energies

MDPI

Article

A Novel Wind Power Outlier Detection Method with Support Vector Machine Optimized by Improved Harris Hawk

Jingtao Huang [1,2,*], Jin Qin [1] and Shuzhong Song [1]

[1] College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
[2] Henan Engineering Laboratory of Power Electronic Devices and Systems, Luoyang 471023, China
* Correspondence: jthuang@haust.edu.cn

Abstract: The accurate detection of wind power outliers plays a crucial role in wind power forecasting, while the inherited strong randomness and high fluctuations bring great challenges to this issue. This work investigates the way to improve the outlier detection accuracy based on support vector machine (SVM). Although SVM can achieve good results for outlier detection in theory, its performance is heavily dependent on the hyper-parameters. Parameter optimization is not an easy task due to its complex nonlinear multi-optimum nature; an improved Harris hawk optimization (IHHO) is proposed to optimize the parameters of SVM for more accurate outlier detection. HHO takes the cooperative behavior and chasing style of Harris' hawks in nature called surprise pounce and can effectively search the optimal one in large parameter space, but it tends to fall into local optimum. To solve this issue, an improved Harris hawk optimization algorithm (IHHO) was proposed to obtain the optimal parameters of SVM. First, Hammersley sequence initialization is carried out to acquire good initial solutions. Then, a nonlinear factor control mode and an adaptive Gaussian–Cauchy mutation perturbation strategy are proposed to avoid getting trapped in local optima. In this way, a novel wind power outlier detection method named IHHO-SVM was constructed. The results on several wind power data with outliers show that IHHO-SVM outperforms SVM and HHO-SVM, which achieves the highest average *F*1 score of 96.63% and exhibits the smallest standard deviation. Compared to commonly used models for detecting outliers in wind power, such as isolation forest (IF), local outlier factor (LOF), SVM with grey wolf optimization (GWO-SVM), and SVM with particle swarm optimization (PSO-SVM), the proposed IHHO-SVM model shows the best overall performance with precision, recall, and *F*1 scores of 95.76%, 96.94%, and 96.35%, respectively.

Keywords: wind power; outlier detection; support vector machine; Harris hawk optimization

Citation: Huang, J.; Qin, J.; Song, S. A Novel Wind Power Outlier Detection Method with Support Vector Machine Optimized by Improved Harris Hawk. *Energies* **2023**, *16*, 7998. https://doi.org/10.3390/en16247998

Academic Editors: Sonia Leva, Emanuele Ogliari and Alessandro Niccolai

Received: 6 October 2023
Revised: 27 November 2023
Accepted: 9 December 2023
Published: 10 December 2023

1. Introduction

The widespread utilization of renewable energy has led to the rapid development of new power systems predominantly relying on clean energy sources globally [1,2]. As a main type of renewable energy, wind power continues to improve its penetration rate in current power systems [3,4]. However, due to the strong volatility and randomness of wind power, wind power prediction has become crucial to help the efficient scheduling and resource optimization of the power system to ensure a stable supply of electric energy.

During the operation of wind power units, factors such as complex environmental conditions, equipment failures, and operational errors can lead to outlier power data [5]. Outliers will mislead the wind power prediction model and decrease the prediction accuracy. If the outliers can be detected accurately and replaced in some way, the power data for prediction model construction are closer to the actual true case, which can make a solid foundation for data-driven prediction [6].

As the inherent characteristics of wind power data, the data are often subject to sudden changes. This means that traditional detection methods often struggle to identify outliers accurately due to rapidly changing data, posing challenges to the precise detection

of anomalies in wind power [7]. Methods for detecting outliers in wind power can be categorized into statistical analysis and machine learning approaches [8]. Statistical analysis methods primarily include the quartile method and statistical metrics method [9–11]. In [9], the quartile method is utilized to identify outlier values with different distribution characteristics, which is efficient but not effective when dealing with a high proportion of outlier data. In [10], a Gaussian mixture copula model is proposed to identify outlier values, but it cannot effectively clean stacked outlier data. Machine learning methods mainly include isolation forest (IF), local outlier factor (LOF), and SVM [12–17]. In [17], a generalized isolation forest was proposed, which can achieve high accuracy in anomaly detection tasks and improve detection speed, but the accuracy of detecting stacked outliers in wind power data is insufficient. In [18], the LOF algorithm is used to identify outlier values in wind power operational data. However, for complex datasets with stacked outlier values, the LOF algorithm may fail to detect them accurately, and its parameters will impact the accuracy. The SVM classification algorithm has demonstrated excellent performance in handling nonlinear and imbalanced data [19,20] and is more suitable for dealing with wind power outliers with randomness and volatility [21].

When identifying wind power outliers using SVM, careful selection of SVM model appropriate hyper-parameters plays a crucial role in achieving optimal outlier classification performance of outliers [22,23]. In the field of SVM hyper-parameters optimization research, scholars have adopted various intelligent search algorithms, including particle swarm optimization (PSO), grey wolf optimization (GWO), and Harris hawks optimization (HHO) [24–29]. In [30], the advantages and disadvantages of the HHO algorithm, application fields, and improved variants are outlined. That is, the diversity of the population and convergence accuracy are maintained through certain strategies, so that the algorithm can play the best role in the application. In [28], the HHO algorithm with two update strategies is used to optimize the hyper-parameters of the SVM to improve the classification accuracy; although the SVM models optimized by the HHO algorithm have shown good performance in outlier detection, they still have the disadvantage of easily falling into local optimum, resulting in low accuracy and stability of the model.

To obtain the optimal parameters of SVM more efficiently, an improved HHO is proposed in this work. Firstly, Hammersley sequence initialization is employed to provide the algorithm with higher-quality initial solutions at the beginning of the iteration, thus enhancing the optimization efficiency. Secondly, a nonlinear factor control mode is proposed to increase the population's opportunities for performing global search. Finally, an adaptive Gaussian–Cauchy mutation perturbation strategy is incorporated to maintain diversity during the population optimization process. Subsequently, the improved IHHO algorithm is utilized to optimize the hyper-parameters of the SVM and construct an IHHO-SVM model for wind power outlier detection. Experiment results show that the proposed IHHO-SVM wind power outlier detection model mitigates the issue of easily getting trapped in local optima during the optimization process. It demonstrates a robust capacity to escape local extrema, thus significantly enhancing the accuracy, stability, and generalization performance of outlier detection in wind power systems.

The proposed outlier detection based on the IHHO-SVM model has the following features:

(1) The hyper-parameters of SVM are initialized by the Hammersley sequence, which ensures a better initial solution at the beginning of the iteration.
(2) A novel nonlinear factor control strategy is designed to make SVM hyper-parameters explore the parameter space globally with a greater chance, which helps to find the global optimal solution.
(3) An adaptive Gauss–Cauchy mutation strategy is proposed to perturb the local optimal solutions to help them jump out of the potential local optimum, which can improve global optimization performance.

