Temperature control mode prediction in a greenhouse based on SMOTETomek-ISSA-CatBoost model

Xiaojuan Mao[1,2,3], Hongyu Lu[2], Zhongyi Yi[1,2], Ni Ren[2,3], Jing Jin[2,3*]

(1. *School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China*;
2. *Agricultural Information Institute, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China*;
3. *Key Laboratory of Smart Agricultural Technology (Yangtze River Delta), Ministry of Agriculture and Rural Affairs of the People's Republic of China, Nanjing 210014, China)*)

Abstract: Temperature is a critical factor influencing crop growth in controlled environment agriculture. Accurate regulation of air temperature within a greenhouse is essential for promoting optimal crop development and enhancing production efficiency. In this study, a Categorical Boosting model based on SMOTETomek mixed sampling method and improved Sparrow Search Algorithm (SMOTETomek-ISSA-CatBoost) was proposed to predict the categories of greenhouse temperature control modes. This study utilized historical temperature control mode data, which had been accumulated by cultivation experts through practical production and demonstrated effective in temperature management. To enhance the model's performance and achieve real-time, precise temperature regulation in greenhouses, firstly, the SMOTETomek mixed sampling method was utilized to expand the original training set, effectively addressing the issue of data imbalance. Secondly, the Latin Hypercube Sampling (LHS) method, the Cauchy mutation perturbation operator, and a greedy rule were employed to refine the Sparrow Search Algorithm to enhance the global search capability. Ultimately, the improved Sparrow Search Algorithm was employed to optimize the hyper-parameters of CatBoost model to improve its predictive accuracy. Compared with SMOTETomek-CatBoost models optimized by Whale Optimization Algorithm (WOA), Fruit Fly Optimization Algorithm (FOA), Particle Swarm Optimization (PSO), and standard Sparrow Search Algorithm (SSA), the SMOTETomek-ISSA-CatBoost model demonstrated better prediction efficacy, with F1-score and AUC values reaching 0.8147 and 0.9629, respectively. The SMOTETomek-ISSA-CatBoost model exhibited the capability to predict the category of temperature control modes in a greenhouse accurately, thereby providing a decision-making foundation for intelligent management of greenhouse environments.

Keywords: greenhouse, temperature control mode, CatBoost, SMOTETomek, ISSA
DOI: 10.25165/j.ijabe.20261903.9630

Citation: Mao X J, Lu H Y, Yi Z Y, Ren N, Jin J. Temperature control mode prediction in a greenhouse based on SMOTETomek-ISSA-CatBoost model. Int J Agric & Biol Eng, 2026; 19(3): 123–132.

1 Introduction

The widespread application of greenhouses has become a crucial method for crop cultivation, especially in regions experiencing adverse climatic conditions such as extreme high temperatures. Greenhouses significantly enhance crop productivity by providing a controllable environment. Within greenhouses, environmental factors have a decisive impact on the growth and development of crops, with temperature emerging as a pivotal environmental factor, directly influencing the photosynthesis, respiration, and transpiration processes of crops[1]. The alteration of temperature has a significant impact on the growth rate of crops, and is closely associated with the yield and quality of crops. Therefore, achieving precise control of greenhouse temperature is of paramount importance for ensuring the healthy growth of crops and optimizing production efficiency.

Conventional methods for controlling environments within greenhouses primarily encompass threshold-based switch control[2,3], Proportional-Integral-Derivative feedback control (PID)[4,5], model predictive control (MPC)[6,7], and robust control[8], etc. Although the threshold-based switch control has the potential to enhance control precision, promote plant growth, and reduce resource consumption to a certain extent, its inability to adapt feedback adjustments to real-time environmental fluctuations introduces certain limitations[9]. The PID control method is capable of achieving feedback adjustment in accordance with environmental monitoring data. However, due to the dynamic nature of greenhouse environments, significant fluctuations may still occur[10]. MPC and robust control can achieve accurate temperature regulation. However, they are affected by issues such as delayed responses, intricate modeling processes, and elevated system costs. Furthermore, in practical applications, the dynamic changes in the environment and various uncertainty factors can adversely affect the regulation performance, further limiting the effectiveness and stability of these methods in large-scale greenhouse cultivation.

In recent years, the rapid development of artificial intelligence technologies such as machine learning and deep learning has rendered data-driven intelligent greenhouse environment regulation a prominent research focus[11-16]. Chen et al.[17] proposed a novel control strategy that integrated the Particle Swarm Optimization algorithm with MPC to regulate air temperature within greenhouses. The simulations demonstrated that this integrated approach was effective in tracking the desired temperature trajectory, even in the

Received date: 2024-12-21 **Accepted date:** 2026-06-02
Biographies: Xiaojuan Mao, PhD candidate, research interest: environmental perception and intelligent control, Email: 20190033@jaas.ac.cn; **Hongyu Lu**, MS candidate, research interest: agricultural data analysis and utilization, Email: rogerluhongyu@163.com; **Zhongyi Yi**, Professor, research interest: intelligent agricultural machinery and equipment, Email: yzy201@163.com; **Ni Ren**, Professor, research interest: smart agriculture, Email: rn@jaas.ac.cn.
***Corresponding author: Jing Jin**, PhD, research interest: data diagnosis, Nanjing 210014, China. Tel: +86-025-84391658, Email: jinjing@jaas.ac.cn.

presence of disturbances. Mahmood et al.[18] proposed a data-driven MPC that employed the multi-layer perceptron algorithm to construct a temperature prediction model in a greenhouse. The findings indicated that this method not only maintained an optimal environment conducive to crop growth but also significantly reduced energy consumption in comparison to conventional control methodologies. Jung et al.[19] employed a neural network to construct a temperature prediction model within a greenhouse, establishing a ventilation control decision model based on an output feedback neural network. This model was designed to regulate the opening and closing of ventilation windows within greenhouses, thereby optimizing temperature control. Mahmood et al.[20] presented a methodology for regulating greenhouse temperatures based on a data-driven model. The method employs artificial neural networks to achieve a dynamic prediction of indoor temperature, and optimizes the prediction model by combining the function of minimizing the operating cost of the greenhouse with Particle Swarm Optimization algorithm. This enables the proposal of a control strategy to ensure the stable operation of the system. However, these methods rely excessively on real-time monitoring data while neglecting the valuable expertise of domain specialists. Consequently, these methods lack the requisite adaptability and flexibility to effectively address the complexities and dynamics inherent to production environments.

The CatBoost algorithm is an open-source machine learning algorithm developed by Yandex[21]. It utilizes symmetric decision trees as the base learner and is capable of efficiently handling categorical features. It addresses the issues of gradient bias and prediction bias, which in turn reduces model overfitting and enhances both model performance and generalization ability. Currently, the CatBoost algorithm has demonstrated significant effectiveness in addressing and solving classification problems[22-25].

The Sparrow Search Algorithm is a swarm intelligence algorithm inspired by the foraging and vigilance behaviors of sparrows. The algorithm exhibits several notable advantages, including a simple underlying principle, strong robustness, rapid convergence speed, minimal parameter tuning requirements, and strong searching capability. The algorithm has shown promising performance in parameter optimization problems[26].

