT model

As mentioned above, the effectiveness of GBDT model was evaluated by comparing it with the commonly used ANN (MLP used in this study) and SVM models. Table 2 lists the hyperparameters obtained through Grid Search, a conventional method to automatically find the optimal parameters, in established models to estimate the AER in NVCH. The quantitative evaluators of different models are also listed in Table 2. R^2 is the fraction of total variation that is explained by the regression, which ranges from 0 to 1. The higher the R^2, the better the explanation effect of the regression model. The authors believe that the model is more reliable when R^2 is higher than 0.8. MSE and MAE evaluate the prediction accuracy, which ranges from 0 to infinity. The lower the MSE and MAE, the better the accuracy of the prediction model. MAPE evaluates the relative errors between the actual and the prediction, which range from 0% to infinity. This indicates a perfect match when MAPE is 0% and a very poor prediction when MAPE is over 100%. As listed in Table 2, the MLP model shows the poorest performance, while the GBDT model had the highest R^2 and lowest MSE, MAE, and MAPE, showing the best performance.

Table 2 Hyperparameters in model based on grid search and evaluation of different models with grid search in predicting AER

Model	Hyperparameter	Range	Optimized value	R^2	MSE	MAE	MAPE
SVM	C	[10^{-3}, 10]	3.0786	0.63	0.41	0.45	11%
	nu	[10^{-3}, 1]	0.5499				
MLP	hidden_layer_sizes	[1, 50]	28	0.35	0.72	0.57	17%
	alpha	[10^{-3}, 1]	0.0016				
	learning_rate_init	[10^{-4}, 1]	0.0656				
GBDT	learning_rate	[10^{-3}, 1]	0.0530	0.84	0.17	0.28	7%
	max_depth	[2, 5]	4				
	n_estimators	[90, 120]	100				
	min_samples_leaf	[1, 3]	2				

Note: SVM: Support Vector Machine; MLP: Multilayer Perceptron; GBDT: Gradient Boosting Decision Tree; MSE: Mean Square Error; MAE: Mean Absolute Error; MAPE: Mean Absolute Percentage Error.

Figure 4 demonstrates the residual of the three models. The absolute residual between the predicted and measured AER can intuitively reflect the performance of each model. The lower the absolute residual, the better. As shown in Figure 4, there were significantly smaller prediction residuals for the GBDT model than for the SVM and MLP models. Its residuals tended to be reduced with increasing measured AER.

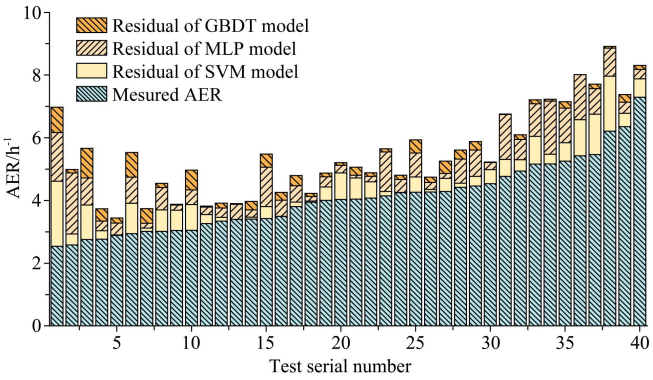


Figure 4 Comparison between measured and predicted AER of different models

3.2 Bayesian optimized GBDT

Grid search was first adopted in the three models introduced in Section 3.1 to obtain the optimal hyperparameter. However, the Grid Search method requires considerable computing resources and takes a long time when searching for hyperparameters. Alternatively, Bayesian optimization was conducted and compared with the grid search method to obtain the hyperparameters in GBDT in estimating AER.