2. Outlier Data Distribution in Wind Turbines

2.1. Wind Speed–Power Curve

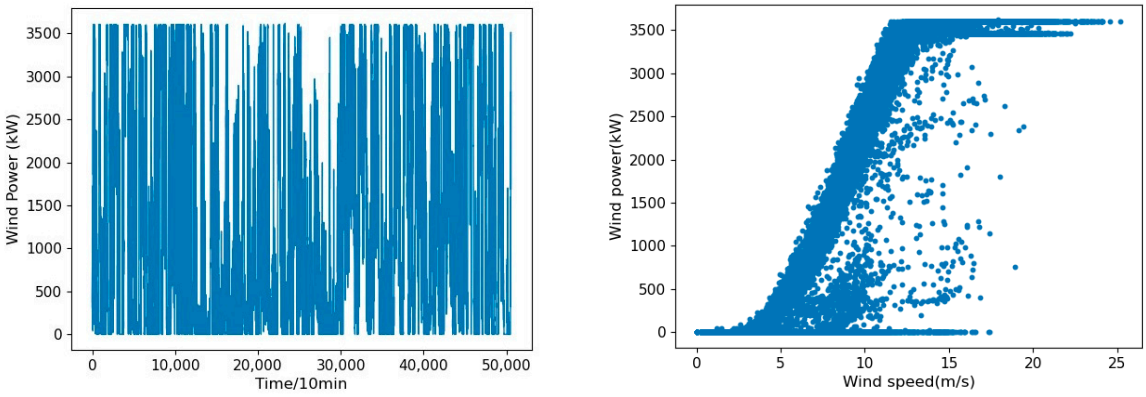
Wind power generation is a renewable energy technology that converts wind energy into electrical energy. In wind power data, the wind speed–power curve is mainly used to analyze the relationship between wind speed and power. Between the cut-in wind speed (the minimum wind speed at which the wind turbine can generate power) and the rated wind speed (the minimum wind speed required to achieve rated power generation), the power output of the wind turbine is proportional to the cube of the wind speed, indicating that the wind speed is the most significant factor influencing power generation. In the ideal state, the output power of the wind turbine can be represented by Equation (1):

$$P = \begin{cases} 0 & v < v_{in} \\ \frac{1}{2} C_P \rho A v^3 & v_{in} \le v < v_n \\ P_n & v_n \le v \le v_{out} \\ 0 & v > v_{out} \end{cases} \quad (1)$$

where v represents the actual wind speed, v_{in} is the cut-in wind speed, v_{out} is the cut-out wind speed, v_n is the rated wind speed, P represents the output power of the wind turbine, and P_n is the rated power.

2.2. Characteristics of Outlier Data Distribution

Wind turbines generally operate in a more complex environment, and wind power data may be affected by conditions, such as adverse weather, unit failures, wind curtailment, and power rationing. Therefore, the wind power distribution sampled by the SCADA system presents high volatility, strong randomness, and intermittency in the time series, as shown in Figure 1a. These factors constitute an important factor leading to outliers in wind power data.



(**a**) (**b**)

Figure 1. Distribution of wind power and wind speed. (**a**) Line diagram of wind power distribution; (**b**) Distribution points of wind speed and power.

By analyzing the scatter diagram of wind speed-power distribution in Figure 1b, these data points can be divided into the following three categories: normal power points, stacked outliers, and discrete outliers. The distribution of normal power points is regular and is positively correlated with the cubic velocity. The accumulation outlier is generally caused by factors such as unit shutdown and power curtailment in the power grid. When the wind speed is greater than zero, the power is at a low level, which is not affected by the wind speed and is generally distributed horizontally at the bottom of the wind power scatter diagram. Discrete outliers are usually caused by signal transmission noise, extreme weather, and other reasons. Compared with stacked outliers, they have strong

uncontrollability and randomness and are usually randomly distributed near the wind power curve.

Wind power data with outliers cannot accurately reflect the real output of the generator unit, which can impact the effectiveness of state monitoring and power prediction for wind turbine units. This, in turn, can lead to unstable system operation and pose risks of overload or inadequate power supply. Therefore, it is crucial to accurately detect outliers in wind power.

3. Principle of the IHHO-SVM Wind Power Outlier Detection Model

3.1. Support Vector Machine

The idea behind support vector machine classification is to find an optimal hyperplane that separates samples of different classes and maximizes the margin between the two classes.

For nonlinear separable sample sets, the classification problem can be formulated as a convex quadratic programming problem, as shown in Equation (2).

$$\begin{cases} \min\limits_{\omega,b,\xi_i} \frac{1}{2}||\omega||_2 + C\sum\limits_{i=1}^{m} \xi_i,\ i = 1,2,\ldots,m \\ s.t. y_i(\omega^T f(x_i) + b) \geq 1 - \xi_i \end{cases} \tag{2}$$

where ω represents the normal vector of the hyperplane, C is the penalty coefficient, m is the number of samples, $y_i \in \{-1, 1\}$ is the class label, $\xi_i (\xi_i \geq 0)$ is the slack variable, and $f(x_i)$ is the mapping function.

By introducing the Lagrange function, the dual principle, and the SVM kernel functions, the decision function can be obtained as shown in Equation (3).

$$f(x, \alpha) = \mathrm{sgn}(\sum_{i=1}^{m} y_i \alpha_i K(x_i, y_i) + b) \tag{3}$$

$$K(x_i, x_j) = \exp(\frac{-||x_i - x_j||^2}{2g^2}) \tag{4}$$

where $\alpha_i (\alpha_i \geq 0)$ is the Lagrange multiplier, $K(x_i, x_j)$ is the Gaussian kernel function, and g is the parameter of the kernel function.

3.2. Harris Hawks Optimization Algorithm

The HHO simulates the predation process within a group of Harris hawks to search for the optimal solution through individual collaboration and competition. This algorithm consists of three stages: the exploration stage, the transition from the exploration to the exploitation stage, and the exploitation stage.

3.2.1. Exploration

When searching for the position of prey, Harris hawks update their position as follows:

$$X(t+1) = \begin{cases} X_{\mathrm{rand}}(t) - r_1 |X_{\mathrm{rand}}(t) - 2r_2 X(t)| & q \geq 0.5 \\ (X_{\mathrm{rabbit}}(t) - X_{\mathrm{mean}}(t)) - r_3(lb + r_4(ub - lb)) & q < 0.5 \end{cases} \tag{5}$$

$$X_{mean}(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t) \tag{6}$$

where t is the current iteration number; X_{rand} is the position of the randomly selected individual in the current population; X_{mean} is the average position of the current population; X_{rabbit} is the best individual in the current population; r_1, r_2, r_3, r_4 and q are random numbers inside (0,1); ub and lb are the upper and lower boundaries of the population exploration space; N denotes the population size; and X_i represents the position of a Harris hawk.

3.2.2. Transition from Exploration to Exploitation

The transition behavior of a Harris hawk from the global exploration phase to the local phase is simulated as the prey energy attenuation process. The mathematical expression is as follows:

$$E = 2(1 - \frac{t}{T}) \tag{7}$$

$$E_1 = EE_0 \tag{8}$$

where T represents the maximum number of iterations, E represents the escape energy factor of the prey, E_0 is a random number in the range of $(-1, 1)$, and E_1 represents the escape energy of the prey, which ranges from $(-2, 2)$. When $|E_1| \geq 1$, global exploration is performed. When $|E_1| < 1$, local development is conducted.

3.2.3. Exploitation

In this phase, the Harris hawk determines four possible attack strategies based on the prey's escape probability r and escape energy E_1.

1. Soft Besiege

When $r \geq 0.5$ and $|E_1| \geq 0.5$, the Harris hawk uses the soft Besiege mode for encircling prey. Its mathematical expression is as follows:

$$X(t+1) = \Delta X(t) - E_1|JX_{rabbit}(t) - X(t)| \tag{9}$$

$$\Delta X(t) = X_{rabbit}(t) - X(t) \tag{10}$$

$$J = 2(1 - r_5) \tag{11}$$

where $\Delta X(t)$ represents the difference between the current population and prey position at iteration t, J represents the prey's jumping distance, and r_5 is a random number within the range of (0,1).