In agricultural production, during the summer’s peak heat season, cooling control often emerges as the primary consumer of energy. Accordingly, this paper employed the CatBoost algorithm to construct a decision-making model for greenhouse cooling control in summer. The model utilized historical temperature control data from cultivation experts with proven effectiveness in actual production. To further improve the performance and generalization ability of the model, this study employed the mixed sampling method to handle the imbalanced dataset, and introduced the improved Sparrow Search Algorithm to optimize the hyper-parameters of the CatBoost model, thereby achieving a more precise and efficient temperature regulation effect. In this study, the Sparrow Search Algorithm was enhanced in the following two aspects: 1) The LHS method was adopted to initialize the sparrow population, yielding a more uniform distribution over the search space and thereby improving the global exploration ability of the algorithm. 2) The Cauchy mutation perturbation operator combined with a greedy rule was introduced to perform Cauchy perturbation on the optimal sparrow individual in the population. This strategy effectively expands the search space, enhances population diversity, and further improves the global search performance of the SSA.

2 Materials and methods

2.1 Data acquisition

In this study, experiments were conducted in a solar panel-equipped greenhouse located at Jiangsu Academy of Agricultural Sciences in Nanjing, China. The greenhouse adopted a single-span, double-peaked Venlo-type structure, where the roof and all four side walls were constructed using 8 mm solar panels. The facility is equipped with a shading system, an insulation system, an intermittent roof ventilation system, a wet-curtain fan cooling system, and other equipment. The internal structure of the greenhouse is illustrated in Figure 1. During the summer cooling operation, the greenhouse operates in one of four distinct control modes: static, natural ventilation, mechanical ventilation, and fan-wet curtain mode. The cooling capacity is gradually increased. The greenhouse switches among these four modes to achieve the desired cooling effect, as shown in Table 1.

Figure 1 Internal view of the Venlo-type greenhouse

Table 1 The operating modes of greenhouse and the corresponding status of equipment

Control mode	Window	Fan	Wet curtain	Classify
Static	Close	Close	Close	0
Natural ventilation	Open	Close	Close	1
Mechanical ventilation	Close	Open	Close	2
Fan-wet curtain	Close	Open	Open	3

Wireless sensors (JXBS-3001, Clear Communication, China) were installed within the greenhouse to collect environmental factors, including air temperature, air relative humidity, photosynthetically active radiation, light intensity, and carbon dioxide concentration. A small weather station (QI, Dongfang Zhigen, China) was set up outside the greenhouse to collect outdoor meteorological data, including air temperature, air relative humidity, wind speed, rainfall, and photosynthetically active radiation. Furthermore, the PLC controller and MQTT gateway were utilized for the real-time monitoring and operation control of the environmental control devices within the greenhouse.

During the experimental period, the cherry tomatoes cultivated within the greenhouse were at the fruiting stage, necessitating the summer cooling management. The dataset used in this study was selected from May 1, 2024 to May 31, 2024, a period associated with high tomato yield and favorable fruit quality. A total of 4464 samples were collected, with a sampling interval of 10 min. In light of the missing and anomalous data in the dataset, this paper introduced the quadratic interpolation method to imputation or replacement. To account for discrepancies in data units and dimensions, the Min-Max normalization method was employed to standardize the data.

2.2 CatBoost model

CatBoost, a gradient boosting algorithm based on Gradient Boosted Decision Trees (GBDT), was developed by the Russian technology company in 2017[21]. It adopts symmetric decision trees as a base predictor and exhibits superior performance in terms of speed and accuracy compared to traditional GBDT algorithms. Its key advantage lies in its ability to process categorical features effectively. By incorporating ordered boosting, CatBoost can effectively address gradient bias and prediction bias, avoid overfitting issues, and further enhance the accuracy and generalization capacity of the model.

The CatBoost employs two principal methodologies:

1) A method of processing categorical features

Categorical features, a common type of discrete non-numerical features, have traditionally been processed using approaches such as one-hot encoding. However, one-hot encoding can result in a considerable increase in feature space dimensionality when dealing with a large number of categories. To address this issue, CatBoost introduces the Greedy Target-based Statistics (TBS) method, which transforms categorical features into numerical values according to target statistics, thereby effectively mitigating dimensionality expansion. Nevertheless, this method may lead to conditional shift when discrepancies exist between the distributions of the training and test datasets, since relying exclusively on label averages can ignore valuable information embedded in the features. To solve this problem, CatBoost employs a methodology that incorporates prior distributions to enhance the original Greedy TBS. This approach prevents the conditional shift caused by using label averages as feature values when distribution differences exist between datasets. Consequently, the model's accuracy is effectively improved. The revised formula is presented below:

$$x_{\sigma_j,k} = \frac{\sum_{j=1}^{p-1}[x_{\sigma_j,k} = x_{\sigma_p,k}]Y_{\sigma_j} + a \cdot p}{\sum_{j=1}^{p-1}[x_{\sigma_j,k} = x_{\sigma_p,k}] + a} \tag{1}$$

where, p represents the prior value; a represents the weight of the prior ($a > 0$). Incorporating a prior is to attenuate the noise associated with the fringe category.

2) Ordered Boosting Method

Within the conventional GBDT algorithm, the model is updated iteratively by computing the gradient of the loss function using the same dataset. Although this strategy simplifies computation, it is prone to gradient estimation bias influenced by data noise. Such disturbance induces estimation bias and further leads to model overfitting. In contrast, the CatBoost algorithm introduces an ordered boosting mechanism and employs disjoint data subsets for gradient estimation during tree construction, which effectively eliminates gradient estimation errors caused by dataset noise. This methodology not only improves the precision of gradient estimation but also refines the model's approach to mitigating prediction bias, thereby augmenting the model's generalization capabilities and predictive performance.

In this study, the indoor air temperature, indoor relative humidity, indoor photosynthetically active radiation, indoor light intensity, indoor carbon dioxide concentration, outdoor air temperature, outdoor air relative humidity, outdoor photosynthetically active radiation, outdoor wind speed, and rainfall were selected as input parameters for the CatBoost model. The output of the CatBoost model was the category of temperature control modes in the greenhouse at the next time step. The model output of 0 represents a static control mode, while a model output of 1 denotes a natural ventilation control mode. Similarly, a model output of 2 represents a mechanical ventilation control mode, and a model output of 3 indicates a fan-wet curtain control mode.

2.3 ISSA

2.3.1 Sparrow Search Algorithm

The Sparrow Search Algorithm (SSA) was proposed in 2020 as an optimization algorithm, inspired by the foraging and anti-predation behaviors of sparrows[27]. In a D-dimensional solution space, each sparrow's position is defined as $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,d}]$, with $f_i = f(x_{i,1}, x_{i,2}, \ldots, x_{i,d})$ serving as its fitness value. The number of sparrows is set to N. In each generation, the best PD individuals are selected as the discoverers, while the remaining N-PD sparrows are classified as participants.

In the SSA, discoverers with higher fitness values are given priority in the food search process. Furthermore, as discoverers are responsible for locating food for the entire sparrow population and providing foraging directions to all participants, they are granted a wider search range for foraging. The position of discoverers in each iteration is updated as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \cdot exp\left(\dfrac{-i}{\alpha \cdot T}\right), & R_2 < \mathrm{ST} \\ X_{i,j}^t + Q \cdot L, & R_2 \geq \mathrm{ST} \end{cases} \tag{2}$$

where, t is the current iteration number; T is the maximum iteration number; $X_{i,j}^t$ represents the position information of the sparrow; α is the random number between 0 and 1; Q represents a random number following a normal distribution; L is a range vector; R_2 is an alert value; ST is a safety value.

