sustainability

MDPI

Article

Environmental Prediction Model of Solar Greenhouse Based on Improved Harris Hawks Optimization-CatBoost

Jie Yang, Guihong Ren, Yaxin Wang, Qi Liu, Jiamin Zhang, Wenqi Wang, Lingzhi Li and Wuping Zhang *

School of Software, Shanxi Agricultural University, Jinzhong 030810, China; yangjie20210228@126.com (J.Y.); endeavorren@163.com (G.R.); wangyx1919@126.com (Y.W.); 297341853@163.com (Q.L.); zjm817107@163.com (J.Z.); 18335493677@163.com (W.W.); lilz008@hotmail.com (L.L.)
* Correspondence: zwping@126.com

Abstract: Solar greenhouses provide a favorable climate environment for the production of counter-seasonal crops in northern China. The greenhouse environment is a key factor affecting crop growth, so accurate prediction of greenhouse environment changes helps to precisely regulate the crop growth environment and helps to promote the growth of fruits and vegetables. In this study, an environmental prediction model based on the combination of a gradient boosting tree and the Harris hawk optimization algorithm (IHHO-Catboost) is constructed, and in response to the problems of the HHO algorithm, such as the fact that the adjustment of the search process is not flexible enough, it cannot be targeted to carry out a stage search, and sometimes it will fall into the local optimum to make the algorithm's search accuracy relatively poor, an algorithm based on the improved Harris hawk optimization (IHHO) algorithm-based parameter identification method is constructed. The model considers the internal and external environmental and regulatory factors affecting crop growth, which include indoor temperature and humidity, light intensity, carbon dioxide concentration, soil temperature and humidity, outdoor temperature and humidity, light intensity, carbon dioxide concentration, wind direction, wind speed, and opening and closing of upper and lower air openings of the cotton quilt, and is input into a prediction model with a time series for training and testing. The experimental results show that the *MAE* (mean absolute error) values of temperature, relative humidity, carbon dioxide concentration, and light intensity of the model are reduced to 49.8%, 35.3%, 72.7%, and 32.1%, respectively, compared with LSTM (Long Short-Term Memory), which is a significant decrease in error. It shows that the proposed multi-parameter prediction model for solar greenhouse environments presents an effective method for accurate prediction of environmental data in solar greenhouses. The model not only improves prediction accuracy but also reduces dependence on large data volumes, reduces computational costs, and improves the transparency and interpretability of the model. Through this approach, an effective tool for greenhouse agriculture is provided to help farmers optimize the use of resources, reduce waste, and improve crop yield and quality, ultimately leading to a more efficient and environmentally friendly agricultural production system.

Keywords: solar greenhouse; IoT; gradient lifting tree; Harris Eagle optimization algorithm; environmental prediction model

Citation: Yang, J.; Ren, G.; Wang, Y.; Liu, Q.; Zhang, J.; Wang, W.; Li, L.; Zhang, W. Environmental Prediction Model of Solar Greenhouse Based on Improved Harris Hawks Optimization-CatBoost. *Sustainability* **2024**, *16*, 2021. https://doi.org/10.3390/su16052021

Academic Editors: Wei Liu and Chia-Huei Wu

Received: 27 December 2023
Revised: 20 February 2024
Accepted: 27 February 2024
Published: 29 February 2024

1. Introduction

In northern China, a solar greenhouse is an important place for winter fruit and vegetable production [1]. The growth of crops in greenhouses is affected by a variety of environmental parameters [2], and a suitable greenhouse environment promotes the healthy and efficient growth of greenhouse crops [3]; therefore, an accurate greenhouse environment prediction model is needed that can predict the trend of the greenhouse environment in advance [4] in order to avoid economic losses, improve crop yields and quality, and solve the problem of global food security [5].

In recent years, many prediction models have been proposed using different techniques to provide effective prediction models [6]. These techniques use statistical, physical,

and deep learning techniques for prediction [7]. Among the environmental factors are time series variables characterized by nonlinearity and time lags [8]. There are several mechanistic models, including 3D simulation models and physical models [9,10], that enable dynamic prediction of greenhouse environments [11]. However, mechanistic models are susceptible to boundary conditions and physical parameter definitions, among others. It is challenging to accurately reflect the actual greenhouse. With the development of sensor technology, the amount of data about the greenhouse environment is increasing. The use of data-driven time series prediction model construction is gaining attention [12].

Deep learning algorithms, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM), have been rapidly developed in recent years [13–15], providing powerful technical support for multivariate time series prediction. Researchers have proposed optimization approaches based on many different algorithms to improve the accuracy and performance of prediction models. Among them, the introduction of a convolutional neural network (CNN) combined with a gated recurrent unit neural network (GRU) achieved higher accuracy in multi-point temperature and humidity predictions of mushroom rooms, surpassing the traditional BP neural network, LSTM, and GRU models [16]. The model, using an adaptive global search algorithm combined with an LSTM recurrent neural network, can achieve high accuracy and fast operation [17]. The LSTM network optimized with the Sparrow Search Algorithm (SSA) was used to achieve accurate predictions of greenhouse environmental data and significantly improve the prediction accuracy [18]. A two-layer LSTM structure was used as an encoder and decoder to improve the characterization and performance of environmental parameter prediction models [19]. The introduction of a Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) architecture, which combines feature extraction and sequential dependency learning, has yielded good results for environmental predictions in greenhouse environments [20]. Each of these different algorithms has different features and advantages in greenhouse environment predictions. Principal component regression [21] provides high prediction accuracy, and BP neural networks and LSTM networks can achieve better performance under multi-feature conditions. The approach of combining different algorithms can further improve prediction accuracy. Although these algorithms have demonstrated significant advantages in multivariate time series forecasting, they also have key limitations. Including the reliance on large amounts of training data, high computational costs, and the opacity of the modeling decision-making process, traditional models show limitations in the face of complex and nonlinear greenhouse environmental data.

Huang Zihui et al. [22] used multi-strategy improvements to solve the overfitting and long-term dependence problems of neural networks, which accelerated the convergence speed and, at the same time, showed higher prediction accuracy and robustness, proving the applicability and generalization of the model. Hossein Moayedi et al. [23] investigated the fuzzy logic approach for predicting heat loads in residential buildings and proved the feasibility of the technique to predict heat loads. The results of the above studies show that HHO-based methods are gradually becoming state-of-the-art group intelligence optimization methods.

In this study, a mixed-strategy improved Harris hawk algorithm (IHHO) is proposed and combined with a short-term prediction model based on the CatBoost algorithm, and, finally, an environmental prediction model based on the combination of a gradient boosting tree and the Harris hawk optimization algorithm (IHHO-Catboost) is proposed. This is the first study to combine the improved Harris hawk optimization algorithm (IHHO) with the Catboost algorithm for the prediction model of daylight greenhouse environments. The model makes full use of the advantages of the IHHO algorithm in global search and optimization capability to improve the convergence speed and optimization efficiency of the model, and the prediction accuracy of the model when dealing with complex nonlinear data is significantly improved by the powerful classification capability and robustness to noisy data of the Catboost algorithm, which aims to improve the prediction accuracy,

reduce the dependence on large data volumes, and lower the computational cost while improving the transparency and interpretability of the model. The method has a better recognition ability for category-type features and reduces the requirement of hyperparameters, and the constructed features are screened using the SHAP (SHaplay Additive Interpretation) method, which further improves the prediction accuracy of the solar greenhouse environment. Finally, the superiority and reliability of the proposed model for daylight greenhouse environment prediction were verified by examples. With the growth of the global population and the decrease in arable land, improving the efficiency and sustainability of greenhouse crop production becomes the key to achieving global food security. By employing advanced prediction models, this study is expected to provide greenhouse agriculture with an effective tool to help farmers optimize resource use, reduce waste, and improve crop yields and quality. Ultimately, these efforts will help realize more efficient and environmentally friendly agricultural production systems that will contribute to meeting future food demands and environmental challenges. Technical support is provided for precise regulation of the greenhouse environment and early warning of damage.