2. Hard Besiege

When $r \geq 0.5$ and $|E_1| < 0.5$, the Harris hawk uses the hard Besiege mode for encircling prey. Its mathematical expression is as follows:

$$X(t+1) = X_{rabbit}(t) - E_1|\Delta X(t)| \tag{12}$$

3. Soft Besiege with Progressive Rapid Dives

When $r < 0.5$ and $|E_1| \geq 0.5$, the Harris hawk needs to adopt a progressive rapid dive with a soft besiege attack strategy. Its mathematical expression is as follows:

$$Y = X_{rabbit} - E_1|JX_{rabbit}(t) - X(t)| \tag{13}$$

$$Z = Y + S \times LF(D) \tag{14}$$

$$X(t+1) = \begin{cases} Y & F(Y) < F(X(t)) \\ Z & F(Z) < F(X(t)) \end{cases} \tag{15}$$

where LF represents the Levy flight function, S is a random vector of dimension D, D represents the dimensionality of the problem, and F denotes the fitness function of the problem.

4. Hard Besiege with Progressive Rapid Dives

When $r < 0.5$ and $|E_1| < 0.5$, the Harris hawk attempts to increase the hunting success rate by reducing the average distance to the prey. The mathematical expression is as follows:

$$Y = X_{rabbit}(t) - E_1|JX_{rabbit}(t) - X_{mean}(t)| \tag{16}$$

$$X(t+1) = \begin{cases} Y & F(Y) < F(X(t)) \\ Z & F(Z) < F(X(t)) \end{cases} \tag{17}$$

3.3. Improved Harris Hawks Optimization (IHHO)

3.3.1. Hammersley Sequence Initialization Populations

Random initialization in the HHO algorithm often leads to uneven population distribution, clustering, or stacking, which adversely affects the quality of initial solutions. Furthermore, during the early stage of global exploration, the algorithm may have limited coverage and fail to adequately explore potential optimal solution regions. To address this issue and improve search traversal, this paper proposes using the Hammersley sequence for population initialization. Hammersley is a uniformly distributed point sequence with low difference properties. Compared with traditional random number generation methods, the Hammersley sequence can cover the whole space more uniformly when filling high-dimensional space. This feature makes the Hammersley sequence more efficient when optimizing the initial population of the algorithm and also improves the distribution of the population in the initial space to be more uniform, which helps the algorithm to obtain better initial values.

The Hammersley sequence maps integer indices to values in different dimensions, generating a set with lower discrepancy. The main steps are as follows:

1. Determine any natural number n by a polynomial of the given prime p:

$$n = \sum_{i=0}^{m} a_i p^i = a_m p^m + \ldots + a_2 p^2 + a_1 p^1 + a_0 p^0 \tag{18}$$

   where $a_i \in [0, p-1]$.
2. Reverse the coefficients a_i in order and mirror them to the right of the decimal point, then calculate their value.

$$\boldsymbol{\phi}_p(n) = a_0 p^{-1} + a_1 p^{-2} + \ldots + a_m p^{-m-1} \tag{19}$$

3. Set the dimension to d and obtain the values of the Hammersley sequence.

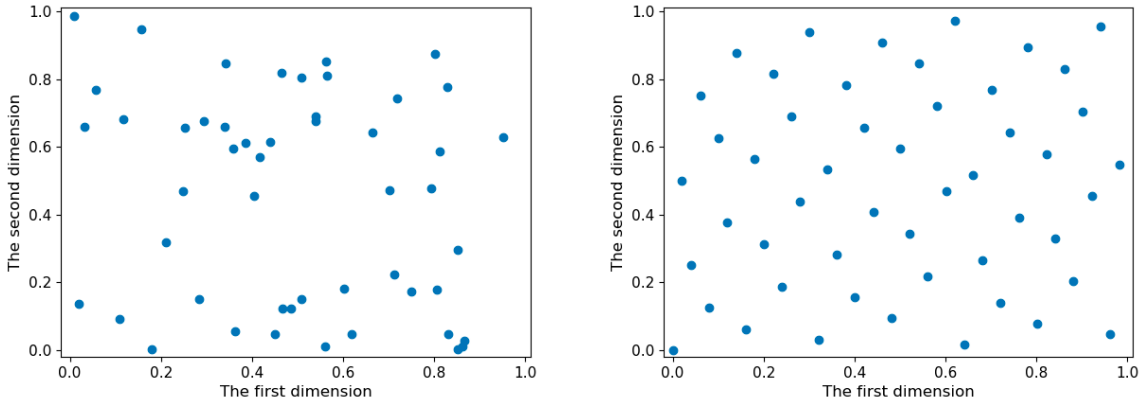
$$H(n) = (\frac{n}{N}, \boldsymbol{\phi}_{p_1}(n), \boldsymbol{\phi}_{p_2}(n), \ldots, \boldsymbol{\phi}_{p_{d-1}}(n)) \tag{20}$$

   where N represents the number of sample points, and p is a prime number determined based on the dimension; $n = 0, 1, 2 \ldots, N-1$.

Figure 2a illustrates the issues related to random initialization, such as clustering, stacking, and uneven distribution. However, Figure 2b shows the effectiveness of employing the Hammersley sequence as an initialization method to achieve a uniform distribution across the spatial extent. This approach takes better consideration of each region, which is beneficial for optimization algorithms to obtain higher-quality initial solutions.

3.3.2. Nonlinear Factor Control Mode

The transition from HHO's global exploration to local exploitation is controlled by the prey's escape energy. In the early stages of iteration, the probability of linear escape energy for global exploration gradually decreases until the middle and later stages of iteration, where the population only focuses on local exploitation. This can easily lead the algorithm to get stuck in a local optimum. In order to overcome these limitations, we consider the periodicity and radian of the inverse triangular function, and the increase and decrease in the exponential function, and propose a nonlinear escape energy update strategy, which enables HHO to have a greater probability of global exploration in the whole iteration process and rapid local exploration in the later stage. The proposed strategy is shown in Equation (21).



(**a**) (**b**)

Figure 2. Comparison of Hammersley population generation and random population generation. (**a**) Random population generation; (**b**) Hammersley population generation.

$$E^* = 2 \times (\arccos\frac{t}{T} \times \frac{2}{\pi})^k \tag{21}$$

$$E_2 = E^* E_0 \tag{22}$$

where E_2 represents the nonlinear escape energy range $(-2, 2)$. E^* is the nonlinear escape energy factor, when $E^* \geq 1$, global exploration is conducted; when $E^* < 1$, local exploitation is performed. k is the decay rate coefficient with $0 < k < 1$. When k is close to 0, it can lead to insufficient local exploration in the later stages of iteration. When k is close to 1, it can result in inadequate global search capability in the later stages of iteration. To balance the local exploitation and global search ability throughout the algorithm iteration process, this paper sets $k = 0.8$.

As shown in Figure 3, the proposed nonlinear update strategy improves the limitations of the original algorithm, which only conducts global exploration after the mid-iteration stage and has a relatively low probability of global exploration in the early stage. This strategy prevents the early stagnation of the algorithm and facilitates a more extensive exploration of the global optimal solution.