The ranges of max_depth, n_estimators, learning_rate, and subsample were set as [10^{-3}, 1], [2, 5], [90, 120], and [1, 3], respectively. Then, the two mentioned optimization methods were used to calculate the optimal combination hyperparameter. It took 12 284 s for Grid Search to find the optimal combination of hyperparameters. Bayesian optimization required 94 s for 97 iterations, dramatically reducing the computation time. Bayesian optimization is an active optimization based on the results of iteration, which is overall stable in the search process and can obtain the optimal combination with less time. However, Grid Search optimization depends on the number of iterations, and the

emergence of an optimal combination has high randomness, resulting in low efficiency. Meanwhile, the R^2, MSE, MAE, and MAPE of the Bayesian optimized GBDT model were close and even better than those of the optimized GBDT (Table 3). This suggests that Bayesian optimization is more effective than grid search. The Bayesian optimized GBDT can be a reliable model that is precise enough to estimate AER and can greatly save on computing resources.

Table 3 Comparisons between Bayesian and Grid Search optimized GBDT

Items	Indicators	Bayesian optimized GBDT	Grid Search optimized GBDT
Performance Evaluation	R^2	0.86	0.84
	MSE	0.17	0.17
	MAE	0.27	0.28
	MAPE	7%	7%
	Computation time	94 s	12 284 s
Hyperparameters	learning rate	0.055	0.053
	max_depth	4	4
	n_estimators	98	100
	min_samples_leaf	2	2

Table 3 also lists the obtained hyperparameters through different methods. They were very close to each other. In the process of hyperparameter optimization, the variations between hyperparameters and validation scores were plotted to illustrate and verify the effectiveness of hyperparameter selection. Figure 5 demonstrates the verification curves of four parameters in the GBDT model: max_depth, n_estimators, learning_rate, and subsample. The max_depth represents the maximum depth of each regression estimator, which limits the number of nodes in the tree. The n_estimators indicate the number of boosting phases to be executed. Learning_rate represents the learning rate, which controls the contribution of each tree. Subsample represents the proportion of samples taken, which is used to fit the score of a single base learner's sample. The plot was drawn using the validation score data from 10-fold cross-validation with its mean (solid curve in Figure 5) and variance (error band in Figure 5).

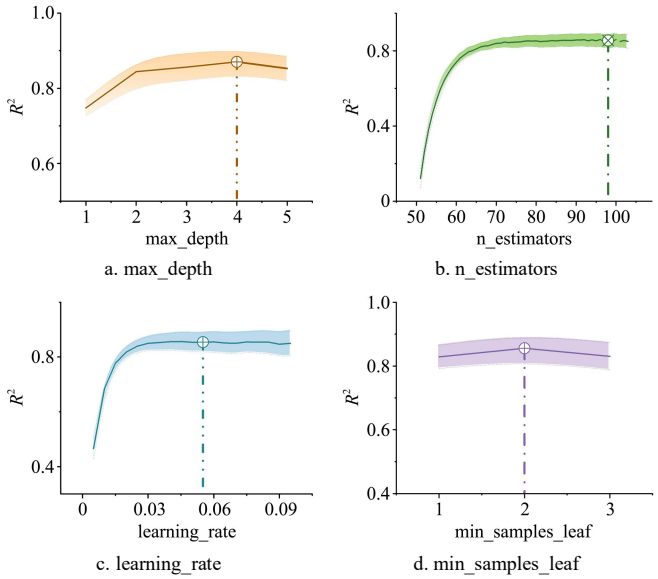


Figure 5 Verification curve of hyperparameters in GBDT model

Figure 5 shows that the value of each hyperparameter falls in the region where R^2 is over 0.8, which ensures the effectiveness of the selected hyperparameters in the GBDT model. With the monotonous change in hyperparameters, R^2 gradually tends to be stable, while the optimal parameter does not appear where the stable R^2 has just been reached. This is because the hyperparameters are not independent of each other, and the optimal combination of a local parameter does not represent the optimal combination of the global parameter.