2. Materials and Methods

2.1. Data Collection

This study was carried out in a solar greenhouse at the solar greenhouse base in Wanggantun Town, Yanggao County, Datong City, Shanxi Province (113°9′34″ E, 40°40′54″ N), in a solar greenhouse. To obtain indoor and outdoor environmental data, the experiment used an IoT data collection system, which consists of a meteorological monitoring host, a four-in-one Stevenson screen, soil sensors, and a small weather station. The collected data includes indoor and outdoor temperature and humidity, carbon dioxide levels, light intensity, outdoor rain and snow weather, wind speed, and wind direction.

The dimensions of the greenhouse structure involved in this study are 82 m in total length, 14 m in width, and 7.5 m in height. To ensure that the data collection adequately reflects the greenhouse environment, sensors should be evenly distributed throughout the greenhouse, covering multiple areas. The sensors are positioned at different heights or levels to provide comprehensive data. The overall system architecture and sensor layout are shown in Figure 1.

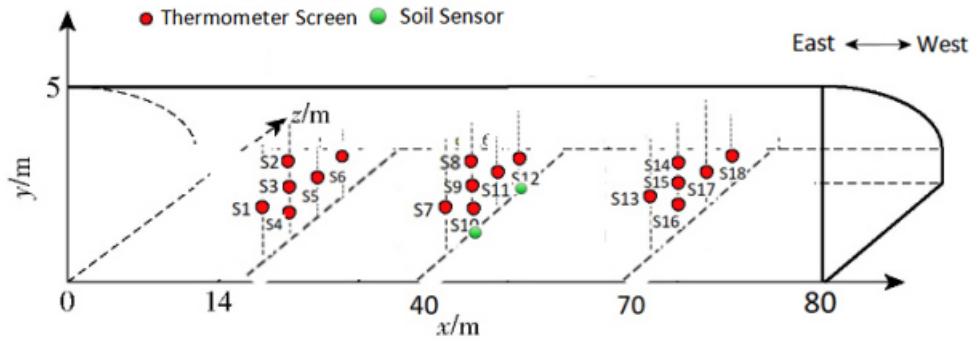

Figure 1. Map of sensor placement points.

The sensors in the study were arranged in three planes along the east–west direction at distances of 14 m, 40 m, and 70 m, with heights set at 1.5 m, 2.5 m, and 3 m. The sensor points were marked as S1–S18.

The sensor mainly uses an integrated louvered box, which can be widely applied to environmental testing, integrating CO_2, temperature and humidity, atmospheric pressure, and light. Installed in the louvered box, the device adopts the standard MODBUS-RTU communication protocol and RS485 signal output. The transmitter is widely used on a variety of occasions that need to measure environmental temperature and humidity, noise, air quality, atmospheric pressure, light, etc. It is safe, reliable, and durable.

During the test period, the motor controlled the opening and closing of the quilts and air vents, recorded the switching of the quilts with the upper and lower air vents, and automatically collected 15 kinds of data on indoor temperature and humidity, light intensity, carbon dioxide concentration, soil temperature, and humidity, along with the outdoor temperature and humidity, light intensity, carbon dioxide concentration, wind direction, and wind speed through the above collection system. The collection equipment uploaded the data to the monitoring and control cloud platform through RS485 communication to store and download the data. The sampling time was from 25 May 2023 to 15 August 2023, with the interval of collection time being 30 min, and a total of 59,760 pieces of data were collected. Given that environmental variables such as temperature and humidity in greenhouses do not change much over a relatively short period of time, a sampling frequency of 30 min is sufficient to capture key environmental dynamics and meet the needs of most greenhouse crop growth monitoring and environmental control. Therefore, the choice of 30 min as the sampling interval is a practical and well-considered option to effectively monitor the greenhouse environment while maintaining the economy and sustainability of the system.

2.2. Data Preprocessing

2.2.1. Missing Value Handling

During data collection, sensors can experience data loss due to equipment malfunctions and sudden power outages, which can significantly impact the accuracy of model predictions. In this study, the method mentioned in the literature [24] is used for handling missing values: when the span of missing time data is large, data with similar weather conditions are used to fill the gaps; when there are fewer missing values, linear interpolation is employed for filling in, thereby obtaining a complete dataset. The calculation formula for this process is as follows:

$$X_{a+i} = X_a + \frac{i(X_{a+i} - X_a)}{j}(0 < i < j) \tag{1}$$

In the formula, X_{a+i} represents the missing value at time $a + i$, while X_{a+j} represent the original data at times a and $a + j$, respectively.

2.2.2. Data Denoising Process

Unprocessed data may contain noise, errors, or incomplete information, which may lead to misleading results or inaccurate analysis. A Savitzky–Golay filter is a smoothing technique based on polynomial fitting [25]. Its basic principle is to fit a local polynomial to the signal and then replace the original signal with the fitted polynomial for smoothing.

2.2.3. Normalization Process

To eliminate the impact of different units of measurement on the prediction model, this study employs the Min–Max normalization method to standardize the data, thereby improving the model's convergence speed and prediction accuracy [26].

$$X_i^* = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

In the formula, X_i represents the i measured value; $X_{\max}$ represents the maximum value of the measurements; $X_{\min}$ represents the minimum value of the measurements; and X_i^* represents the normalized measured value.

2.3. Greenhouse Environment Prediction Model

2.3.1. IHHO-Optimized CatBoost

CatBoost, along with XGBoost and LightGBM, are the main algorithms of GBDT (Gradient Boosting Decision Tree). Compared to traditional GBDT, the model obtained from

XGBoost is simpler, LightGBM trains faster, and CatBoost offers higher accuracy. CatBoost requires fewer machine learning parameters and supports categorical variables with high precision, showing significant advantages in handling categorical data. Compared to other algorithms, it offers better accuracy and improves generalization ability [27]. In the prediction process, there might be gradient bias and prediction shift, leading to overfitting issues, which CatBoost can effectively address. Additionally, CatBoost reduces the need for extensive hyperparameter tuning, exhibiting high robustness [28]. Unlike XGBoost and LightGBM, CatBoost can automatically process categorical features and convert them into numerical features. By combining categorical feature processing, it enriches the feature dimensions through different relationships between features. Furthermore, CatBoost uses ordered boosting to handle noise points in datasets, thus solving the prediction shift issue. It uses completely symmetric trees as the base model, effectively avoiding overfitting problems, increasing model reliability, and speeding up the prediction process.

In this paper, a division of training, validation, and test sets is used to ensure the generalization ability of the model. First, the IHHO-Catboost algorithm is trained using the training set, while the parameters of the model are tuned using the validation set to obtain the optimal model configuration. After identifying the optimal model, we further evaluate the predictive performance of the model using an independent test set. This process ensures that the performance of the model on unseen data truly reflects its predictive ability, avoids the overfitting problem, and enhances the credibility of the model in real-world applications.

2.3.2. Harris Hawk Optimization Algorithm

The Harris hawk optimization algorithm has the advantages of requiring a few parameters to tune, being simple and easy to implement, and having strong adaptability [29]. The hunting and chasing process of the Harris hawk is divided into two stages: global exploration and local exploitation.