During the foraging process, some participants maintain visual contact with the discoverers. Once they perceive that the discoverers have found better food, they immediately leave their current position to compete for the food. The position of participants is updated as follows:

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\dfrac{X_{\mathrm{worst}} - X_{i,j}^t}{i^2}\right), & i > \dfrac{n}{2} \\ X_p^{t+1} + \left|X_{i,j}^t - X_p^{t+1}\right| \cdot A^+ \cdot L, & \text{otherwise} \end{cases} \tag{3}$$

where, X_{worst} is the current global worst position; X_p^{t+1} indicates the position of the sparrow with the best fitness in the $t+1$ iterations; A is a $1 \in d$ matrix with its elements randomly assigned as 1 or –1; and $A^+ = A^T(AA^T)^{-1}$.

When sparrows perceive a threat, they abandon the current food source and engage in an alarm behavior. The position of vigilant is updated as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{\mathrm{best}}^t + \beta \cdot \left|X_{i,j}^t - X_{\mathrm{best}}^t\right|, & f_i > f_g \\ X_{i,j}^t + K \cdot \left(\dfrac{|X_{i,j}^t - X_{\mathrm{worst}}^t|}{(f_i - f_w) + \varepsilon}\right), & f_i = f_g \end{cases} \tag{4}$$

where, X_{best}^t is the current global best position; K is a random number between –1 and 1; β is a step control parameter, following a normal distribution with a mean of 0 and a variance of 1; ε is the smallest constant to avoid division by zero; f_i is the fitness values of current individual sparrows; f_w and f_g represent the worst and best global fitness values, respectively.

However, the original sparrow search algorithm still exhibits certain deficiencies, including low precision and a proclivity for local optimal solutions. This paper improved the algorithm, increasing the diversity of the initial population individuals and utilizing Cauchy perturbation for the optimal population individuals, thereby enhancing the algorithm's accuracy.

2.3.2 Improved Sparrow Search Algorithm

LHS is a statistical method designed to generate random samples in multi-dimensional spaces[28]. The objective is to achieve a uniform coverage of the entire sample space with a more even distribution, thereby enhancing the efficiency of the sampling process, particularly in high-dimensional problems. Currently, LHS has been widely adopted in various fields including computer experiments, numerical simulations, and optimization. The standard SSA employs a random initialization method for population particles, which may lead to severe clustering and poor uniformity of initial individuals. In this paper, the LHS method is employed for initial population sampling to ensure a uniform distribution within the search space, thus enhancing the algorithm's global search capability.

However, since the iterative update of food positions is directed by the optimal individual and the population tends to move toward the optimal position, population diversity may gradually decrease in the later stages of iteration. The population convergence can lead the algorithm to become trapped in local optima. To address this issue, a Cauchy mutation perturbation operator is introduced in this paper. This operator applied Cauchy perturbations to the optimal sparrow individuals within the population, thereby expanding the search space of the algorithm and increasing the diversity of the population. The objective of this enhancement is to improve the global search capability of the SSA algorithm.

The Cauchy mutation is derived from the Cauchy distribution, which is a classical continuous probability distribution. The probability density function curve of the Cauchy distribution is illustrated in Figure 2. Compared to the standard normal distribution, it has a lower peak at the origin and a slower decline from the peak to zero. This yields a more uniform mutation range and enhances perturbation capability. By incorporating the Cauchy mutation to perturb the optimal individual, it could enable the algorithm to escape the confines of the local optimal solution. The model for introducing the Cauchy mutation into the optimal individual is as follows:

$$X^t_{\text{new\_best}} = X^t_{\text{best}} \cdot (1 + \text{Cauchy}(0,1)) \tag{5}$$

where, X^t_{best} is the optimal sparrow individual; $X^t_{\text{new\_best}}$ is the optimal sparrow individual after mutation; and Cauchy(0,1) is the standard Cauchy distribution. The random variable of the Cauchy distribution is generated by $\tan[(u-0.5)\pi]$ function, where u is the random number between [0,1].

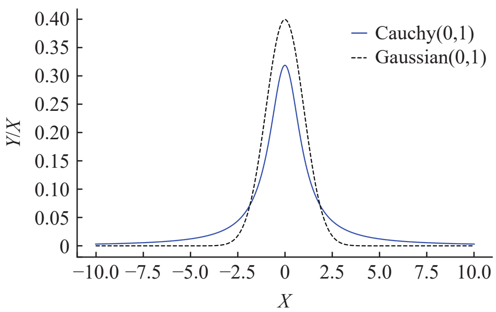


Figure 2 The probability density function curve

To ensure that the mutated optimal solution is superior to the original one, the greedy rule is adopted. By comparing the fitness values of the optimal solution prior to and subsequent to the mutation, the mathematical formulation of the greedy strategy is expressed as follows:

$$X^t_{\text{best}} = \begin{cases} X^t_{\text{best}}, & f(X^t_{\text{new\_best}}) \geq f\left(X^t_{\text{best}}\right) \\ X^t_{\text{new\_best}}, & f(X^t_{\text{new\_best}}) < f\left(X^t_{\text{best}}\right) \end{cases} \tag{6}$$

where, $f(X^t_{\text{best}})$ is the fitness value of the optimal individual sparrow; and $f(X^t_{\text{new\_best}})$ is the fitness value of the individual sparrow after mutation.

In conclusion, this paper addressed the limitations of the standard SSA by improving its population initialization and optimal individual selection. The specific steps of the ISSA are as follows:

Step 1: Initialize the population using the LHS method, and define the maximum number of iterations, and other relevant parameters.

Step 2: Calculate the individual fitness values within the population and arrange in ascending order to identify the initial optimal and worst fitness values, and their respective positions. The sparrow with the highest fitness value is designated as the current food location, with its position corresponding to the hyper-parameter of the CatBoost model.

Step 3: Introduce a Cauchy perturbation operator and update the global optimal solution according to Equation (5). Then compare the fitness values of the perturbed optimal individual and the original optimal individual using a greedy selection criterion. If the fitness value of the mutated individual is lower than that of the original optimal individual, the current optimal individual is replaced by the mutated one; otherwise the original optimal individual is retained.

Step 4: Designate the top 20% of sparrows in the sorted list as discoverers, and the remaining 80% as participants. The positions of all individuals are updated using Equations (2) and (3), respectively.

Step 5: Select a random subset of 10% of sparrows from the total population as vigilant and update their individual positions according to Equation (4).

Step 6: Repeat steps 2 to 5 iteratively until the maximum number of iterations is reached, and the global optimal solution is thereby obtained.

2.4 SMOTETomek-ISSA-CatBoost model

The CatBoost model has been widely employed in classification problems. In this study, the CatBoost model was applied to predict the category of the temperature control modes within a greenhouse. To achieve an enhanced prediction effect, an improved Sparrow Search Algorithm was proposed to optimize CatBoost hyper-parameters including Iterations, Depth, Learning_rate, and L2_leaf_reg. Five-fold cross-validation was adopted to evaluate model performance and determine the optimal parameter combination, thereby mitigating the risk of overfitting. The original training set was randomly partitioned into five equal subsets. One subset was designated as the validation data for the model, while the remaining four subsets were used as the training data. Subsequently, each subset was validated exactly once in rotation, with the F1-score on the validation set being documented. In each iteration of the ISSA, the negative mean of the five F1-score values was utilized as the fitness value. The cross-validation strategy was employed to conduct repeated experiments for each parameter combination of the classifier, aiming to identify the optimal parameter set based on the average F1-score. The sparrow individual exhibiting the lowest fitness value was selected as the optimal parameter configuration for the CatBoost model.