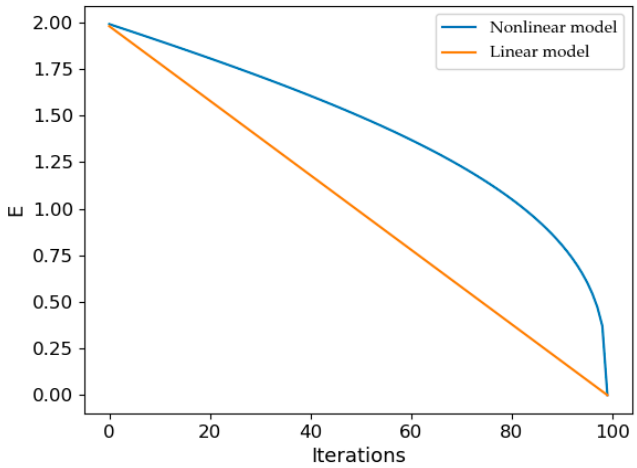


Figure 3. Comparison of two escape energy update strategies.

3.3.3. Adaptive Gaussian–Cauchy Perturbation Strategy

In the HHO algorithm, the population tends to converge towards the region of the current best solution, which reduces the diversity of the population and can lead to the

problem of the algorithm getting trapped in local optima. In order to overcome this problem, we consider the characteristics of Gaussian and Cauchy distributions, set an adaptive weight coefficient by increasing the number of iterations, and use this strategy to perturb the optimal individual that remains unchanged for two successive iterations, as shown in Equations (23) and (24). If the perturbed position of the individual improves on the current best position, it will be incorporated in the next iteration.

$$X(t+1) = X^*_{rabbit}(t)[1 + (1 - \frac{t^2}{T^2})Cauchy(0,1) + \frac{t^2}{T^2}Gauss(0,1)] \tag{23}$$

$$X_{rabbit}(t+1) = \begin{cases} X(t+1) & F(X(t+1)) < F(X_{rabbit}(t)) \\ X_{rabbit}(t) & F(X_{rabbit}(t)) < F(X(t+1)) \end{cases} \tag{24}$$

where X^*_{rabbit} represents the position information of an individual whose fitness value remains unchanged for two consecutive iterations; $X(t+1)$ denotes the individual's position after Gaussian–Cauchy perturbation; $Cauchy(0,1)$ represents the standard Cauchy distribution; $Gauss(0,1)$ represents the standard Gaussian distribution; $X_{rabbit}(t+1)$ represents the updated position of the best individual; F represents the fitness value of the problem.

In Equation (23), the standard Cauchy distribution, with its wide numerical distribution range, can induce significant disturbances to individual positions. Thus, in the early stages, it is assigned a higher weight to escape local optima. Conversely, the standard Gaussian distribution possesses a more concentrated numerical distribution, resulting in smaller perturbation values. Therefore, in the later stages of iteration, it is assigned a higher weight to facilitate the population to explore the vicinity of the current best individual and discover the optimal solution.

Figure 4 shows the flow chart of the IHHO algorithm. By integrating the above three enhancement strategies, the limitations of inadequate global search and susceptibility to local optima in the original HHO algorithm are mitigated, resulting in a more precise identification of the global optimal solution.

3.4. IHHO-SVM Wind Power Outlier Detection Model

In the detection of wind power outliers with SVM, the parameters C and g play a crucial role in determining the final detection results. A larger value of C tends to favor overfitting, while a smaller value tends to favor generalization ability. A larger value of g focuses more on the local data structure, which may lead to overfitting. Conversely, a smaller value makes the model pay more attention to the global data structure, which can result in underfitting.

The IHHO-SVM wind power outlier detection model uses the parameters C and g of SVM as position information of individuals in the population. The fitness value is calculated to determine whether it is the optimal model. In the outlier detection task, the precision, recall, and $F1$ score can reflect the performances from different views; so in order to evaluate the overall performance, the fitness function is constructed as follows:

$$F = 3 - (P + R + F1) \tag{25}$$

where P represents precision, R represents recall, and $F1$ represents $F1$ score. When precision, recall, and $F1$ score are close to 1, it indicates that the model can accurately identify anomalies while maintaining low false-positive and false-negative rates. To facilitate the observation and comparison of experimental results, a constant of 3 is introduced in the formula.

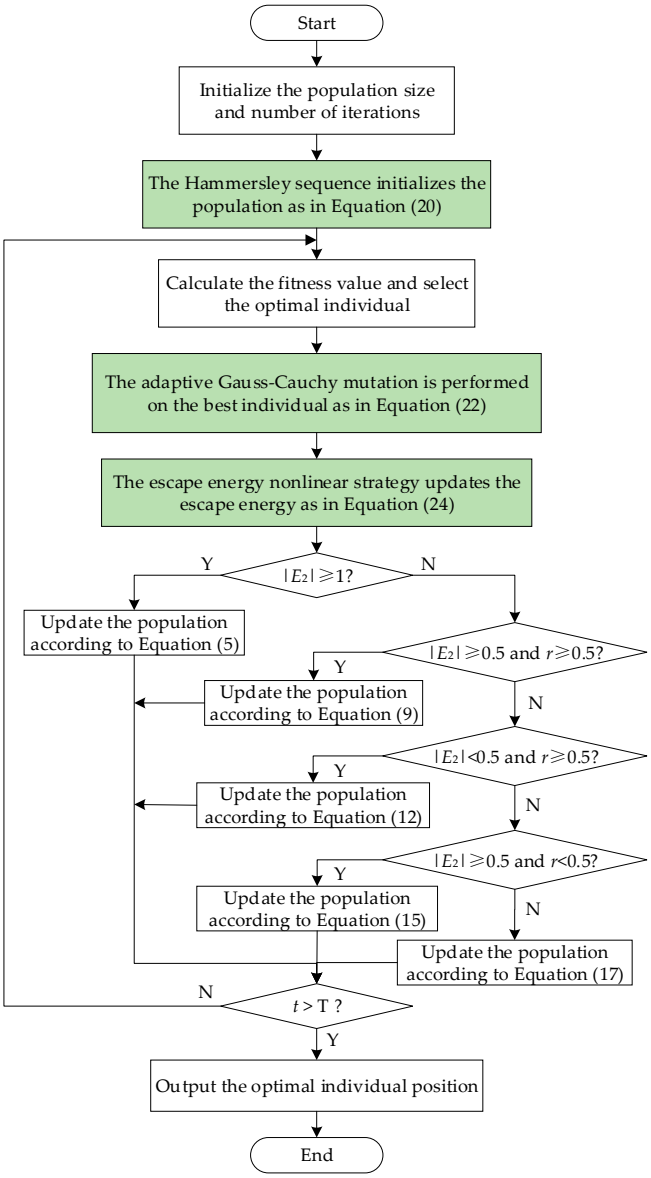


Figure 4. Flow chart of IHHO.

The wind power outlier detection model based on IHHO-SVM is depicted in Figure 5. It consists of two main components. Firstly, the wind turbine data are preprocessed as part of the process. Secondly, the parameters of the SVM model are optimized utilizing IHHO. Finally, the outlier value of wind power is detected.

In the wind power outlier detection algorithm based on the IHHO-SVM, several steps are performed. Firstly, the wind power data are divided and normalized, and the parameters of IHHO are set, including the maximum number of iterations and the number of populations. Then, the population is initialized using the Hammersley sequence, taking into account the value range of SVM hyper-parameters C and g. Subsequently, the IHHO algorithm is used to optimize the SVM parameters, and the fitness value is obtained by 5-fold cross-validation of the model in the training samples. The process assesses whether the maximum number of iterations is reached and determines the optimal SVM model with the minimum fitness. Finally, the test samples are fed into the optimal SVM model to obtain the results of anomaly detection.