Global exploration phase: When the prey's escape energy is $|E| \geq 1$, the algorithm executes global exploration behavior. In this phase, the Harris hawk population is perched in a location, waiting, detecting, and inspecting the search space [*lb*, *ub*] to locate prey. The position update formula in this phase is as follows:

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1|X_{rand}(t) - 2r_2X(t)|, \ q \geq 0.5 \\ (X_{rabbit}(t) - X_m(t)) - r_3(lb + r_4(ub - lb)), \ q < 0.5 \end{cases} \quad (3)$$

In the formula, $X(t)$ and $X(t+1)$ represent the position of the hawk at the t and t + 1 iterations, respectively; $X_{rand}(t)$ is a random position of the hawk at the t iteration; $X_{rabbit}(t)$ is the current optimal individual position; r and q are random numbers between 0 and 1; and lb and ub represent the lower and upper bounds of the search space, respectively. $X_m(t)$ denotes the average position of all individuals in the Harris hawk population, and its expression is as follows:

$$X_m(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t) \quad (4)$$

In the formula, N represents the number of individuals in the Harris hawk population; $X_i(t)$ denotes the position of the i Harris hawk individual during the t iteration. During the iterative search process, whether the algorithm performs global exploration or local exploitation depends on the magnitude of the linearly decreasing prey energy value E. The escape energy E of the prey linearly decreases during the escape process, and its expression is as follows:

$$E = 2E_0(1 - \frac{t}{T}) \quad (5)$$

where E represents the energy of the escaping prey, T is the maximum number of iterations, and E_0 is the initial state of its energy, calculated as $E_0 = 2 \times rand - 1$, where rand is a random number in the interval [0, 1].

When $|E| < 1$, it indicates that the Harris hawk population has located the prey. In this phase, as the escape energy of the prey is relatively small, the algorithm performs local exploitation behavior. Based on the Harris hawk's surprise pounce hunting strategy, four different encircling strategies are formed by comparing the random number r and the magnitude of the prey energy $|E|$ with 0.5.

Strategy 1: Soft besiege. When $|E| < 0.5$ and $r \geq 0.5$, at this time, the prey has abundant escape energy E, and the Harris hawk employs a soft besiege strategy. The update of the hawk's position is shown in Equation (6).

$$X(t+1) = \Delta X(t) - E|JX_{rabbit}(t) - X(t)| \tag{6}$$

In the formula, J is a random number between 0 and 2.

Strategy 2: Hard besiege. When $|E| < 0.5$ and $r \geq 0.5$, the prey's escape energy E is low, and it does not have sufficient energy to escape. In this case, the Harris hawk employs a hard besiege strategy, and its position update is shown in Equation (7).

$$X(t+1) = X_{rabbit}(t) - E|\Delta X(t)| \tag{7}$$

Strategy 3: Rapid dive soft besiege. When $|E| \geq 0.5$ and $r < 0.5$, the position is updated based on Equation (8), and it is compared with the fitness of the current position. If the fitness does not improve, indicating a failed besiege, the hawk population performs a random walk based on Levy flight and updates their positions using Equation (9).

$$X(t+1) = X_{rabbit}(t) - E|JX_{rabbit}(t) - X(t)| \tag{8}$$

$$Z = Y + S \times LF(D) \tag{9}$$

In the formula, D represents the dimension of the problem, and S is a random vector. The expression for $LF(D)$ is as follows:

$$LF(x) = 0.01 \times \frac{\mu \times \sigma}{|v|^{\frac{1}{\beta}}},\ \sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\frac{\pi\beta}{2})}{\Gamma(\frac{1+\beta}{2}) \times \beta \times 2^{\frac{\beta-1}{2}}} \right) \tag{10}$$

In the formula, μ and ν represent random vectors, and α and β are constants with values of 0.69 and 1.5, respectively.

Strategy 4: Rapid dive hard besiege. When $|E| < 0.5$ and $r < 0.5$, and the prey's escape energy is low, the hawk population performs a rapid dive hard besiege. The position update is shown in Equation (11). If the rapid dive fails, a Levy flight random walk is executed using Equation (9) for position updates.

$$Y = X_{rabbit}(t) - E|JX_{rabbit}(t) - X_m(t)| \tag{11}$$

2.4. Multi-Strategy Improvement of HHO

To address the issue of the HHO algorithm being prone to getting stuck in local optima, various improvement strategies are utilized to increase the diversity of the Harris hawk population. Considering the HHO algorithm's inflexibility in adjusting the search process and its inability to conduct phase-specific searches, which sometimes leads to getting trapped in local optima, thereby relatively diminishing the search precision, the parameter identification method of the improved Harris hawk optimization (IHHO) algorithm is enhanced to expand the search space and improve the global search capability of the HHO algorithm.

2.4.1. Improvement Point One: Latin Hypercube Population Initialization Strategy

Latin Hypercube sampling is a multidimensional stratified sampling technique that efficiently samples within the distribution intervals of variables. It essentially divides the interval [0, 1] into N non-overlapping sub-intervals of equal width and independently

samples from each sub-interval with equal probability, ensuring that sample points are uniformly distributed across the entire distribution interval. Random sampling, on the other hand, follows a uniform distribution in the interval [0, 1], and in cases with a small number of samples, random distribution may not spread the samples evenly across the entire interval. Unlike random sampling, Latin Hypercube sampling ensures that variables cover the entire distribution space. Random sampling and Latin Hypercube sampling each select 10 points from the interval [0, 1]. Latin Hypercube sampling can distribute the samples across the entire space, especially when dealing with a small number of samples.

2.4.2. Improvement Point Two: Normal Cloud Model

The cloud model is described by three parameters: expectation, entropy, and hyper-entropy. When the hyper-entropy increases, the range of cloud droplets' distribution increases accordingly, and when the hyper-entropy increases, the dispersion degree of cloud droplets also increases. This reflects the randomness and fuzziness of cloud droplet distribution. The positive normal cloud generator is an algorithm used to generate cloud droplets that follow a normal distribution. It generates cloud droplets based on the specified parameters until the desired number of cloud droplets is generated. The process of generating normal cloud droplets can be defined as follows:

Among them, Nd is the expected number of cloud droplets.

$$X[x_1, x_2, x_3, \cdots x_{Nd}] = Gnc(Ex, En, He, Nd) \tag{12}$$

Introduce the normal cloud model as a new updating mechanism for Harris's hawk position. By using the expected value of the normal cloud model to refine the optimal position solution and adjusting the dispersion of Harris's hawk positions by manipulating the remaining position solutions, the formula is as follows:

$$Position' = Gnc(Position_{best}, En, He, Nd) \tag{13}$$

$$En = \lambda \times (\frac{T-t}{T})^{T} \tag{14}$$

$$He = En \times 10^{-\varepsilon} \tag{15}$$

2.4.3. Improvement Point Three: Reverse Learning

The concept of reverse learning has become a commonly used improvement strategy in optimization algorithms since its inception. It primarily involves the reverse learning of feasible solutions by evaluating the original solution and its reverse counterpart and selecting the better solution to incorporate into the algorithm's iterations [28]. To achieve this, the idea of random reverse learning is introduced to update the worst Harris's hawk position, as shown in the following formula:

$$X_{worst,t+1} = ub_1 + rand \times (lb_1 - X_{worst,t}) \tag{16}$$

Among them, $X_{worst,t+1}$ represents the worst position of Harris's hawk; ub_1 and lb_1 are the dynamic boundaries and upper and lower bounds, respectively.

As the algorithm iterates, the update of the worst position is performed through Equation (16), yielding a reverse random solution, thereby enhancing the diversity of the Harris's hawk population and increasing the chances of finding the global optimum. At the same time, random reverse learning uses dynamic boundaries ub_1 and lb_1, reducing the problem of losing search information associated with traditional fixed boundaries ub and lb, as well as reducing the computational complexity of the improved algorithm.