The algorithm flow of temperature control mode prediction method in a greenhouse based on the SMOTETomek-ISSA-

CatBoost model is shown in Figure 3. The specific steps are as follows:

Step 1: Collect the real-time data using the team's independent development platform. The raw data is subjected to preprocessing, which encompasses the identification and elimination of anomalies, the imputation of missing values, and the normalization of the dataset. The dataset is split into training and testing sets, with the first 70% of samples designated as the training set and the last 30% as the testing set.

Step 2: Expand the training set with the SMOTETomek mixed sampling method and divide the training set into five subsets to ensure no duplicate samples exist across different subsets.

Step 3: Take the negative value of the mean F1-score of the five-fold cross-validation as the individual fitness value of the sparrow. Through repetition of the iterations of ISSA, the optimal parameter set of the CatBoost model is constantly updated.

Step 4: Upon reaching the maximum number of iterations, the optimal solution of the CatBoost model parameters is obtained; otherwise, repeat Step 3.

Step 5: Construct the SMOTETomek-ISSA-CatBoost model based on the optimal parameters to obtain the optimal predicted value of the category of the greenhouse temperature control mode.

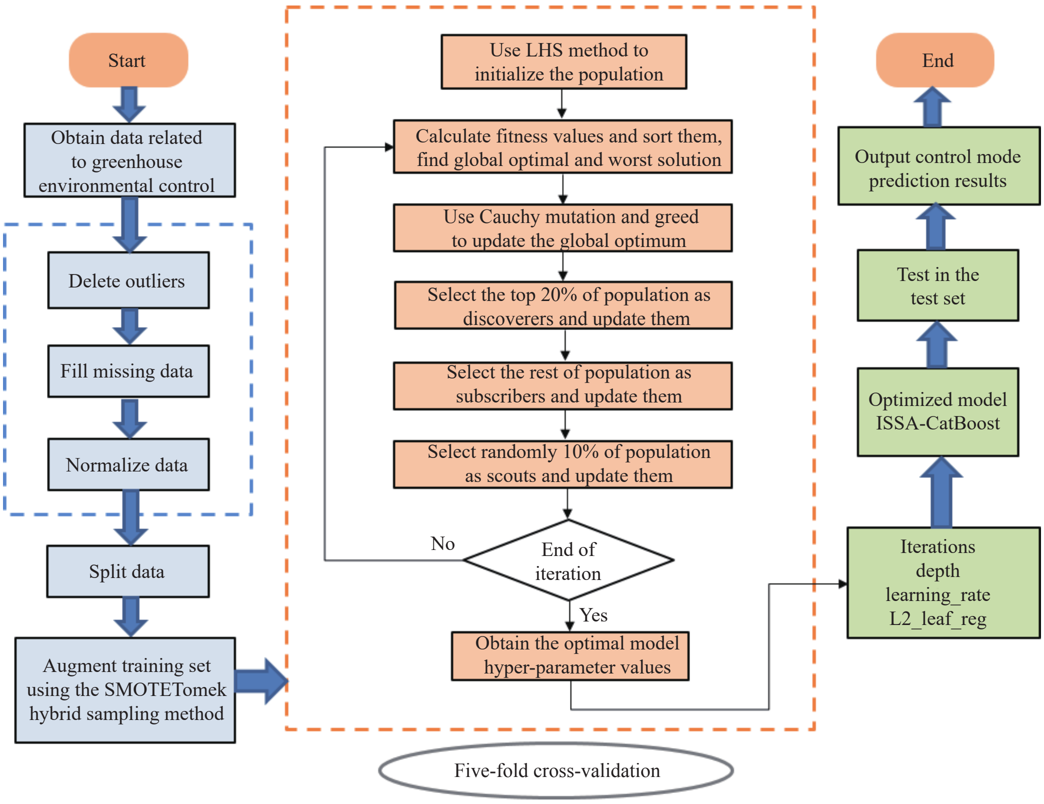


Figure 3 The flow of SMOTETomek-ISSA-CatBoost model

2.5 Evaluation index

To validate the prediction performance for greenhouse temperature control mode classification, the five evaluation indices of accuracy, recall, precision, F1-score, and area under the Receiver Operating Characteristic (ROC) curve (AUC) are selected to comprehensively evaluate the prediction results. The confusion matrix is presented in Table 2, and the equations are defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{F1-score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{10}$$

where, Accuracy denotes the proportion of instances correctly predicted by the model; Recall represents the proportion of positive instances that are correctly identified; Precision indicates the proportion of positive instances predicted correctly; F1-score is the harmonic mean of Precision and Recall, which comprehensively considers both indicators and avoids the bias caused by a single metric.

Table 2 Confusion matrix

Real situation	Predicted result	
	Predicted positive	Predicted negative
Real positive	TP	FN
Real negative	FP	TN

The AUC represents the area under the ROC curve, and the value is between 0 and 1. This provides a direct assessment of the classifier's performance, with higher values indicating better performance. The ROC curve is constructed based on true positive rate (TPR) and false positive rate (FPR). In the ROC curve, the horizontal axis represents the FPR, which is the proportion of negative instances incorrectly classified as positive among all negative instances. The vertical axis represents the TPR, which is the proportion of positive instances correctly classified among all positive instances.

$$\text{FPR} = \frac{FP}{TP+FP} \tag{11}$$

$$\text{TPR} = \frac{TP}{TP+FN} \tag{12}$$

In datasets characterized by an imbalanced class distribution, the accuracy, precision, and recall metrics may not fully assess the model's overall performance. In particular, when a model performs well in predicting the majority class, these metrics may be misleading. Therefore, this paper used ROC curves and AUC values, which are highly reliable measures, to evaluate the model's overall performance. Although ROC curves and AUC values are typically stable indicators of a model's overall performance, they may not fully capture the model's predictive capability for the minority class under conditions of extreme class imbalance. To more comprehensively assess the model's performance, this paper also incorporated the F1-score as an evaluation metric. The F1-score, by considering both precision and recall, effectively reflects the model's predictive ability for minority classes. Consequently, this paper primarily relied on a combined assessment of the F1-score and AUC value to select the optimal model, thereby guaranteeing the predictive accuracy and stability of the model on imbalanced datasets.

3 Results and analysis

To enhance the predictive performance of the model, this study employed an optimization strategy encompassing two principal dimensions: data preprocessing and methodological advancement. Furthermore, to comprehensively evaluate the performance of the proposed model, five evaluation metrics were utilized for testing, and four types of comparative analysis were conducted. Firstly, the basic classification model used in this study, CatBoost, was compared with several classical machine learning classifiers to evaluate its performance in classification tasks. Secondly, the SMOTETomek mixed sampling method was compared with other commonly used sampling methods to validate the effectiveness of the data processing approach. Thirdly, the ISSA was compared with other swarm intelligence optimization algorithms to optimize the performance of the CatBoost model. Finally, the superiority of the proposed model improvement was validated through ablation experiment analysis.