3.3 Importance analysis of inputs

The influence of the input was compared in the process of building the GBDT model. Figure 6 shows the importance of inputs in the Bayesian-optimized GBDT. When all 11 inputs were used in building the model, the south opening ratio, wind speed, north opening ratio, and ΔTa had a great impact on the model. This was followed by outdoor Ta, ΔRH, wind direction, indoor Ta, outdoor RH and indoor RH. The temperature difference and outdoor wind speed are the driving forces of natural ventilation[5]. At the same time, the outdoor wind speed and direction affect the differences in Ta, RH, and gas concentration inside and outside the NVCH[8]. The airflow rate in an NVCH is dominated by the outdoor wind speed when the opening ratio and wind speed are high enough, making the temperature difference decrease and the impact of other environmental factors small enough to be ignored. The opening ratio, wind speed, and temperature difference were the most important inputs to the GBDT model, and the temperature difference was less important than the opening ratio and wind speed. This is consistent with the mechanism of natural ventilation, which verifies the reliability of the model. The influence of the season on the modelling was almost 0. This is because there is a strong correlation between season and temperature inside and outside the house, resulting in the least importance in the model.

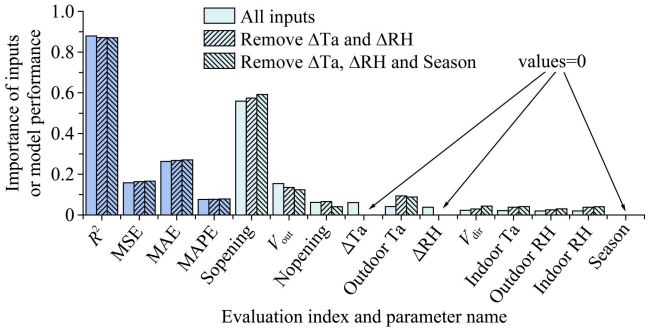


Figure 6 Importance of inputs and comparison of input combinations in GBDT

Two groups of comparative tests were carried out to explore the influence of relevant inputs in modelling. One test removed the inputs of ΔTa and ΔRH, as they can be calculated through Ta or RH inside and outside the dairy house. The other test removed the ΔTa, ΔRH, and season, as the season shows the least importance in modelling. As shown in Figure 6, the removal of relevant inputs has little effect on the R^2, MSE, MAE, and MAPE of the established GBDT model in estimating AER, suggesting that ΔTa, ΔRH, and season can be excluded to make full use of the remaining information to make the model more concise. However, the importance of the remaining factors changed slightly after the removal of inputs. The importance of outdoor Ta greatly increased after ΔTa was removed and became the third most important input in the GBDT model. As shown in Figure 6, the outdoor Ta had larger importance of inputs than indoor Ta, which suggests that AER was more dependent on outdoor Ta than indoor Ta when using the GBDT model.

3.4 Comparisons with existing models

The established GBDT model was validated and compared

with the existing models using the verification dataset from field measurements. The mechanism model based on the Bernoulli equation has a good explanation for the results and estimates the AER according to the opening area, wind speed at the opening, and empirical coefficients[5,7].

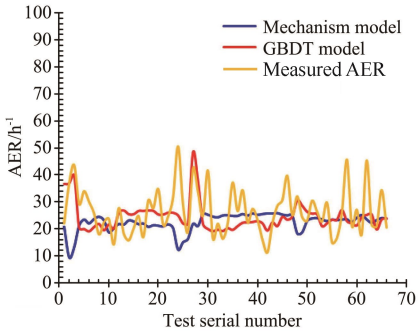


Figure 7 Comparison of estimated AER by different models

As shown in Figure 7, the curves between the AER predicted by the GBDT model and the mechanism model are very close to each other. The average difference in AER estimated between the mechanism model and the Bayesian-optimized GBDT model was 5.0 h^{-1} (20.6%). Using the measured AER from field measurements as the baseline, the averaged estimation errors of the mechanism model and the GBDT model in this study were 26.1% and 23.5%, respectively. The developed GBDT model showed a slightly lower error than the mechanism model. This further verifies the validity of the Bayesian-optimized GBDT model in this study. Compared with the mechanism model, the GBDT model does not rely on the empirical coefficient whose value varies with the opening and only needs to measure the outdoor wind speed instead of the wind speeds at multiple openings. This makes the developed GBDT model much easier to use for real-time environmental control in practical applications.