2.4.4. Improvement Point Four: Dynamic Perturbation Strategy

When the prey's energy $|E|$ is less than 1, the algorithm enters the development stage, but it cannot guarantee that the population is close to the global optimum at this point,

which may lead to premature convergence or getting stuck in local optima [30]. Therefore, a dynamic perturbation strategy is introduced among the four predation strategies, ensuring optimization precision while allowing for a quick escape from local optima [31].

$$\psi = -\cos\left(\frac{\pi t}{2T} + \pi\right) \tag{17}$$

$$X_{\psi,rabbit} = \psi \times X_{rabbit} \tag{18}$$

2.5. Model Evaluation

To assess the predictive capability and accuracy of the model, an analysis of the prediction model is performed using the coefficient of determination (R^2), root mean square error (*RMSE*), and mean absolute error (*MAE*). A higher R^2 value, along with smaller *RMSE* and *MAE* values, indicates a more accurate prediction model. The calculation formulas are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widetilde{y}_i\right| \tag{19}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i-1}^{N}(y_i - \widetilde{y}_i)^2} \tag{20}$$

$$R^2 = 1 - \frac{\sum_i (\widetilde{y_i} - y_i)^2}{\sum_i (\overline{y_i} - y_i)^2} \tag{21}$$

3. Results and Discussion

3.1. SHAP Analysis

SHAP, based on game theory, is used to explain the impact of various features of a specific sample on the prediction results in complex algorithms [32]. SHAP explains the importance of features on model predictions through attribution values [33]. If the attribution value is positive, it indicates a positive impact of the feature on the model prediction; otherwise, it is a negative impact. The model's prediction value is obtained by adding the model's predicted average value and the feature's attribution value [34].

As shown in Figure 2, the importance of different features in influencing indoor temperature is ranked from top to bottom. Red represents a positive attribution value, while blue represents a negative attribution value. According to the SHAP analysis results, the variables that most significantly affect indoor temperature are carbon dioxide concentration, outdoor air temperature, the degree of opening and closing of the windward opening, soil temperature, and outdoor illumination, in that order. The importance of these variables is indicated by their color, with red indicating a stronger positive influence and blue indicating a stronger negative influence.

Carbon dioxide concentration has the most significant and negative impact on indoor temperature. This means that when indoor carbon dioxide concentration increases, indoor temperature decreases. This may be due to the higher carbon dioxide concentration leading to a decrease in indoor air quality, which negatively affects indoor temperature. Therefore, it is important to control carbon dioxide concentration in the greenhouse to maintain good air quality and suitable temperatures.

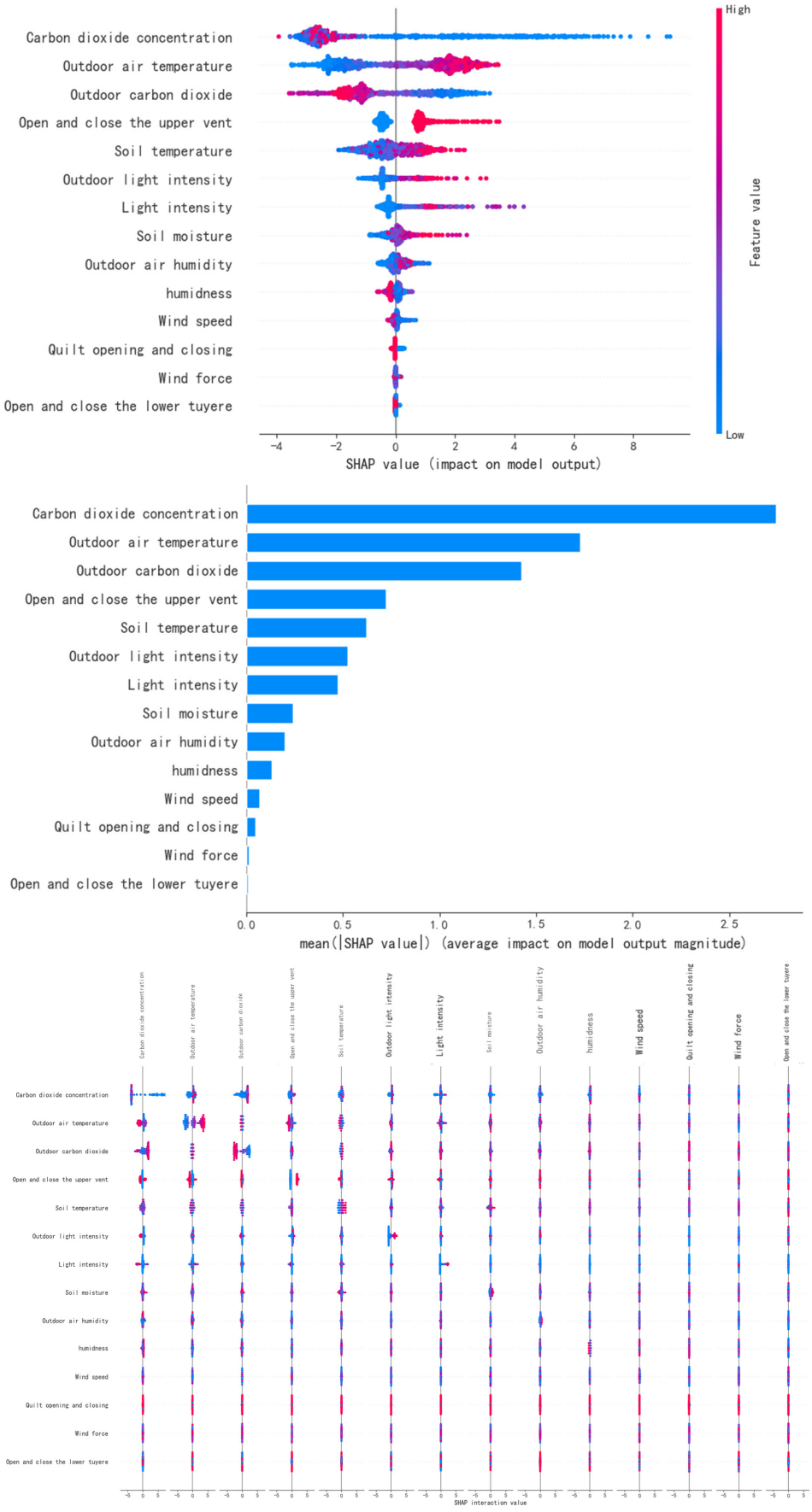

Figure 2. Feature importance ranking based on the SHAP method.

Outdoor air temperature has a relatively large positive impact on indoor temperature. When the outdoor air temperature rises, the indoor temperature also increases. The degree of opening and closing of the windward opening also has some influence on indoor temperature. A larger opening degree of the windward opening will increase indoor temperature, while a smaller opening degree will decrease it. This is because changes in the degree of opening and closing of the windward opening affect air circulation and ventilation, thus influencing indoor temperature regulation. Therefore, it is essential to adjust the degree of opening and closing of the windward opening reasonably in the greenhouse to control indoor temperature fluctuations.

In addition, factors like soil temperature and outdoor illumination have a relatively smaller impact, and their specific effects may be dependent on the particular model and experimental environment.

Through SHAP analysis, a better understanding of the impact of different variables on indoor temperature has been achieved. In a greenhouse environment, effectively controlling carbon dioxide concentration, outdoor air temperature, and the degree of opening and closing of the windward opening can help regulate the indoor environment, providing an optimal growing environment and improving the greenhouse's efficiency and comfort. Additionally, considering the influence of other factors comprehensively allows for the holistic optimization of the greenhouse environment.

The relationship between carbon dioxide concentration and temperature is demonstrated in the Figure 3. The experimental data points are visually depicted in the form of a scatterplot, and the linear regression relationship between carbon dioxide concentration and temperature is represented by the red straight line. As can be seen from the graph, the temperature shows a certain upward trend as the carbon dioxide concentration increases, and this trend is quantitatively represented by the slope of the regression line. In particular, data points in the region of high concentration show a greater contribution to the regression line, indicating that high concentrations of CO_2 have a significant effect on temperature prediction.