3.1 Comparison of different machine learning models

Before exploring the effectiveness of the proposed method, five classical classification algorithms were empirically analyzed, namely Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, utilizing the identical test set. The selection of these benchmark models was based on their sufficient representativeness and comprehensiveness, which can fully cover the mainstream solutions for this type of classification problem. Specifically, SVM is a classic supervised learning algorithm with strong generalization ability and excellent performance in small sample and high dimensional data classification tasks. RF, as a typical ensemble learning model, has the advantages of anti-overfitting, strong robustness, and ability to handle non-linear relationships, and is also one of the most commonly used benchmark models in classification tasks[29]. In addition, XGBoost, LightGBM, and CatBoost, as representative algorithms of the gradient boosting decision tree (GBDT) family, have outstanding performance in handling complex data and improving classification accuracy, and are mainstream solutions for current classification problems[30]. Collectively, these five algorithms cover different types of classical classification models (single models and ensemble models), and their wide application in related research ensures the rationality and comprehensiveness of the benchmark model selection, laying a solid foundation for the subsequent comparative analysis of the proposed method. The specific forecast results and their evaluation indicators are detailed in Table 3. In order to facilitate a visual comparison of each algorithm, all results are presented in the form of bar chart in Figure 4 and ROC curve in Figure 5.

Table 3 Prediction error evaluation indicators of different models

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM	0.8920	0.7723	**0.7468**	0.7255	0.9588
RF	0.8732	0.7689	0.6782	0.7156	0.9565
XGBoost	0.9031	0.7962	0.6743	0.7108	0.9462
LightGBM	0.9024	0.7957	0.6580	0.6867	0.9458
CatBoost	**0.9087**	**0.8523**	0.7103	**0.7612**	**0.9614**

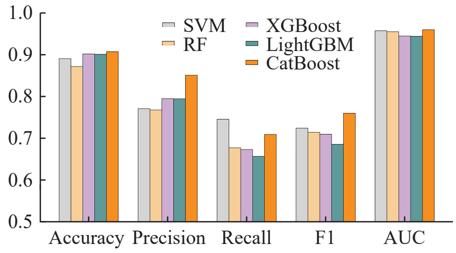


Figure 4 Five model performance comparison histograms

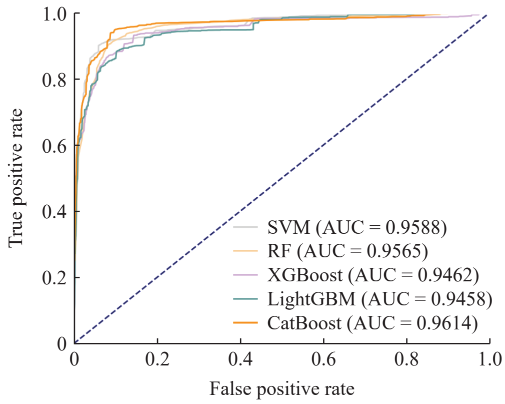


Figure 5 ROC curves of five models

As illustrated in Table 3 and Figure 4, the CatBoost model exhibited the highest accuracy in terms of the Accuracy index, reaching 0.9087, among all the evaluated basic classification models. The accuracy of SVM, RF, XGBoost, and LightGBM models were 0.8920, 0.8732, 0.9031, and 0.9024, respectively. With regard to precision, the CatBoost model continued to demonstrate superior performance in comparison to the other four models, attaining a value of 0.8523. While the CatBoost model demonstrated a slight deficit in recall relative to the SVM model, it nevertheless exhibited superior efficacy compared to the remaining three basic classification models.

F1-score correlates with Accuracy rate and Recall rate. As shown in Table 3 and Figure 4, the CatBoost model attained the highest F1-score among the comparative models, reaching 0.7612. In comparison, the F1-scores of SVM, RF, XGBoost, and LightGBM were 0.7255, 0.7156, 0.7108, and 0.6867, respectively. The RF model exhibited the lowest F1-score, while the CatBoost model demonstrated a significantly superior performance compared to the other models. In terms of the AUC index, the AUC values for SVM, RF, XGBoost, LightGBM, and CatBoost models were 0.9588, 0.9565, 0.9462, 0.9458, and 0.9614, respectively. The AUC value of the CatBoost model was the highest (Figure 5).

In conclusion, although the CatBoost model exhibited slightly inferior performance in terms of Recall compared to the SVM, it achieved superior performance in other key indicators, including the crucial F1-score and AUC. This superior comprehensive performance is rooted in the inherent advantages of CatBoost, which effectively mitigates the impact of category feature bias through ordered boosting and adaptive learning rate adjustment. These inherent advantages enable the CatBoost model to better capture the complex non-linear relationships between input features and temperature control modes in greenhouses, thereby demonstrating superior classification performance compared to other conventional classification models such as SVM. Moreover, these findings were consistent with those of previous studies[31], providing further evidence for the selection of CatBoost as the fundamental model for predicting greenhouse temperature control modes in this study.

3.2 Data balance processing

The initial training dataset, as presented in this paper, comprises a total of 3023 samples. The sample sizes corresponding to the static, natural ventilation, mechanical ventilation, and fan-wet curtain modes are 225, 2553, 233, and 12, respectively. From the sample distribution, it can be observed that the training set employed in this study constitutes a typical imbalanced dataset.

The prevalence of data imbalance in classification problems represents a significant factor contributing to inefficient model training and suboptimal prediction performance[32]. To address this class imbalance issue, the data sampling method is typically used to balance the class distribution of the dataset prior to the model training. Currently, several methods are employed at the dataset level for data balancing, including undersampling, oversampling, and hybrid sampling methods. To improve the performance and generalization ability of the proposed model, this study employed several resampling techniques: undersampling methods including Edited Nearest Neighbors (ENN) and Repeated Edited Nearest Neighbors (RENN); oversampling methods including Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Technique (SMOTE); and mixed sampling methods including SMOTEENN and SMOTETomek[33]. The six sampling methods were applied to the original training set in order to achieve data balancing, with the results presented in Table 4.

Table 4 Six different sampling methods for imbalanced dataset

Type	Number of training set	Static	Natural ventilation	Mechanical ventilation	Fan-wet curtain
Base	3023	225	2553	233	12
ENN	2963	202	2542	207	12
RENN	2895	193	2501	189	12
ADASYN	10 230	2554	2553	2571	2552
SMOTE	10 212	2553	2553	2553	2553
SMOTEENN	10 097	2551	2441	2552	2553
SMOTETomek	10 212	2553	2553	2553	2553

To evaluate the effects of different data balancing processing methods, based on the same test set, an empirical analysis was conducted on the category prediction model for temperature control mode, utilizing the CatBoost model. The specific evaluation results are listed in Table 5. As can be observed from the table, the undersampling method resulted in the removal of only a small number of samples from the majority of classes, with a notable decline in the overall classification efficacy, particularly evident in the F1-score. The F1-score for the RENN-CatBoost method was only 0.7017. Conversely, the oversampling method exhibited superior overall classification efficacy compared to the undersampling method, particularly in terms of Recall and F1-score, which demonstrated notable enhancement. However, due to the random repeated sampling process, there was a possibility of overfitting. Mixed sampling method had been demonstrated to outperform the single undersampling or oversampling method on multiple indicators, particularly in terms of Precision, Recall, and F1-score. Among them, the F1-score of SMOTETomek-CatBoost reached 0.8052, and the Precision and Recall were significantly higher than those of the other five models. Although the AUC value was slightly lower than that of the CatBoost model, with a difference of only 0.0018, this difference had a minimal impact on the overall classification effect.