4 Conclusions

Field measurements and CFD simulations were conducted to obtain the AER in NVCHs under different environmental conditions. A Bayesian-optimized GBDT model was proposed to predict AER to adjust the side opening for ventilation. The proposed model can improve the applicability to small samples and avoid the drawbacks of easily falling into a local optimum. Its effectiveness and reliability were examined, and the results showed the following:

1) Compared with Grid Search, Bayesian optimization can make full use of historical information, greatly reduce the computation time, and improve the efficiency of hyperparameter search in GBDT modelling;

2) The GBDT model can be more concise with effective performance in predicting the AER after removing the inputs of season, air temperature, and humidity difference from the original 11 parameters. The model was greatly dependent on the opening ratio, outdoor wind speed, and outdoor air temperature;

3) The proposed GBDT model had an R^2 of 0.84, showing a better performance than traditional machine learning models such as MLP and SVM. Compared with the mechanism model, it had a similar or even slightly higher estimation accuracy but much easier obtained inputs. This makes it more applicable for use in practical applications for real-time environmental control.

Acknowledgements

This work was financially supported by the National Key Research and Development Program of China (2019YFE0125400), the Beijing Natural Science Foundation (Grant No. 6194037), and the Youth Personnel Project of Beijing Outstanding Talents.

[References]

[1] Samer M, Ammon C, Loebsin C, Fiedler M, Berg W, Sanftleben P, et al. Moisture balance and tracer gas technique for ventilation rates measurement and greenhouse gases and ammonia emissions quantification in naturally ventilated buildings. Building & Environment, 2012; 50: 10–20.

[2] Yi Q Y, Li H, Wang X S, Zong C, Zhang G Q. Numerical investigation on the effects of building configuration on discharge coefficient for a cross-ventilated dairy building model. Biosystems Engineering, 2019; 182: 107–122.

[3] Saha C K, Ammon C, Berg W, Loebsin C, Fiedler M, Brunsch R, et al. The effect of external wind speed and direction on sampling point concentrations, air change rate and emissions from a naturally ventilated dairy building. Biosystems Engineering, 2013; 114(3): 267–278.

[4] Li Q F, Yao C X, Ding L Y, Yu L G, Ma W H, Gao R H, et al. Numerical investigation on effects of side curtain opening behavior on indoor climate of naturally ventilated dairy buildings. Int J Agric & Biol Eng, 2020; 13(5): 63–72.

[5] Rong L, Bjerg B, Batzanas T, Zhang G Q. Mechanisms of natural ventilation in livestock buildings: Perspectives on past achievements and future challenges. Biosystems Engineering, 2016; 151: 200–217.

[6] Yi Q Y, Zhang G Q, Koenig M, Janke D, Hempel S, Amon T. Investigation of discharge coefficient for wind-driven naturally ventilated dairy barns. Energy & Buildings, 2018; 165: 132–140.

[7] Ding L Y, E L, Li Q F, Yao C X, Wang C C, Yu L G, et al. Mechanism analysis and airflow rate estimation of natural ventilation in livestock buildings. Transactions of the CSAE, 2020; 36(15): 189–201. (in Chinese)

[8] Wang X, Ndegwa P M, Joo H S, Neerackal G M, Stckle C O, Liu H P, et al. Indirect method versus direct method for measuring ventilation rates in naturally ventilated dairy houses. Biosystems Engineering, 2016; 144: 13–25.

[9] Edouard N, Mosquera J, van Dooren H, Mendes L B, Ogink N W M. Comparison of CO_2- and SF_6- based tracer gas methods for the estimation of ventilation rates in a naturally ventilated dairy barn. Biosystems Engineering, 2016; 149: 11–23.

[10] Kiwan A, Berg W, Fiedler M, Ammon C, Glaser M, Mueller H-J, et al. Air exchange rate measurements in naturally ventilated dairy buildings using the tracer gas decay method with ^{85}Kr, compared to CO_2 mass balance and discharge coefficient methods. Biosystems Engineering, 2013; 116(3): 286–296.