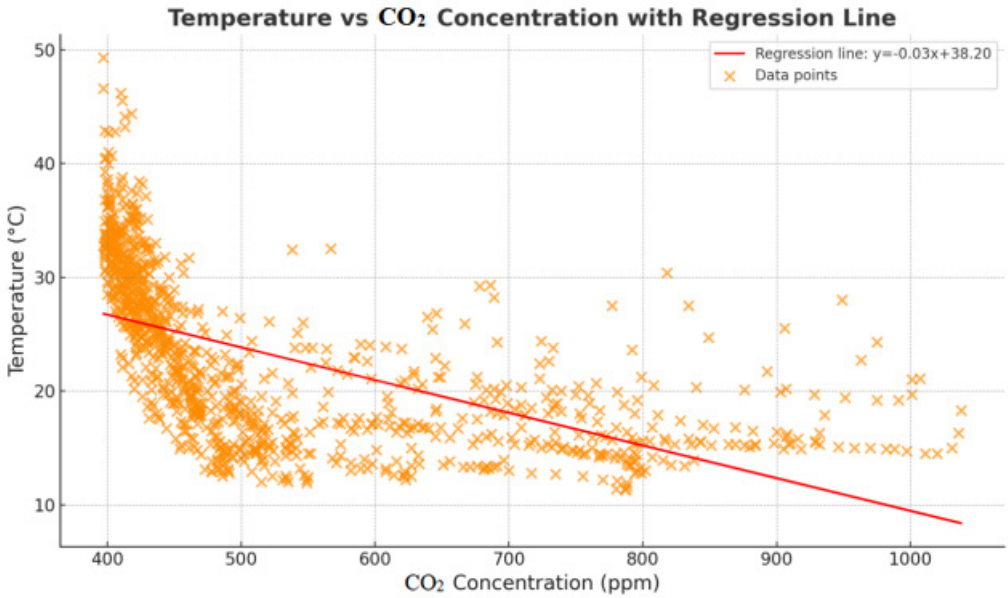

Figure 3. Regression scatter plot of carbon dioxide concentration versus temperature.

3.2. Prediction versus Actual Values

This analysis focuses on four key environmental parameters: temperature, humidity, carbon dioxide, and light intensity. Figure 4 demonstrates the relationship between the model's predicted and actual values through time series comparisons, revealing the model's ability to capture variations in these environmental parameters. Although deviations are

observed at some extreme values, overall, the model is able to accurately track actual fluctuations in the vast majority of cases.

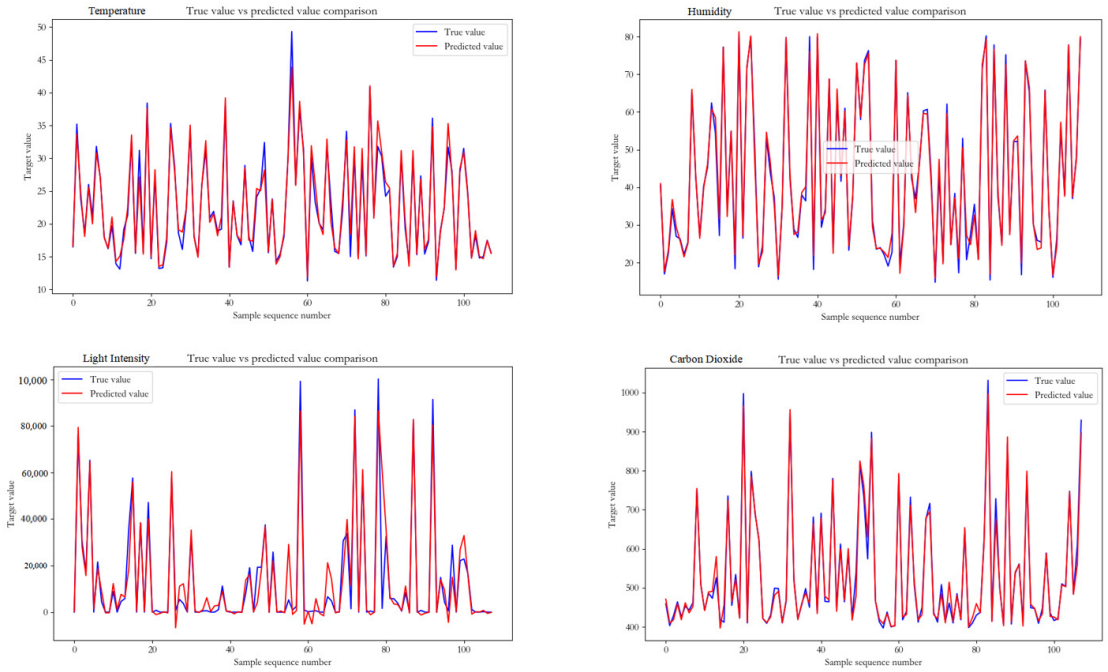

Figure 4. Dynamic prediction and comparison of environmental parameters.

A detailed analysis of the data in Figure 4 reveals that the model performs particularly well in the prediction of temperature and humidity, with the time series fluctuations highly consistent with the actual records, suggesting that the model is able to effectively reflect the effects of day–night temperature differences and seasonal variations on the greenhouse environment. In the prediction of carbon dioxide concentration, the model also showed high accuracy and was able to capture the subtle adjustment of environmental parameters by biological processes such as plant photosynthesis, thus providing a scientific basis for environmental control inside the greenhouse.

For the prediction of light intensity, although the overall trend was consistent with the actual situation, deviations at certain times of day (e.g., on days with thick cloud cover or at sunset) pointed out the challenges of the model in dealing with extreme light conditions. These biases may stem from the model's insufficient sensitivity to changes in light intensity or the inability of the available data to adequately reflect complex natural light patterns.

Overall, despite some localized biases, the present model demonstrated good performance in tracking the actual fluctuations of key environmental parameters in the greenhouse. Future work will focus on further optimizing the model, especially improving the prediction accuracy for extreme values and complex lighting conditions, to ensure the reliability and usefulness of the model in a wider range of application scenarios. Through comprehensive analysis and careful adjustment of the model, it is expected that more precise control and management of the greenhouse environment will be realized, providing strong technical support for efficient and sustainable agricultural production.

The high statistical accuracy of the model is observed in the scatterplot presented in Figure 5 by comparing the correlation between the actual and predicted values. The linear fit line as well as the high R^2 values in the plot indicate the statistical accuracy of the model. Specifically, the predicted values of temperature and humidity are very close to the actual values, with R^2 values of 0.971 and 0.991, respectively, showing the model's ability to predict these two parameters with extreme accuracy. In addition, the predictions of carbon dioxide and light intensity also showed a high degree of accuracy, with R^2 values of 0.987 and 0.891, respectively, which demonstrated that the model also had good predictive ability

in these parameters. Overall, the model shows convincing accuracy and reliability in all parameters, especially humidity and CO_2, which are almost perfectly predicted. These results not only demonstrate the efficiency of the IHHO-CatBoost model in environmental monitoring and control systems but also reveal that there is room for further optimization of the model for light intensity prediction. Although the R^2 values for light intensity are relatively low, they are still within the acceptable range, suggesting that there is potential to improve the model's prediction accuracy on this parameter through further algorithmic adjustments and data analysis.