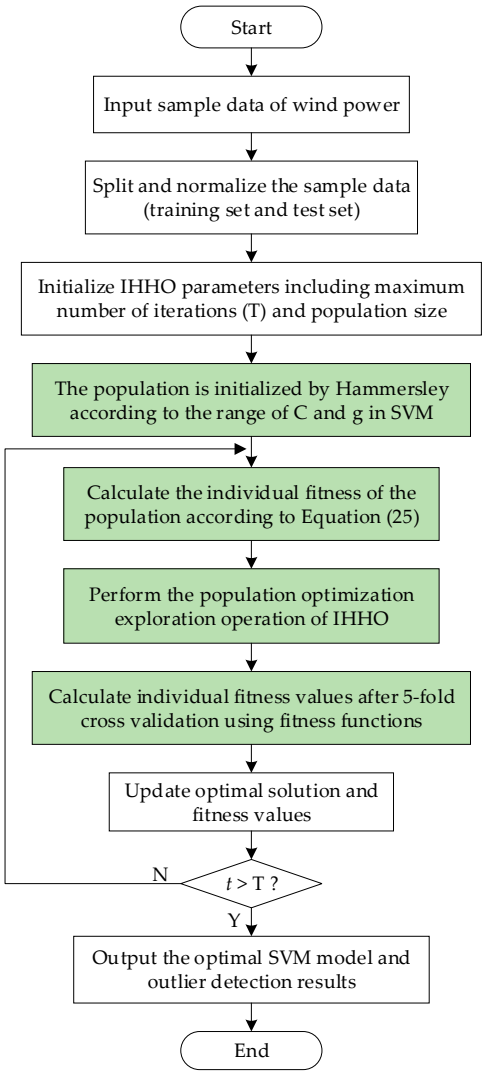


Figure 5. Flow chart of IHHO-SVM.

4. Experiment and Result Analysis

4.1. Experimental Environment and Dataset

Experimental environment: The computer operating system is Windows 10, equipped with an Intel Core i7-11800H CPU and 16GB RAM. Algorithm development is performed using a Python 3.8 interpreter in the PyCharm environment.

The study utilizes data from the SCADA system to evaluate the performance of the proposed IHHO-SVM model. The dataset consists of variables such as wind speed, wind direction, and power generation. These data are sampled every 10 min from the SCADA system of a wind turbine in Turkey. In order to better understand the distribution of data used in the experiment, the index of wind speed in the dataset is given, as shown in Table 1.

Table 1. Wind speed distribution index of experimental data.

Dataset	Sample Size	Minimum (m/s)	Maximum (m/s)	Mean (m/s)	Variance	Standard Deviation
1	10,000	0	25.20	8.85	24.96	4.99
2	2000	0	17.00	5.95	13.44	3.66
3	2000	0.35	18.43	8.51	11.61	3.41
4	2000	0	14.12	6.78	7.70	2.77
5	2000	0	16.55	6.01	14.25	3.77

By analyzing the distribution of wind power data collected by the SCADA system, the actual power data not only strongly correlated with wind speed but also related to random nonlinear factors, such as measuring instrument accuracy (such as temperature, humidity, and air pressure), communication reliability, and control problems. If the actual power deviates from the theoretical value is large enough, we can take it as an outlier. However, the absolute deviation changes large with the wind speed. Thus, we design a threshold as shown in Equation (26).

$$\mathrm{y} = \begin{cases} 1, & \text{if } P_a < 0 \text{ and } v < v_{in} \\ 1, & \text{if } \left|\frac{P_t - P_a}{v}\right| > k \text{ and } v_{in} \le v < v_n \\ 1, & \text{if } \frac{P_t - P_a}{v_n} > \frac{k}{2} \text{ and } v_n \le v \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

where y is the label, 1 is the outlier value, 0 is the normal value, v is the actual wind speed, v_{in} is the cut-in wind speed, v_n is the rated wind speed, P_t is the theoretical power value, P_a is the actual power value, and k is the boundary coefficient. Considering the actual working conditions of the fan and the influence of nonlinear factors, the value of k in this experiment is 60. When $v < v_{in}$, the actual power may be affected by the inertia of the fan blade to output a small amount of power, but the output power cannot be a negative value, so the actual power is less than 0 when the wind speed is less than v_{in}, which is labeled as an outlier. When $v_{in} \le v < v_n$, the actual power is related to the wind speed and other nonlinear factors; therefore, when the deviation between the actual power and the theoretical value is greater than a certain threshold relative to the wind speed, it is labeled as an outlier. When $v_n < v$, the theoretical power is the maximum generation power, which will not change with the change in the current wind speed. In contrast, the actual power will be affected by system loss and other factors, which will only be less than the theoretical power and have small fluctuations; therefore, we set a small threshold to judge whether it is an outlier. In the end, about 4% of outliers are labeled in their annual data.

To enhance the model's training and enable a comprehensive evaluation, a data augmentation strategy is designed to increase the proportion of outliers in the dataset, simulating actual power outliers. A certain amount of data are randomly selected from the actual power labeled as normal, and the data are changed as outliers according to Equation (27):

$$\begin{aligned} & P_{a^*} = P_{\mathrm{a}}(1 + h\%) \\ s.t. \quad & h \sim Gauss(0, 40) \\ & |h| > 25 \end{aligned} \tag{27}$$

where P_{a^*} is to simulate the outlier of the actual power, h follows the Gaussian distribution with the mean of 0 and the variance of 40, and $|h| > 25$ makes the simulated outlier as real as possible.

To mitigate the influence of different data scales, the data are normalized using the min–max normalization method as shown in Equation (28).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{28}$$

where x' is the normalized data, x is the original data, $x_{\min}$ is the minimum value in the original data, and $x_{\max}$ is the maximum value in the original data.

4.2. Evaluation Index

In the problem of outlier detection, commonly used evaluation metrics are precision, recall, and $F1$ score. Their expressions are as follows:

$$P = \frac{TP}{TP + FP} \tag{29}$$

$$R = \frac{TP}{TP + FN} \tag{30}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{31}$$

where TP is the number of correct positive class classifications in the result, while FP and FN are the total number of sample misclassifications in the results.

4.3. Experiments and Discussion

4.3.1. Experiment 1

To validate the advantages of IHHO-SVM in the detection of outliers in wind power, this experiment was annotated and enhanced using 10,000 data points from dataset 1 as the training set, in which the proportion of outliers was about 15%. Separate IHHO-SVM and HHO-SVM wind power outlier detection models were established. The performance of the optimization algorithm was assessed by monitoring variations in average fitness values through 5-fold cross-validation. The relevant initial parameters of the algorithm were set as follows: a population size of 30, maximum iteration of 30, variable dimension of 2, penalty factor C search range of $[1, 10^4]$, and kernel function parameter g search range of $[10^{-3}, 1]$.

As shown in Figure 6, in the context of wind power outlier detection, the IHHO-SVM has a smaller initial fitness value, demonstrating that the proposed Hammersley sequence initialization method can provide higher-quality initial solutions for the model. Throughout the iteration process, the fitness value of IHHO-SVM undergoes multiple changes until it reaches the minimum, signifying that the proposed nonlinear factor control mode and adaptive perturbation strategy help the algorithm in evading local optima and enhance its likelihood of exploring various regions to achieve the global optimum. Therefore, in comparison to the HHO algorithm, the IHHO algorithm demonstrates improved global optimization capability and convergence accuracy during the SVM parameter optimization process.