[11] Kiwan A, Berg W, Brunsch R, Zcan S, Berckmans D. Tracer gas technique, air velocity measurement and natural ventilation method for estimating ventilation rates through naturally ventilated barns. Agricultural Engineering International: The CIGR e-journal, 2012; 14(4): 22–36.

[12] Van Overbeke P, De Vogeleer G, Pieters J G, Demeyer P. Development of a reference method for airflow rate measurements through rectangular vents towards application in naturally ventilated animal houses: Part 3: Application in a test facility in the open. Computers and Electronics in Agriculture, 2015; 115: 97–107.

[13] Xiong S, Zhang G Q, Bjerg B. Comparison of different methods for estimating ventilation rates through wind driven ventilated buildings. Energy & Buildings, 2012; 54(6): 297–306.

[14] Wang X S, Wu J G, Yi Q Y, Zhang G Q, Amon T, Janke D, et al. Numerical evaluation on ventilation rates of a novel multi-floor pig building using computational fluid dynamics. Computers and Electronics in Agriculture, 2021; 182: 106050. doi: 10.1016/j.compag.2021.106050.

[15] Shen X, Zhang G Q, Wu W T, Bjerg B. Model-based control of natural ventilation in dairy buildings. Computers & Electronics in Agriculture, 2013; 94: 47–57.

[16] Ayata T, Arcaklıoğlu E, Yıldız O. Application of ANN to explore the potential use of natural ventilation in buildings in Turkey. Applied Thermal Engineering, 2007; 27(1): 12–20.

[17] De Vogeleer G, Van Overbeke P, Brusselman E, Mendes L B, Pieters J G, Demeyer P. Assessing airflow rates of a naturally ventilated test facility using a fast and simple algorithm supported by local air velocity measurements. Building & Environment, 2016; 104: 198–207.

[18] Erdik T, Şen Z. Comments on 'A comparative study of ANN and neuro-fuzzy for the prediction of dynamic constant of rockmass'. Journal of Earth System Science, 2008; 117(6): 973–974.

[19] Momeni E, Nazir R, Armaghani D J, Maizir H. Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN. Measurement, 2014; 57: 122–131.

[20] Friedman J H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001; 29(5): 1189–1232.

[21] Pan C, Tan J, Feng D D. Identification of power quality disturbance sources using gradient boosting decision tree. In: 2018 Chinese Automation Congress (CAC), Xi'an: IEEE, 2018; 2589–2592. doi: 10.1109/CAC.2018.8623162.

[22] Zhang L K, Pan H, Fan Q, Ai C Q, Jing Y Q. 1GBDT, LR & deep learning for turn-based strategy game AI. In: 2019 IEEE Conference on Games (CoG), London: IEEE, 2019; pp.1–8. doi: 10.1109/CIG.2019.8848103.

[23] Shahriari B, Swersky K, Wang Z Y, Adams R P, Freitas N D. Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE, 2015; 104(1): 148–175.

[24] Pedersen S, Sllvik K. Climatization of animal houses: Heat and moisture production at animal and house levels. Horsens: Research Centre Bygholm, Danish Institute of Agricultural Sciences, 2002; 42p.

[25] Si M X, Du K. Development of a predictive emissions model using a gradient boosting machine learning method. Environmental Technology & Innovation, 2020; 20: 101028. doi: 10.1016/j.eti.2020.101028.

[26] Brochu E, Brochu T, Freitas N D. A Bayesian interactive optimization approach to procedural animation design. The Eurographics Association, 2010; pp.103–112. doi: 10.2312/SCA/SCA10/103-112.

[27] Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 2020; 415: 295–316.

[28] Pontes F J, Amorim G F, Balestrassi P P, Paiva A P, Ferreira J R. Design of experiments and focused grid search for neural network parameter optimization. Neurocomputing, 2016; 186: 22–34.

[29] Vakharia V, Gujar R. Prediction of compressive strength and portland cement composition using cross-validation and feature ranking techniques. Construction and Building Materials, 2019; 225: 292–301.