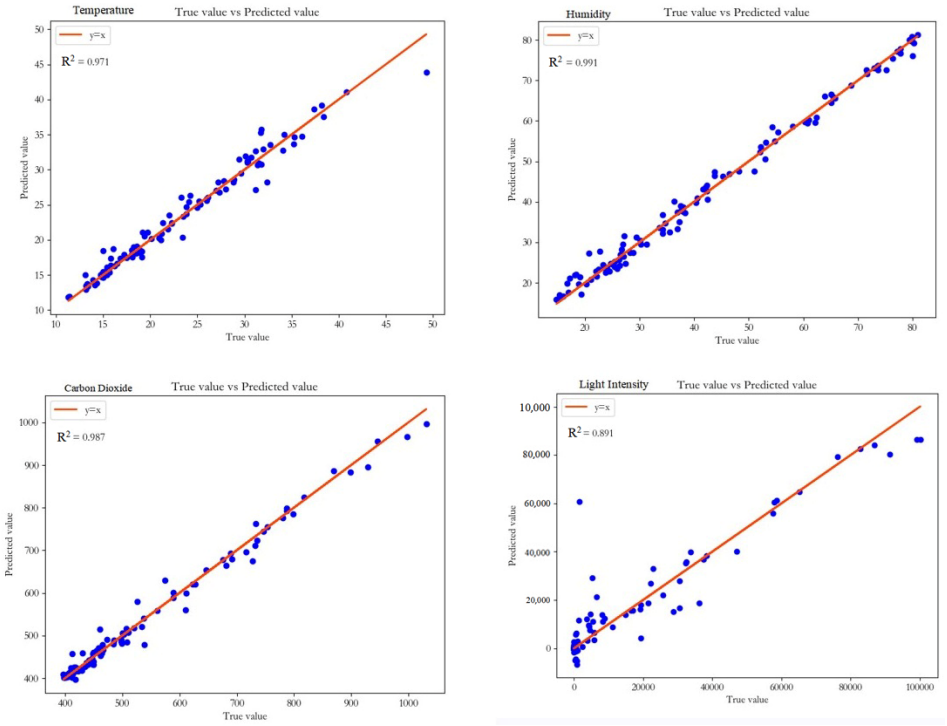

Figure 5. Prediction accuracy analysis of environmental parameters.

3.3. Comparison and Analysis of Different Model Predictions

Based on the optimization results of IHHO, the parameters of the CatBoost model were set to make the final prediction for the temperature in the sunlight greenhouse. To verify the performance of the model proposed in this paper, experiments were conducted to predict the sunlight greenhouse environment using the BP neural network, LSTM model, and HHO-CatBoost model, respectively. For a comparative analysis, the IHHO-LSTM model used in the literature was also selected for experimentation. To ensure the accuracy and reproducibility of the study results, Table 1 shows the settings for the four parameters.

Table 1. Different model parameter settings.

Model	Learning Rate	Number of Iterations	Storey	Number of Neurons per Layer
LSTM	0.01	1000	2	64, 32
BP neural network	0.05	200	3	128, 64, 32
HHO-LSTM	0.01	50	2	50, 50

To reduce the workload of manually adjusting hyperparameters and ensure optimal model performance, this study utilizes the IHHO to optimize key parameters in the CatBoost model, such as learning rate, depth, and L2 regularization. The optimization range for the learning rate is between [0.01, 0.5], the tree depth is between [2, 10], and the L2

regularization parameter is between [0.1, 10]. During the training process, the population size of the IHHO algorithm is set to 30, with 20 iterations. The model's performance is measured by the mean square error (MSE) of the prediction results; the lower the MSE, the better the model's fit to the data.

The results of each predictive model are shown in Figure 6.

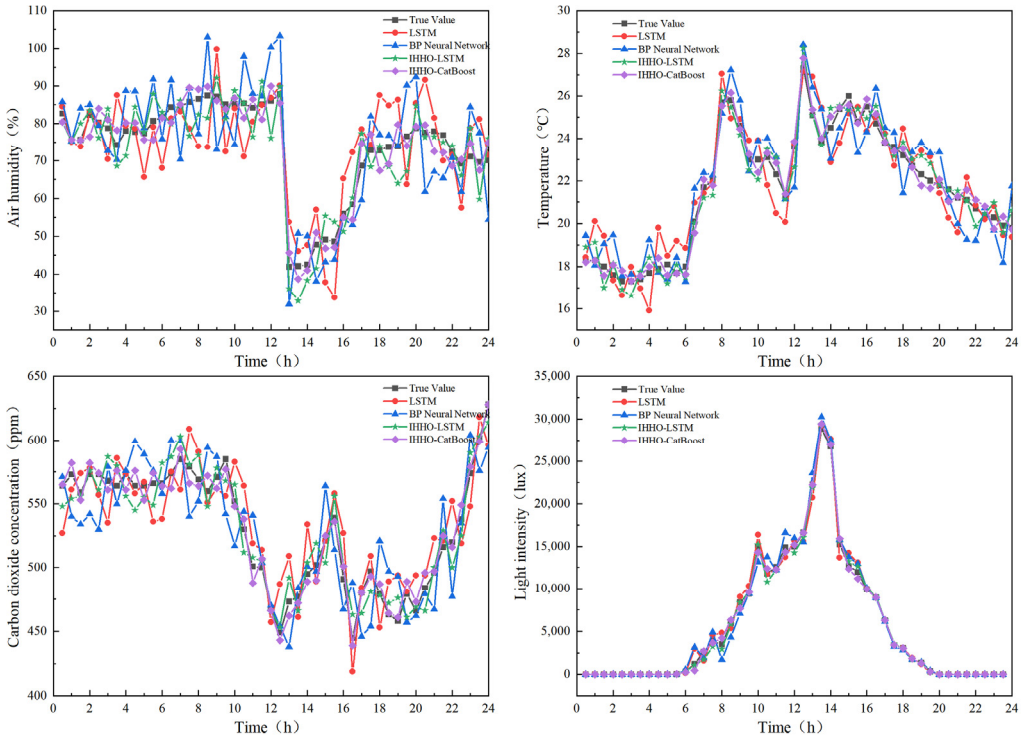

Figure 6. Prediction of environmental parameters by different models.

Intuitively, the prediction results of the BP neural network and LSTM model are not very satisfactory. In the BP neural network model, there is a significant difference between the predicted results and the actual values, while in the LSTM model, despite some predictions being similar to the actual values, deviations often occur after the 13th time point, leading to suboptimal predictions. In contrast, the predictions based on the IHHO-CatBoost model are better and largely consistent with the actual values, showing more accuracy across the entire time series. The structure and characteristics of the BP neural network and LSTM model may lead to difficulties in handling time series data and capturing long-term dependencies within the series. The IHHO-CatBoost model, which utilizes the CatBoost algorithm, effectively handles nonlinear relationships and complex patterns between features. Therefore, predictions based on the IHHO-CatBoost model are more accurate and consistent with actual values, indicating its superior ability to capture and predict the patterns and trends of indoor temperature variations.

In summary, based on intuitive observation and comparative analysis, it can be concluded that the prediction results of the BP neural network and LSTM model are less satisfactory, while the IHHO-CatBoost model shows better predictions, closely aligning with actual values and improving prediction accuracy compared to the other four models. This suggests that selecting suitable models and algorithms can significantly enhance the accuracy and reliability of predictions for indoor temperature forecasting problems.

3.4. Assessing IHHO-CatBoost's Predictive Efficacy

Based on the results in Table 2, it can be concluded that the IHHO-CatBoost model shows the best performance in terms of prediction accuracy, with an *MAE* of 0.8751 °C, an *RMSE* of 1.1799 °C, and an R^2 of 0.9707. This indicates that the model's predicted values

have a very small average absolute error and root mean square error relative to the actual values, and the R^2 value close to 1 suggests the model can accurately interpret the variations in the actual data.

Table 2. Statistical results of prediction error.