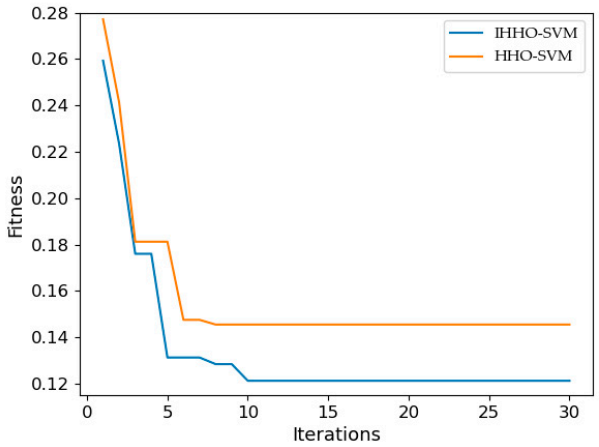


Figure 6. Comparison of fitness changing with HHO-SVM and IHHO-SVM.

4.3.2. Experiment 2

To validate the detection accuracy and generalization performance of IHHO-SVM in the detection of wind power outliers, three groups of labeled wind power datasets in different months were selected as test sets, namely, datasets 2, 3, and 4. Each dataset contains 2000 samples, and then data enhancement generates outliers of different proportions to evaluate the stability of the model. Penalty factor C and the parameter of the kernel function g in the algorithm were obtained from the training model in Experiment 1. The parameter optimization results for the IHHO-SVM model are C = 9801 and g = 1, while for the HHO-SVM model, the parameter optimization results are C = 3649 and g = 0.98.

Table 2 presents the comprehensive *F*1 scores of the IHHO-SVM, HHO-SVM, and SVM models on three test sets. The average *F*1 score of SVM is the highest at 90.39% and the lowest at 84.95%, with the worst standard deviation. This indicates that the parameter settings in the SVM model significantly affect the detection results of outlier data with different proportions. The HHO-SVM model performs better than the SVM model in terms of the average *F*1 values and standard deviations for all three sets, suggesting that the optimization of the HHO-based SVM hyper-parameters overcomes the limitations of manually selecting parameters. Furthermore, the average *F*1 scores of the IHHO-SVM model are higher than those of the HHO-SVM model, reaching a maximum of 97.16%. This indicates that the HHO-SVM model has fallen into local extremum points when optimizing parameters, while the IHHO-SVM model improves the ability to jump out of local extremum points during parameter optimization by nonlinear factor control mode and adaptive Gaussian–Cauchy mutation perturbation strategy, resulting in the best optimization effect and improving the accuracy of outlier detection. Furthermore, the standard deviation of the IHHO-SVM model is the smallest at 0.18 and consistently the best in all three test sets, indicating its low sensitivity to imbalanced distributions of outlier values. In summary, the IHHO-SVM model demonstrates the best detection accuracy, as well as strong stability and generalization performance.

Table 2. Performance of the models on three test sets with different proportions of outliers.

Dataset	Model	*F*1 Score(%)						
		Outlier Ratio					Mean Value	Standard Deviation
		5%	10%	15%	20%	25%		
2	SVM	82.61	87.05	84.95	84.92	85.23	84.95	1.58
	HHO-SVM	92.00	92.93	94.20	93.64	93.76	93.31	0.86
	IHHO-SVM	94.23	95.27	94.99	94.37	94.66	94.70	0.18
3	SVM	90.09	87.32	88.40	89.40	89.61	88.96	1.11
	HHO-SVM	96.61	96.44	95.70	95.50	94.95	95.84	0.69
	IHHO-SVM	97.15	96.77	96.34	96.04	95.48	96.36	0.64
4	SVM	91.33	91.08	90.23	89.32	89.99	90.39	0.82
	HHO-SVM	96.29	95.12	96.51	95.42	95.74	95.82	0.58
	IHHO-SVM	97.05	96.07	97.16	96.61	96.28	96.63	0.47

4.3.3. Experiment 3

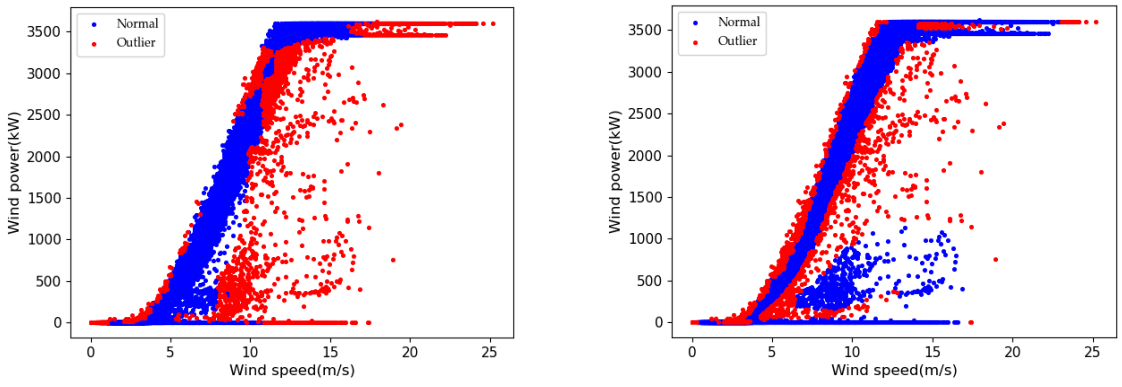
In order to further validate the performance of the IHHO-SVM model in wind power outlier detection, comparative test was conducted by selecting commonly employed machine learning models for outlier detection, including isolation forest (IF) [17], local outlier factor (LOF) [18], SVM, as well as combined models of SVM with widely applied optimization algorithms, such as GWO-SVM [27], PSO-SVM [24], and HHO-SVM, in comparison with the proposed IHHO-SVM model. The test uses the labeled dataset 5, which contains 2000 samples with an outlier content of about 4%, and the results are given in Table 3.

Table 3. Comparison of outlier detection performance with different models.

Model	Precision (%)	Recall (%)	*F*1 Score(%)
IF	83.03	51.89	63.86
LOF	85.11	54.16	66.20
SVM	91.56	89.48	90.51
GWO-SVM	95.31	95.21	95.22
PSO-SVM	95.24	94.08	94.66
HHO-SVM	95.70	95.51	95.60
IHHO-SVM	95.76	96.94	96.35

As shown in Table 3, compared to IF and LOF, SVM achieved the highest evaluation metric values, making it more suitable for outlier detection tasks. However, SVM recall precision and *F*1 score were only 89.48%, 91.56%, and 90.51%, respectively. This phenomenon could be attributed to the artificial selection of hyper-parameters. Nevertheless, based on the SVM model of the general optimization algorithm, the overall *F*1 score is up to 95.60%, which reduces the randomness of hyperparameter selection and improves the overall detection accuracy. In addition, the IHHO-SVM model proposed in this paper is compared with the SVM model based on general optimization algorithms. It is found that the precision of the IHHO-SVM model is 95.76%, the recall is 96.94%, and the *F*1 score is 96.35%. Compared with the GWO-SVM, PSO-SVM, and HHO-SVM, the accuracy rate of the IHHO-SVM model is the best. It shows that the general optimization algorithm is easy to fall into local optimal when SVM parameters are optimized. In the IHHO-SVM proposed in this paper, Hammersley sequence initialization is used to reduce the possibility of falling into local extreme value in the initial optimization, nonlinear factor control mode, and adaptive Gaussian–Cauchy mutation perturbation strategy, which make the model easily jump out of local extreme values and find better parameters in the parameter optimization process. Therefore, IHHO-SVM shows high accuracy in wind power outlier detection.