Table 5 Comparison of evaluation indices for improving methods of imbalanced data

Type	Model	Accuracy	Precision	Recall	F1-score	AUC
Base	CatBoost	0.9087	0.8523	0.7103	0.7612	**0.9614**
Under sampling	ENN-CatBoost	0.8941	0.8222	0.7004	0.7492	0.9597
	RENN-CatBoost	0.9031	0.7699	0.6600	0.7017	0.9607
Over sampling	ADASYN-CatBoost	**0.9164**	0.7898	0.7463	0.7665	0.9549
	SMOTE-CatBoost	0.9129	0.8161	0.7468	0.7766	0.9586
Mixed sampling	SMOTEENN-CatBoost	0.9087	0.8215	0.7645	0.7897	0.9578
	SMOTETomek-CatBoost	0.9136	**0.8653**	**0.7646**	**0.8052**	0.9596

In general, SMOTETomek demonstrates obvious advantages in terms of comprehensive performance. In light of the above conclusion, the mixed sampling method SMOTETomek was selected in this study to carry out data balance processing on the original training set. The expanded data was then employed as the training dataset for the subsequent category prediction model of temperature control mode.

3.3 Comparison of model parameter optimization algorithms

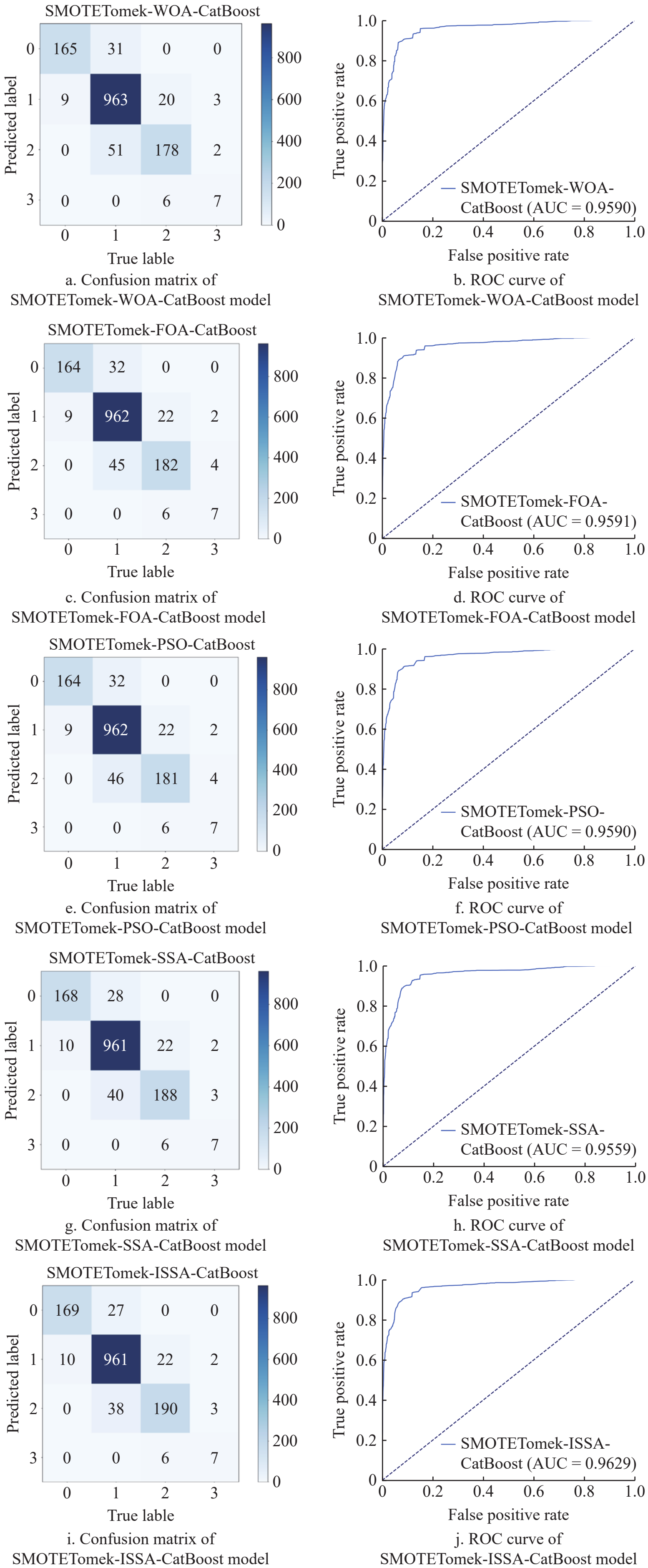
To further improve the classification performance of the CatBoost model, this paper proposed to use the ISSA to optimize the hyper-parameters of CatBoost model. Meanwhile, the ISSA was compared with four other classical swarm intelligence optimization algorithms (WOA, FOA, PSO, and standard SSA). All algorithms were evaluated and compared under the same experimental conditions.

The hyper-parameters of the optimized CatBoost model presented in this paper were the number of algorithm iterations (Iterations), the depth of the tree (Depth), the learning rate (Learning_rate), and the regular sub-parameter (L2_leaf_reg). Table 6 illustrates the range of values for each parameter. The value of Iterations ranged from 100 to 2000, while Depth spanned from 3 to10. Learning_rate varied from 0.01 to 0.3, and the L2_leaf_reg extended from 1 to 10.

Table 6 Hyper-parameter selection range of CatBoost model

Hyper-parameter name	Scope
Iterations	(100, 2000)
Depth	(3, 10)
Learning_rate	(0.01, 0.3)
L2_leaf_reg	(1, 10)

The optimal model hyper-parameters (Table 7) obtained by each optimization algorithm were used to predict the category of temperature control mode within a greenhouse in the test set. Table 8 presents five evaluation metrics of ISSA and other optimization algorithms on the test set. To more intuitively compare the performance of each method, Figure 6 shows the confusion matrix and ROC curves for each model.



a. Confusion matrix of SMOTETomek-WOA-CatBoost model

b. ROC curve of SMOTETomek-WOA-CatBoost model

c. Confusion matrix of SMOTETomek-FOA-CatBoost model

d. ROC curve of SMOTETomek-FOA-CatBoost model

e. Confusion matrix of SMOTETomek-PSO-CatBoost model

f. ROC curve of SMOTETomek-PSO-CatBoost model

g. Confusion matrix of SMOTETomek-SSA-CatBoost model

h. ROC curve of SMOTETomek-SSA-CatBoost model

i. Confusion matrix of SMOTETomek-ISSA-CatBoost model

j. ROC curve of SMOTETomek-ISSA-CatBoost model

Figure 6 Confusion matrix and ROC curve of five models

As illustrated in Table 8 and Figure 6, the SMOTETomek-SSA-CatBoost model exhibited notable Accuracy, Precision, Recall, and F1-score, reaching 0.9226, 0.8329, 0.7938, and 0.8123, respectively. The accuracy, precision, recall, and F1-score of the SMOTETomek-SSA-CatBoost model were superior to those of the SMOTETomek-PSO-CatBoost, SMOTETomek-FOA-CatBoost, and SMOTETomek-WOA-CatBoost models. This indicated that the SSA algorithm had superior capacity for optimizing parameters in comparison to WOA, FOA, and PSO.