Evaluation Index	Model			
	LSTM	**BP Neural Network**	**HHO-LSTM**	**IHHO-CatBoost**
MAE/°C	1.2523	1.5652	0.9852	0.8751
RMSE/°C	1.7654	1.8913	1.3324	1.1799
R^2	0.9245	0.9051	0.9337	0.9707
MAPE/%	4.18	4.62	2.39	1.43

Compared to the LSTM model, the IHHO-CatBoost model shows significant improvements in all metrics. The *MAE* of the IHHO-CatBoost model is reduced by 30.1%, *RMSE* by 33.1%, MAPE by 65.7%, and R^2 increased by 4.7% compared to the LSTM model. This demonstrates that the IHHO-CatBoost model has better prediction accuracy and stability, capturing the trends and patterns in the data more precisely.

When compared to the BP neural network model, the IHHO-CatBoost model also exhibits significant advantages. Relative to the BP neural network, the IHHO-CatBoost model reduces *MAE* by 44.0%, *RMSE* by 37.6%, MAPE by 76.3%, and increases R^2 by 6.7%. The IHHO-CatBoost model excels at capturing nonlinear relationships and complex patterns in the data, outperforming the traditional BP neural network in predicting target variables.

Furthermore, compared to the HHO-LSTM model, the IHHO-CatBoost model also shows better performance. Relative to the HHO-LSTM model, the IHHO-CatBoost model reduces *MAE* by 11.1%, *RMSE* by 11.4%, MAPE by 40.1%, and increases R^2 by 3.8%. The IHHO-CatBoost model predicts target variables more accurately and performs better at reducing prediction errors compared to the HHO-LSTM model.

In summary, based on the comparison results, it can be concluded that the IHHO-CatBoost model demonstrates the best performance in this study, offering high prediction accuracy.

Although this study utilizes machine learning algorithms to predict greenhouse environmental conditions, there are a few limitations: First, the use of grid search for automatic tuning of the model parameters has its shortcomings. Future work will involve the use of AI-based automatic optimization algorithms for model parameter tuning. Second, due to the limited collection time, only one season's worth of environmental data was available, resulting in a relatively small dataset. In future studies, data from an entire year will be collected for further refinement and validation of the model.

The results of the solar greenhouse environmental prediction model based on the IHHO-CatBoost algorithm applied in this study show that the model has significant advantages in prediction accuracy and processing efficiency. The IHHO-CatBoost algorithm is excellent in capturing the complex nonlinear correlation between environmental variables by comparison with the traditional model, and the gradient enhancement mechanism of the CatBoost algorithm further improves the accuracy of data processing, especially in the face of the data noise. The algorithm is also able to improve the accuracy of prediction, which is of great significance for the real-time monitoring and management of the greenhouse environment. This is of great significance for the real-time monitoring and management of the greenhouse environment.

Future research will be devoted to optimizing the solar greenhouse environment prediction model based on the IHHO-Catboost algorithm and improving the prediction accuracy and generalization ability of the model by introducing more advanced optimization techniques and multidimensional data fusion. It is planned to combine the model with IoT technology to develop a real-time monitoring system to realize instant prediction and adaptive management of greenhouse environments. In addition, the adaptability

of the model will be validated under different geographic and climatic conditions to ensure its generalizability to all types of greenhouses. Meanwhile, the economic benefits and ecological impacts of the model in practical applications will be assessed, aiming to develop user-friendly applications that facilitate efficient and accurate environmental management and decision-making by agricultural producers and greenhouse managers and, thus, promote sustainable agricultural development.

4. Conclusions

In this study, a method for predicting the environment in a solar greenhouse is proposed based on the IHHO-CatBoost model, combining the optimized HHO and CatBoost models. The model provides advanced prediction of environmental data in greenhouses. Compared with the traditional LSTM, BP neural network, and HHO-LSTM models, the *MAE* was reduced by 30.1%, 44.0%, and 11.1%, the *RMSE* by 33.1%, 37.6%, and 11.4%, and the MAPE by 65.7%, 76.3%, and 40.1%, respectively. The study also examined the error between the model and the actual measurements. The results showed that the IHHO-Catboost model performed the best in terms of prediction accuracy, achieving 98.5% accuracy, which is an improvement of 5.2% and 6.7% compared to the LSTM model and BP neural network, respectively. This indicates that the IHHO-CatBoost model is more accurate and reliable for these prediction tasks.

The IHHO-CatBoost model shows high accuracy and low peak error in greenhouse environment prediction, which can reduce labor costs, provide a theoretical basis for the application of precise greenhouse environment regulation and damage warning, and improve crop quality and greenhouse production efficiency. By introducing the IHHO-CatBoost model, this study not only improves the prediction accuracy of solar greenhouse environmental parameters but also provides new methodological support for the precise regulation of greenhouse crop growth environments. In addition, this study has important practical applications for optimizing agricultural production and improving crop yield and quality. Combined with the prediction results, farmers are reminded in advance to make corresponding adjustments, and the automatic control system adjusts quilt opening and closing, ventilation equipment, or shading facilities in advance to optimize crop growth conditions. This not only improves the precision and efficiency of crop management but also helps to reduce resource waste, enhance the ability to cope with extreme weather events, and reduce the risk of pests and diseases. Further research can delve into the industrial structure and industrial output of facility agriculture to improve industrial profitability, so as to achieve increased production and income for its practitioners, sustainable development of the industry, further lead the development of agricultural intelligence, and promote the process of agricultural digitization.

Author Contributions: Conceptualization, J.Y.; Methodology, Q.L., J.Z. and W.Z.; Software, J.Y., Q.L., L.L. and W.Z.; Validation, J.Y. and Y.W.; Formal analysis, J.Y., G.R. and J.Z.; Investigation, J.Y.; Resources, W.W. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by XDHZFQY2022-02 and 202103021224123 (Wuping Zhang); 202202140601021 (Fuzhong Li).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not available to the public because of their confidential nature.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

CNNs	Convolutional Neural Networks
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
GRU	Gated Recurrent Neural Network
BP	Back Propagation