In order to evaluate the performance of IHHO-SVM in a long span, the detection performance of five algorithms in one-year wind power data was visualized, as shown in Figure 7. It is evident that these five detection algorithms can broadly detect outliers in raw SCADA data. In terms of detection performance, the IHHO-SVM proposed in this paper performs the best in wind power data. In comparison, LOF and IF algorithms generate more false alarms for data points near the edge of normal data. Additionally, the LOF algorithm exhibits lower sensitivity to stacked outliers and partial discrete outliers at the bottom of wind power data, resulting in more cases of missed detection, while the IF algorithm performs poorly in detecting discrete outliers near the middle and bottom of normal data. Although SVM accurately identified the main portion of normal wind power data, some outlier data points in the upper middle of the curve were missed due to the artificial determination of the hyperparameters. Compared with the IHHO-SVM algorithm, the HHO-SVM algorithm has some missing detection near the middle edge and upper part of wind power data. Therefore, comparing the results of IF, LOF, and SVM, it can be seen that SVM has the best effect on processing data with nonlinear characteristics. According to the results of the SVM outlier detection model based on the optimization algorithm, in IHHO, Hammersley sequence initialization, nonlinear factor control mode, and adaptive Gaussian–Cauchy mutation perturbation strategy improve the optimization of hyper-parameters in the SVM model to avoid falling into a local optimal solution. Therefore, it achieves high precision, good stability, and high generalization performance in wind power anomaly detection task and provides strong support for wind power condition monitoring and wind power prediction.



(**a**) (**b**)

Figure 7. *Cont.*

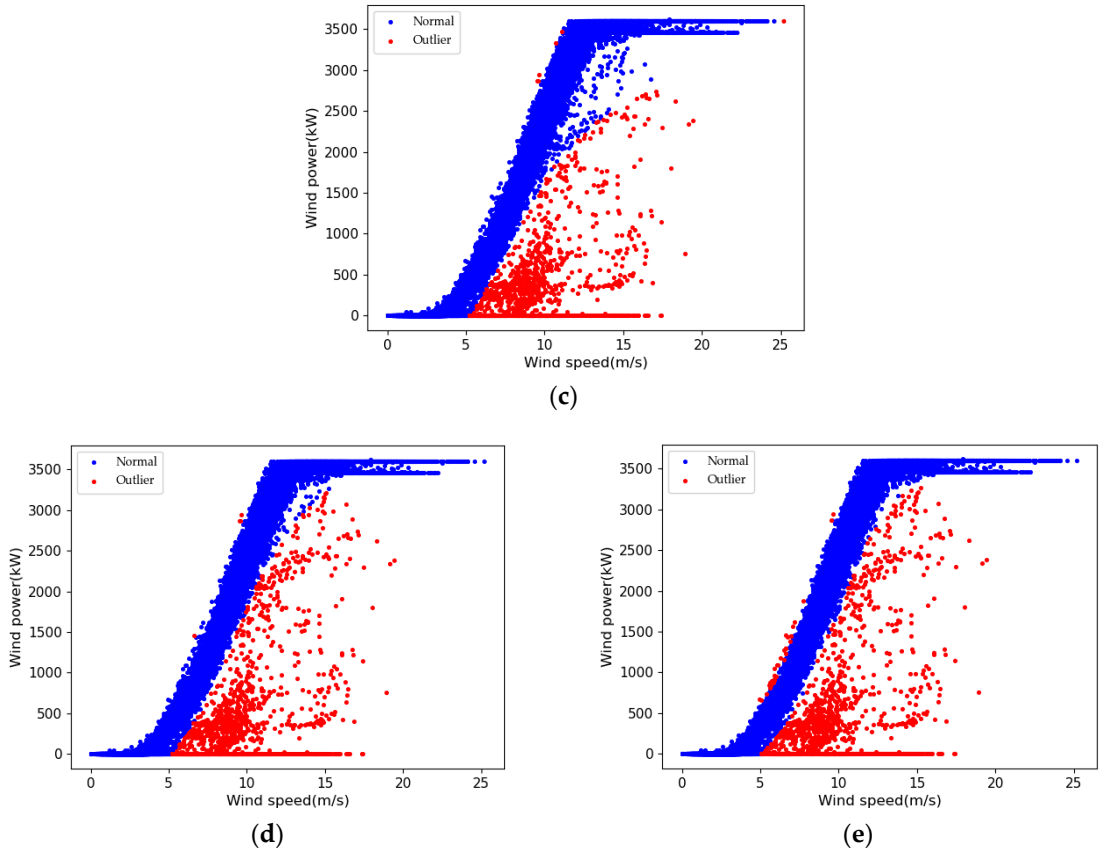


Figure 7. Outlier detection results of wind power data. (**a**) IF; (**b**) LOF; (**c**) SVM; (**d**) HHO-SVM; (**e**) IHHO-SVM.

5. Conclusions

Aiming to address the issue of wind power outlier detection, this paper establishes an SVM-based wind power outlier detection model using the IHHO algorithm. The model is built on the HHO algorithm as its foundation. Initially, the hyper-parameters are initialized using the Hammersley sequence, ensuring a superior initial solution during the iterative process. Subsequently, the algorithm's global exploration ability and optimization accuracy are enhanced by introducing a nonlinear factor control mode and an adaptive Cauchy–Gaussian perturbation strategy. Finally, these techniques are applied to optimize the parameters C and g of the SVM, resulting in an optimized model for the outlier detection model. The experimental results show that IHHO-SVM maintains higher detection scores than other outlier detection models in multiple sets of test results.

In conclusion, the IHHO-SVM model demonstrates high accuracy in the task of wind power outlier detection, while also exhibiting commendable generalization performance and stability. As such, it holds significant relevance in the realms of wind power curve modeling and power prediction tasks. However, it should be noted that this method primarily performs offline training using historical data and can identify outlier points based on fundamental characteristics. Nevertheless, when a wind turbine is operating in real-time, it may not be possible to accurately identify these outliers when extreme operating conditions are present and accompanied by outliers that produce new features. So, the investigation of online training with real-time data and continuous model updates is an important issue, which will aid in addressing the real-time outlier detection of wind turbines under exceptional conditions and can provide more technical support for ultra-short-term wind power prediction.

Author Contributions: Conceptualization, S.S.; formal analysis, J.H.; funding acquisition, J.H.; investigation, J.Q.; methodology, J.H. and J.Q.; project administration, S.S.; software, J.Q.; supervision, J.H. and S.S.; visualization, J.Q.; writing—original draft, J.Q.; writing—review and editing, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Nature Science Foundation of China, grant number U1504617.

Data Availability Statement: The data used in this work are from the publicly archived datasets on Kaggle and are available at https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset, accessed on 7 May 2022.