Table 7 Optimal solution of CatBoost model hyper-parameters under different optimization algorithms

Model	Iterations	Depth	Learning_rate	L2_leaf_reg
SMOTETomek-WOA-CatBoost	1591	5	0.2749	3.6123
SMOTETomek-FOA-CatBoost	1705	4	0.3000	1.0000
SMOTETomek-PSO-CatBoost	2000	4	0.3000	1.0000
SMOTETomek-SSA-CatBoost	1932	3	0.2899	1.2167
SMOTETomek-ISSA-CatBoost	1649	4	0.2368	6.0670

Table 8 Comparison of evaluation indices of different model parameter optimization algorithms

Model	Accuracy	Precision	Recall	F1-score	AUC
SMOTETomek-WOA-CatBoost	0.9150	0.8314	0.7797	0.8036	0.9590
SMOTETomek-FOA-CatBoost	0.9164	0.8197	0.7825	0.7997	0.9591
SMOTETomek-PSO-CatBoost	0.9157	0.8194	0.7814	0.7989	0.9590
SMOTETomek-SSA-CatBoost	0.9226	0.8329	0.7938	0.8123	0.9559
SMOTETomek-ISSA-CatBoost	**0.9247**	**0.8339**	**0.7973**	**0.8147**	**0.9629**

Meanwhile, a comprehensive analysis of the results across all metrics indicated that the proposed SMOTETomek-ISSA-CatBoost model achieved the best performance in terms of Accuracy, Precision, Recall, F1-score, and AUC. Specifically, the SMOTETomek-ISSA-CatBoost model achieved an Accuracy of 0.9247 and a Precision of 0.8339, both of which surpassed those attained by the other four parameter optimization algorithms. In terms of Recall, the SMOTETomek-ISSA-CatBoost model achieved a value of 0.7973, which was significantly higher than those of the other four parameter optimization algorithms. Among these algorithms, SMOTETomek-WOA-CatBoost performed the worst, with a Recall of only 0.7797. Regarding the F1-score, the SMOTETomek-ISSA-CatBoost model had a notable increase compared to the SMOTETomek-SSA-CatBoost model, and had an increase of 0.0158, 0.015, and 0.0111 when compared with SMOTETomek-PSO-CatBoost, SMOTETomek-FOA-CatBoost, and SMOTETomek-WOA-CatBoost, respectively. In terms of the AUC index, the SMOTETomek-ISSA-CatBoost model achieved a value of 0.9629, which was significantly higher than that of the SMOTETomek-SSA-CatBoost model, and slightly higher than those of the other three parameter optimization algorithms.

In conclusion, the ISSA proposed in this paper had been further improved in terms of global search ability, thereby achieving superior optimization results in comparison to the standard sparrow search algorithm.

3.4 Ablation experiment analysis

To further verify the effectiveness of the model proposed in this paper, SMOTETomek-ISSA-CatBoost was compared with three other ablation models. The results of the ablation test are presented in Table 9. Compared with SMOTETomek-CatBoost, the Accuracy, Recall, F1-score, and AUC of SMOTETomek-ISSA-CatBoost increased by 0.0111, 0.0327, 0.0095, and 0.0033, respectively. Compared with ISSA-CatBoost, the Accuracy, Recall, F1-score, and AUC increased by 0.0111, 0.0619, 0.0348, and 0.001, respectively. Among them, the SMOTETomek-ISSA-CatBoost model outperformed the CatBoost model with the most significant performance improvement. The above test results demonstrated that the combined application of the SMOTETomek and ISSA methods effectively compensated for the limitations of a single model in sample distribution and parameter optimization. This approach not only significantly enhanced the prediction accuracy and generalization capability of environment control mode in the greenhouse, but also validated the rationality and effectiveness of the improvement strategies proposed in this paper.

Table 9 Ablation experiment results

Model	Accuracy	Precision	Recall	F1-score	AUC
CatBoost	0.9087	**0.8523**	0.7103	0.7612	0.9614
ISSA-CatBoost	0.9136	0.8436	0.7354	0.7799	0.9619
SMOTETomek-CatBoost	0.9136	0.8653	0.7646	0.8052	0.9596
SMOTETomek-ISSA-CatBoost	**0.9247**	0.8339	**0.7973**	**0.8147**	**0.9629**

4 Conclusions

In this paper, an improved Sparrow Search Algorithm was proposed to optimize the parameters of the CatBoost model. The standard SSA was enhanced primarily in two key areas: Initially, the LHS method was employed to sample the initial population, thereby ensuring an even distribution across the search space. This approach enhanced the algorithm's global search capability. Furthermore, by employing the Cauchy mutation operator and the greedy rule, the optimal individual from the population of sparrows was subjected to Cauchy perturbation. This expansion of the search space improved the diversity of the population, thereby further strengthening the global search performance of the SSA algorithm. Furthermore, the initial training set was expanded based on the SMOTETomek mixed sampling method, and the hyper-parameters of the CatBoost model were optimized using the ISSA algorithm, then verified on the test set.

This paper explored the research from four primary aspects. Firstly, the standard CatBoost model was evaluated in comparison with various commonly used machine learning models, including RF, SVM, XGBoost, and LightGBM, to evaluate the rationale behind the selection of the basic classification models. Secondly, the SMOTETomek mixed sampling method was employed to expand the training set, thereby alleviating the issue of class imbalance within the dataset. Furthermore, ISSA was compared with four other common optimization algorithms (WOA, FOA, PSO, and standard SSA) in optimizing CatBoost model performance. Finally, the ablation test analysis demonstrated the superiority of the method proposed in this paper.

During the training phase, the training set was used for five-fold cross validation to adjust the hyper-parameters of the CatBoost model. Once the optimization process was completed, the performance of each model was evaluated using an independent test set. The results demonstrated that the SMOTETomek-ISSA-CatBoost model exhibited robust prediction capability in the category prediction of temperature control modes in the greenhouse, and had obvious advantages compared with other traditional classification models and optimization algorithms.

This study presents a data-driven solution for cooling control in greenhouses during the summer season. Conventional control methods such as PID and MPC suffer from high computational complexity, high deployment costs, and excessive reliance on precise physical models. In contrast, the method proposed in this paper learns the internal operating mechanism of greenhouses from agricultural expert experience, equipment operation rules, and empirical data in greenhouses via machine learning. It features lower computing overhead, as well as stronger adaptability to nonlinear and time-varying greenhouse environments, thus highlighting obvious advantages in agricultural production applications. The study proposed in this paper is of great significance in the field of greenhouse environmental control, and the proposed method can be further extended to various types of greenhouse in the future, including solar greenhouses, film greenhouses, and others.

However, this study focused solely on temperature control for greenhouse cooling in the summer and did not include experimental validation for other seasonal types, such as winter heating. In addition, the dataset was only collected from a single month, which limited the generalizability of the proposed method. In future research, collecting high-quality expert data under a broader range of climatic scenarios will be considered. This involves integrating superior control data for winter heating, spring and autumn insulation, and summer cooling. The objective is to develop a temperature control method within greenhouses suitable for the full growth cycle of crops.