HHO	Harris Hawk Optimization
CatBoost	Gradient Boosting + Categorical Features

References

1. Zhang, R.; Liu, Y.; Zhu, D.; Ge, M. Construction and Application of Front Roof Lighting Efficiency Model of Solar Greenhouse Considering Film Ash Deposition. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 190–199.
2. Lei, W.; Lu, H.; Qi, X.; Tai, C.; Fan, X.; Zhang, L. Field Measurement of Environmental Parameters in Solar Greenhouses and Analysis of the Application of Passive Ventilation. *Sol. Energy* **2023**, *263*, 111851. [CrossRef]
3. Martínez-Gómez, P.; Rahimi Devin, S.; Salazar, J.A.; López-Alcolea, J.; Rubio, M.; Martínez-García, P.J. Principles and Prospects of Prunus Cultivation in Greenhouse. *Agronomy* **2021**, *11*, 474. [CrossRef]
4. Xu, D.; Ren, L.; Zhang, X. Predicting Multidimensional Environmental Factor Trends in Greenhouse Microclimates Using a Hybrid Ensemble Approach. *J. Sens.* **2023**, *2023*, 6486940. [CrossRef]
5. Dou, Z.; Feng, H.; Zhang, H.; Abdelghany, A.E.; Zhang, F.; Li, Z.; Fan, J. Silicon application mitigated the adverse effects of salt stress and deficit irrigation on drip-irrigated greenhouse tomato. *Agric. Water Manag.* **2023**, *289*, 108526. [CrossRef]
6. Fadel, M.M.; El-Ghamrawy, S.M.; Ali-Eldin, A.M.T.; Hassan, M.K.; El-Desoky, A.I. The proposed hybrid deep learning intrusion prediction IoT (HDLIP-IoT) framework. *PLoS ONE* **2022**, *17*, e0271436. [CrossRef]
7. Vijaya Kumar, S.; Chakraborty, S.; Revi, K.; Munet, R. Simulation of Telemetry Signals of a Car Using Machine Learning. *SAE Int. J. Adv. Curr. Pract. Mobil.* **2022**, *5*, 1465–1472.
8. Zhou, S.; Guo, S.; Du, B.; Huang, S.; Guo, J. A Hybrid Framework for Multivariate Time Series Forecasting of Daily Urban Water Demand Using Attention-Based Convolutional Neural Network and Long Short-Term Memory Network. *Sustainability* **2022**, *14*, 11086. [CrossRef]
9. Zhang, Q.; Yu, H.; Zhang, Z.; Dong, L.; Zhang, Q.; Shao, C. Study on Air Flow in Solar Greenhouse Using CFD Model. *Trans. Chin. Soc. Agric. Eng.* **2012**, *28*, 166–171. (In Chinese with English Abstract)
10. Fu, Q.; Li, X.; Zhang, G.; Li, X. A Temperature and Vent Opening Couple Model in Solar Greenhouses for Vegetable Cultivation Based on Dynamic Solar Heat Load Using Computational Fluid Dynamics Simulations. *J. Food Process Eng.* **2022**, *46*, e14240. [CrossRef]
11. Yuan, Y.L.; Sheng, W.Y. Prediction Method of Dynamic Change of Stem Diameter Based on Principal Component Regression. *Trans. Chin. Soc. Agric. Mach.* **2015**, *46*, 306–314. (In Chinese with English Abstract)
12. Yang, Y.; Gao, P.; Sun, Z.; Wang, H.; Lu, M.; Liu, Y.; Hu, J. Multistep Ahead Prediction of Temperature and Humidity in Solar Greenhouse Based on FAM-LSTM Model. *Comput. Electron. Agric.* **2023**, *213*, 108261. [CrossRef]
13. Hardaha, S.; Edla, D.R.; Parne, S.R. A Survey on Convolutional Neural Networks for MRI Analysis. *Wirel. Pers. Commun.* **2022**, *128*, 1065–1085. [CrossRef]
14. Wang, S.; Xia, P.; Chen, K.; Gong, F.; Wang, H.; Wang, Q.; Zhao, Y.; Jin, W. Prediction and optimization model of sustainable concrete properties using machine learning, deep learning and swarm intelligence: A review. *J. Build. Eng.* **2023**, *80*, 108065. [CrossRef]
15. Zhu, Z.; Lin, K.; Jain, A.K.; Zhou, J. Transfer Learning in Deep Reinforcement Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13344–13362. [CrossRef] [PubMed]
16. Zhao, Q.; Song, Z.; Li, Q.; Zheng, W.; Liu, Y.; Zhang, Z. Research on Multi-Point Temperature and Humidity Prediction Method of Mushroom House Based on CNN-GRU. *Trans. Chin. Soc. Agric. Mach.* **2019**, *51*, 294–303.
17. Liu, L.; Yang, S.; Wang, Z.; He, X.; Zhao, W.; Liu, S.; Du, W.; Mi, J. Coke Quality Prediction Based on Improved WOA-LSTM. *Acta Chem. Sin.* **2022**, *73*, 1291–1299. (In Chinese with English Abstract)
18. Zu, L.; Liu, P.; Zhao, Y.; Li, T.; Li, H. Study on the Environmental Prediction Model of Sunlight Greenhouse Based on SSA-LSTM. *Trans. Chin. Soc. Agric. Mach.* **2023**, *54*, 351–358. (In Chinese)
19. Ji, R.; Shi, S.; Zhao, Y.; Liu, Z.; Wu, Z. Based on LSTM-Seq2Seq Multi-Parameter Prediction of Rabbit House Environment. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 396–401+409. (In Chinese)
20. Lu, J.; Zhang, Q.; Yang, Z.; Tu, M.; Lu, J.; Peng, H. Short-Term Load Forecasting Method Based on CNN-LSTM Hybrid Neural Network Model. *Autom. Electr. Power Syst.* **2019**, *43*, 131–137. (In Chinese with English Abstract)
21. Li, P.; Li, M.; Liu, D. Power Load Forecasting Based on Improved Regression Method. *Power Syst. Technol.* **2006**, *30*, 99–104. (In Chinese)
22. Huang, Z.; Gu, C.; Peng, J.; Wu, Y.; Gu, H.; Shao, C.; Zheng, S.; Zhu, M. A Statistical Prediction Model for Sluice Seepage Based on MHHO-BiLSTM. *Water* **2024**, *16*, 191. [CrossRef]
23. Moayedi, H.; Le Van, B. Feasibility of Harris Hawks Optimization in Combination with Fuzzy Inference System Predicting Heating Load Energy Inside Buildings. *Energies* **2022**, *15*, 9187. [CrossRef]
24. Cai, T.; Tang, H. Overview of the Least Squares Fitting Principle of Savitzky-Golay Smoothing Filters. *Digit. Commun.* **2011**, *38*, 63–68+82. (In Chinese)
25. Guo, S.; Lü, Z.; Feng, Y. Structural Non-probabilistic Reliability Model Based on Interval Analysis. *Chin. J. Comput. Mech.* **2001**, *14*, 56–60. (In Chinese)
26. Zhang, Y.; Feng, B.; Chen, Y.; Liao, W.; Guo, C. Fault Diagnosis Method for Oil-Immersed Transformers Based on Genetic Algorithm Optimized XGBoost. *Electr. Power Autom. Equip.* **2021**, *41*, 200–206. (In Chinese)

27. Xuan, L.; Lin, Z.; Liang, J.; Huang, X.; Li, Z.; Zhang, X.; Zou, X.; Shi, J. Predictionofresilienceandcohesionofdeep-friedtofubyultrasonicdetectionandLightGBMregression. *Food Control* **2023**, *154*, 110009. [CrossRef]
28. Almotairi, S.; Badr, E.; Abdul Salam, M.; Dawood, A. Three Chaotic Strategies for Enhancing the Self-Adaptive Harris Hawk Optimization Algorithm for Global Optimization. *Mathematics* **2023**, *11*, 4181. [CrossRef]
29. Zhao, Y.; Si, D.; Pei, J.; Yang, X. Geodesic Basis Function Neural Network. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. *early access*.
30. Zhou, G.; Chen, Z.; Zhang, C.; Chang, F. An adaptive ensemble deep forest based dynamic scheduling strategy for low carbon flexible job shop under recessive disturbance. *J. Clean. Prod.* **2022**, *337*, 130541. [CrossRef]
31. Fu, W.; Zhang, X.R.; Zhang, H.R.; Fu, Y.C.; Liu, X.T. Research on Ultra-Short-Term Wind Speed Prediction Based on INGO-SWGMN Hybrid Model. *J. Sol. Energy* **2023**, *45*, 123–130.
32. Huang, L.; Jiang, B.; Lu, S.; Liu, Y.; Li, D. Review of Recommendation Systems Based on Deep Learning. *J. Comput.* **2018**, *41*, 1619–1647. (In Chinese with English Abstract)
33. Bao, H.N.; Xiong, J.; Zhang, C.Y.; Guo, X.X.; Zhao, Y.; Zhou, Y.Y. Study on Influencing Factors and Model Prediction of Bioavailability of Arsenic and Benzo[a]pyrene in Contaminated Soils. *J. Environ. Eng.* **2023**, *17*, 3392–3399. (In Chinese with English Abstract)
34. Chen, H.; Yi, Y.; Huang, S.; Chen, J. Short-Term Photovoltaic Power Prediction Method Based on CatBoost Algorithm. *Zhejiang Electr. Power* **2023**, *42*, 67–75. (In Chinese with English Abstract)