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

References

1. Vargas, S.A.; Esteves, G.R.T.; Maçaira, P.M.; Bastos, B.Q.; Cyrino Oliveira, F.L.; Souza, R.C. Wind Power Generation: A Review and a Research Agenda. *J. Clean. Prod.* **2019**, *218*, 850–870. [CrossRef]
2. Ahmad, T.; Madonski, R.; Zhang, D.; Huang, C.; Mujeeb, A. Data-Driven Probabilistic Machine Learning in Sustainable Smart Energy/Smart Energy Systems: Key Developments, Challenges, and Future Research Opportunities in the Context of Smart Grid Paradigm. *Renew. Sustain. Energy Rev.* **2022**, *160*, 112128. [CrossRef]
3. Sundarapandi Edward, I.E.; Ponpandi, R. Challenges, Strategies and Opportunities for Wind Farm Incorporated Power Systems: A Review with Bibliographic Coupling Analysis. *Env. Sci. Pollut Res.* **2023**, *30*, 11332–11356. [CrossRef] [PubMed]
4. Dessouky, S.S.; Abdellatif, W.S.E.; Abdelwahab, S.A.M.; Ali, M.A. Maximum Power Point Tracking Achieved of DFIG-Based Wind Turbines Using Perturb and Observant Method. In Proceedings of the 2018 Twentieth International Middle East Power Systems Conference (MEPCON), Cairo, Egypt, 18–20 December 2018; pp. 1121–1125.
5. Wang, S.; Huang, Y.; Li, L.; Liu, C. Wind Turbines Abnormality Detection through Analysis of Wind Farm Power Curves. *Measurement* **2016**, *93*, 178–188. [CrossRef]
6. Wang, Y.; Hu, Q.; Li, L.; Foley, A.M.; Srinivasan, D. Approaches to Wind Power Curve Modeling: A Review and Discussion. *Renew. Sustain. Energy Rev.* **2019**, *116*, 109422. [CrossRef]
7. Morrison, R.; Liu, X.; Lin, Z. Anomaly Detection in Wind Turbine SCADA Data for Power Curve Cleaning. *Renew. Energy* **2022**, *184*, 473–486. [CrossRef]
8. Boukerche, A.; Zheng, L.; Alfandi, O. Outlier Detection: Methods, Models, and Classification. *ACM Comput. Surv.* **2020**, *53*, 1–37. [CrossRef]
9. Shen, X.; Fu, X.; Zhou, C. A Combined Algorithm for Cleaning Abnormal Data of Wind Turbine Power Curve Based on Change Point Grouping Algorithm and Quartile Algorithm. *IEEE Trans. Sustain. Energy* **2019**, *10*, 46–54. [CrossRef]
10. Wang, Y.; Infield, D.G.; Stephen, B.; Galloway, S.J. Copula-Based Model for Wind Turbine Power Curve Outlier Rejection. *Wind. Energy* **2014**, *17*, 1677–1688. [CrossRef]
11. Zhao, Y.; Ye, L.; Wang, W.; Sun, H.; Ju, Y.; Tang, Y. Data-Driven Correction Approach to Refine Power Curve of Wind Farm Under Wind Curtailment. *IEEE Trans. Sustain. Energy* **2018**, *9*, 95–105. [CrossRef]
12. Chen, K.; Wang, H.; Ying, Z.; Zhang, C.; Wang, J. Online Cleaning Method of Power Grid Energy Anomaly Data Based on Improved Random Forest. *J. Phys. Conf. Ser.* **2021**, *2108*, 012067. [CrossRef]
13. Alghushairy, O.; Alsini, R.; Soule, T.; Ma, X. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. *Big Data Cogn. Comput.* **2021**, *5*, 1. [CrossRef]
14. Qiu, Y.; Dong, T.; Lin, D.; Zhao, B.; Cao, W.; Jiang, F. Fault Diagnosis for Lithium-Ion Battery Energy Storage Systems Based on Local Outlier Factor. *J. Energy Storage* **2022**, *55*, 105470. [CrossRef]
15. Zeng, A.; Yan, L.; Huang, Y.; Ren, E.; Liu, T.; Zhang, H. Intelligent Detection of Small Faults Using a Support Vector Machine. *Energies* **2021**, *14*, 6242. [CrossRef]
16. Hosseinzadeh, M.; Rahmani, A.M.; Vo, B.; Bidaki, M.; Masdari, M.; Zangakani, M. Improving Security Using SVM-Based Anomaly Detection: Issues and Challenges. *Soft Comput.* **2021**, *25*, 3195–3223. [CrossRef]
17. Lesouple, J.; Baudoin, C.; Spigai, M.; Tourneret, J.-Y. Generalized Isolation Forest for Anomaly Detection. *Pattern Recognit. Lett.* **2021**, *149*, 109–119. [CrossRef]
18. Zheng, L.; Hu, W.; Min, Y. Raw Wind Data Preprocessing: A Data-Mining Approach. *IEEE Trans. Sustain. Energy* **2015**, *6*, 11–19. [CrossRef]
19. Hu, C.; Albertani, R. Wind Turbine Event Detection by Support Vector Machine. *Wind. Energy* **2021**, *24*, 672–685. [CrossRef]
20. Turkoz, M.; Kim, S.; Son, Y.; Jeong, M.K.; Elsayed, E.A. Generalized Support Vector Data Description for Anomaly Detection. *Pattern Recognit.* **2020**, *100*, 107119. [CrossRef]
21. Chen, B.; Yu, S.; Yu, Y.; Zhou, Y. Acoustical Damage Detection of Wind Turbine Blade Using the Improved Incremental Support Vector Data Description. *Renew. Energy* **2020**, *156*, 548–557. [CrossRef]
22. Benmahamed, Y.; Kherif, O.; Teguar, M.; Boubakeur, A.; Ghoneim, S.S.M. Accuracy Improvement of Transformer Faults Diagnostic Based on DGA Data Using SVM-BA Classifier. *Energies* **2021**, *14*, 2970. [CrossRef]

23. Jeong, K.; Choi, S.B.; Choi, H. Sensor Fault Detection and Isolation Using a Support Vector Machine for Vehicle Suspension Systems. *IEEE Trans. Veh. Technol.* **2020**, *69*, 3852–3863. [CrossRef]
24. Yu, W.; Yu, R.; Li, C. An Information Granulated Based SVM Approach for Anomaly Detection of Main Transformers in Nuclear Power Plants. *Sci. Technol. Nucl. Install.* **2022**, *2022*, e3931374. [CrossRef]
25. Wang, D.; Tan, D.; Liu, L. Particle Swarm Optimization Algorithm: An Overview. *Soft Comput.* **2018**, *22*, 387–408. [CrossRef]
26. Zeng, B.; Guo, J.; Zhu, W.; Xiao, Z.; Yuan, F.; Huang, S. A Transformer Fault Diagnosis Model Based On Hybrid Grey Wolf Optimizer and LS-SVM. *Energies* **2019**, *12*, 4170. [CrossRef]
27. Ahmed, Q.I.; Attar, H.; Amer, A.; Deif, M.A.; Solyman, A.A.A. Development of a Hybrid Support Vector Machine with Grey Wolf Optimization Algorithm for Detection of the Solar Power Plants Anomalies. *Systems* **2023**, *11*, 237. [CrossRef]
28. Nong, Y.; Chen, Z.; Huang, C.; Zhou, Z.; Pan, J.; Liang, D.; Wei, Y.; Li, Z.; Lu, Y. Support Vector Machine Classification Based on Improved Harris Hawk Optimization Algorithm. *J. Phys. Conf. Ser.* **2022**, *2219*, 012050. [CrossRef]
29. Heidari, A.A.; Mirjalili, S.; Faris, H.; Aljarah, I.; Mafarja, M.; Chen, H. Harris Hawks Optimization: Algorithm and Applications. *Future Gener. Comput. Syst.* **2019**, *97*, 849–872. [CrossRef]
30. Tripathy, B.K.; Reddy Maddikunta, P.K.; Pham, Q.-V.; Gadekallu, T.R.; Dev, K.; Pandya, S.; ElHalawany, B.M. Harris Hawk Optimization: A Survey onVariants and Applications. *Comput. Intell. Neurosci.* **2022**, *2022*, 1–20. [CrossRef]