Acknowledgements

This work was supported by Science and Technology Project of the Ministry of Agriculture and Rural Affairs, Jiangsu Provincial Key Research and Development Program (Grant No. BE2021379), Key Laboratory of Smart Agricultural Technology (Yangtze River Delta) Ministry of Agriculture and Rural Affairs Open Fund (Grant No. KSAT-YRD2024002), and Jiangsu Agricultural Science and Technology Independent Innovation Fund (Grant No. CX(22)3113).

[References]

[1] Moore C E, Meacham-Hensold K, Lemonnier P, Slattery R A, Benjamin C, Bernacchi C J, et al. The effect of increasing temperature on crop photosynthesis: From enzymes to ecosystems. Journal of Experimental Botany, 2021; 72(8): 2822–2844.

[2] Omid M, Shafaei A. Temperature and relative humidity changes inside greenhouse. International Agrophysics, 2005; 19(2): 153–158.

[3] Pinho P, Hytonen T, Rantanen M, Elomaa P, Halonen L. Dynamic control of supplemental lighting intensity in a greenhouse environment. Lighting Research & Technology, 2013; 45(3): 295–304.

[4] Pasgianos G D, Arvanitis K G, Polycarpou P, Sigrimis N. A nonlinear feedback technique for greenhouse environmental control. Computers and Electronics in Agriculture, 2003; 40: 153–177.

[5] Espinoza K, Valera D L, Torres J A, Lopez A, Molina-Aiz F D. An auto-tuning PI control system for an open-circuit low-speed wind tunnel designed for greenhouse technology. Sensors, 2015; 15(8): 19723–19749.

[6] Blasco X, Martínez M, Herrero J M, Ramos C, Sanchís J. Model-based predictive control of greenhouse climate for reducing energy and water consumption. Computers and Electronics in Agriculture, 2007; 55(1): 49–70.

[7] Liang M H, Chen L J, He Y F, Du S F. Greenhouse temperature predictive control for energy saving using switch actuators. IFAC PapersOnLine, 2018; 51(17): 747–751.

[8] Moreno J C, Berenguel M, Rodríguez F, Baños A. Robust control of greenhouse climate exploiting measurable disturbances. IFAC Proceedings Volumes, 2002; 35(1): 271–276.

[9] Ariffin M A M, Ramli M I, Amin M N M, Ismail M, Zainol Z, Ahmad N D, et al. Automatic climate control for mushroom cultivation using IoT approach. 10th International Conference on System Engineering and Technology (ICSET), Shah Alam, Malaysia: IEEE, 2020; pp.123–128.

[10] Montoya-Ríos A P, García-Mañas F, Guzmán J L, Rodríguez F. Simple tuning rules for feedforward compensators applied to greenhouse daytime temperature control using natural ventilation. Agronomy, 2020; 10(9): 1327.

[11] Azaza M, Tanougast C, Fabrizio R, Mami A. Smart greenhouse fuzzy logic based control system enhanced with wireless data monitoring. ISA Transactions, 2016; 61: 297–307.

[12] Lachouri C E, Mansouri K, Lafifi M M, Belmeguenai A. Adaptive neuro-fuzzy inference systems for modeling greenhouse climate. Int J of Advanced Computer Science and Applications, 2016; 7(1): 96–100.

[13] Pezeshki Z, Mazinani S M. Comparison of artificial neural networks, fuzzy logic and neuro fuzzy for predicting optimization of building thermal consumption: a survey. Artificial Intelligence Review, 2018; 52(1): 495–525.

[14] Wang L, Zhang H H. An adaptive fuzzy hierarchical control for maintaining solar greenhouse temperature. Computers and Electronics in Agriculture, 2018; 155: 251–256.

[15] Lin D, Zhang L J, Xia X H. Hierarchical model predictive control of Venlo-type greenhouse climate for improving energy efficiency and reducing operating cost. Journal of Cleaner Production, 2020; 264: 121513.

[16] Chen W H, You F Q. Semiclosed greenhouse climate control under uncertainty via machine learning and data-driven robust model predictive control. IEEE Transactions on Control Systems Technology, 2021; 30(3): 1186–1197.

[17] Chen L J, Du S F, He Y F, Liang M H, Xu D. Robust model predictive control for greenhouse temperature based on particle swarm optimization. Information Processing in Agriculture, 2018; 5(3): 329–338.

[18] Mahmood F, Govindan R, Bermak A, Yang D, Khadra C, Al-Ansari T. Energy utilization assessment of a semi-closed greenhouse using data-driven model predictive control. Journal of Cleaner Production, 2021; 324: 129172.

[19] Jung D H, Kim H J, Kim J Y, Lee T S, Park S H. Model predictive control via output feedback neural network for improved multi-window greenhouse ventilation control. Sensors, 2020; 20(6): 1756.

[20] Mahmood F, Govindan R, Bermak A, Yang D, Al-Ansari T. Data-driven robust model predictive control for greenhouse temperature control and energy utilisation assessment. Applied Energy, 2023; 343: 121190.

[21] Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V, Gulin A. Catboost: unbiased boosting with categorical features. In: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2018; pp.6639–6649.

[22] Chang W F, Wang X, Yang J, Qin T. An improved CatBoost-based classification model for ecological suitability of blueberries. Sensors, 2023; 23: 1811.

[23] Ge Z W, Feng S, Ma C C, Wei K, Hu K, Zhang W J, et al. Quantifying and comparing the effects of key chemical descriptors on metal-organic frameworks water stability with CatBoost and SHAP. Microchemical Journal, 2024; 196: 109625.

[24] Li L B, Qiao J D, Yu G, Wang L Z, Li H Y, Liao C, et al. Interpretable tree-based ensemble model for predicting beach water quality. Water Research, 2022; 211: 118078.

[25] Nasir N, Kansal A, Alshaltone O, Barneih F, Sameer M, Shanableh A, et al. Water quality classification using machine learning algorithms. Journal of Water Process Engineering, 2022; 48: 102920.

[26] Jia Y, Su Y J, Zhang R, Zhang Z N, Lu Y K, Shi D X, et al. Optimization of an extreme learning machine model with the sparrow search algorithm to estimate spring maize evapotranspiration with film mulching in the semiarid regions of China. Computers and Electronics in Agriculture, 2022; 201: 107298.

[27] Xue J K, Shen B. A novel swarm intelligence optimization approach: Sparrow search algorithm. Systems Science & Control Engineering, 2020; 8(1): 22–34.

[28] Heltona J C, Davis F J. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering and System Safety, 2003; 81(1): 23–69.

[29] Bulut C, Arslan E. Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment. Artificial Intelligence Review, 2024; 57: 252.

[30] Long L H, Shi Q L, Zhang Q J, Hu J D, Zhang H M. Dual-warning model for coal spontaneous combustion temperature prediction and risk classification based on BO-LightGBM. Process Safety and Environmental Protection, 2025; 201: 107624.

[31] Wei X, Rao C J, Xiao X P, Chen L, Goh M. Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. Expert Systems With Applications, 2023; 219: 119648.

[32] Coser A, Maer-matei M M, Albu C. Predictive models for loan default risk assessment. Economic Computation and Economic Cybernetics Studies and Research, 2019; 53(2): 149–165.

[33] Kotb M H, Ming R. Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models. International Journal of Advanced Computer Science and Applications, 2021; 12(9): 621–629.