Research Article – Vol. 03, Issue 01, No. 08

Advanced in Engineering and Intelligence Systems

Journal Web Page: https://aeis.bilijipub.com

Enhancing Solar Energy Prospects: Predicting Direct Normal Irradiance in Qinghai Province Using ALO-RF Modeling

MD ABDUL MUNNAF[1], Tawhidul Islam[2, *]

[1] *Collage of Economics and Management, China Three Gorges University, Hubei China*

[2] *College of Mechanical Design, Manufacturing, and Automation, China Three Gorges University, Hubei, China*

Highlights

- Significance of Direct Normal Irradiance in atmospheric processes, climate, and renewable energy generation.
- Utilization of Genetic algorithm, Moth Flame Optimization, and Ant Lion Optimizer to enhance Random Forest model.
- Importance of precise Direct Normal Irradiance prediction for maximizing solar power plant efficiency.
- Novel hybrid approach combining Ant Lion Optimizer with Random Forest yields superior performance outcomes.

Article Info

Received: 19 February 2024
Received in revised: 16 March 2024
Accepted: 28 March 2024
Available online: 30 March 2024

Keywords

Solar Power Grid Integration
Direct Normal Irradiance
Forecasting
Qinghai Province
Random Forest.

Abstract

One important aspect of solar radiation that has a direct impact on atmospheric processes, climatic conditions, and energy generation is direct normal irradiance. The integration of solar geometry, geographic location, and atmospheric characteristics is required for the prediction of Direct Normal Irradiance. Predicting Direct Normal Irradiance is essential for maximizing the efficiency of solar power plants in the field of renewable energy. Precise predictions facilitate the efficient assimilation of solar electricity into the electrical grid, augmenting energy production and maintaining system stability. This study uses the Genetic algorithm, Moth Flame Optimization, and Ant Lion Optimizer to optimize the Random Forest model. The basic approach for this study is provided as a novel hybrid method by combining Ant Lion Optimizer with Random Forest, which has the best performance outcome compared to other created models. The data used is from June 1, 2022, to July 30, 2023. In presenting this study, many aspects have been considered, including the coefficient of determination, root mean square error, mean absolute percentage error, and mean absolute error. The proposed model's findings with the highest amount of R-squared have shown satisfactory performance.

1. Introduction

The issues related to the depletion of fossil energy, environmental decline, and global warming are escalating due to the continuous growth in energy consumption driven by population increase, rapid industrialization, and economic development [1]. Consequently, there is a heightened emphasis on optimizing the utilization of abundant and sustainable renewable energy sources, with a specific focus on solar and wind energy [2]. In many countries, renewable energy serves as a viable alternative to fossil fuels, playing a crucial role in diminishing greenhouse gas emissions. Photovoltaic (PV) power generation, in particular, has gained widespread popularity due to its numerous benefits such as sustainability, minimal site installation demands, user-friendliness, and safety [3]. The primary challenge in producing solar energy is the intermittent power generation of photovoltaic systems, which is mostly caused by weather. Fundamentally, the quality of electric power output may be significantly impacted by variations in temperature and irradiance [4]. Predicting solar irradiance may be a useful tool for estimating power production since it is closely linked to solar power harvesting [5]. The photovoltaic system's power imbalance may result in a considerable loss of

* Corresponding Author: Tawhidul Islam
Email: tawhidurrahman658@gmail.com

economic profit for large-scale solar farms. In order to minimize the effects of uncertainty and energy prices and to facilitate the appropriate integration of photovoltaic systems in a smart grid, it is becoming more important to anticipate solar irradiance accurately. Numerous research has been conducted on models and algorithms to forecast sun irradiance based on regularly collected meteorological variables like humidity and temperature [6],[7].

Ensuring precise forecasts of future solar radiation, a key factor affecting PV power generation, is essential for preventing unnecessary energy losses. Solar radiation prediction models are typically categorized into two groups: physical models and data-driven models. By utilizing physical equations that consider the interactions among solar geometric elements such as azimuth and solar altitude angle and meteorological factors including cloud cover and temperature, the physical model anticipates solar irradiance for the next day [8],[9]. However, the complex definitions of meteorological variables like temperature, cloud cover, and solar irradiance pose challenges for this model. The efficacy of the physical model is additionally hindered by the inherent nonlinearity of relational expressions and the necessity for frequent adjustments to boundary conditions and correlation coefficients in response to changing external conditions [10]. Without knowledge of solar geometry, data-driven models may be constructed via learning processes and anticipate solar irradiance based on the statistical or probabilistic associations observed in measured data [11],[12]. In some studies, Artificial Neural Network (ANN) approaches were used to build early versions of data-driven models [13][14][15]. According to [16], data-driven solar forecasting models now have much higher accuracy thanks to ANNs' learning abilities, which are especially useful for handling nonlinear data. Their limited structure, however, makes it difficult for them to manage massive amounts of data and interpret a variety of meteorological factors. When compared to ANNs, ensemble techniques like Random Forests and Gradient Boosting are often more resilient to noisy data and outliers. The effect of individual faults is lessened when many models are combined. Multiple weak regressors, also known as base learners, are used in ensemble learning models. These learners are taught on the training set and then integrated by a meta-learner on the validation set. Recently, there has been a lot of interest in ensemble learning-based solar irradiance predictions. [17] applied ensemble pruning across three distinct ensemble learning methodologies, namely bagging, random subspace, and boosting. [18] introduced the Integrated Bayesian Multi-Model Uncertainty Estimation Framework (IBMUEF) to concurrently assess the uncertainty associated with both the model structure and input parameters. For the prediction of half-hourly Global Horizontal Irradiance (GHI), [19] integrated the foundational persistent model with four novel models: static, dynamic, moving average, and amplified persistent. The aggregation parameters of these models were enhanced using a particle swarm optimizer. Additionally, [20] employed multi-task representation learning within various customized groups to diversify the array of base learners, specifically Gated Recurrent Unit (GRU). One popular class of machine learning algorithms is random forests. Their foundation is an ensemble learning technique that generates predictions by combining many decision trees. A portion of the data and accessible characteristics are used to train each tree in the random forest. This lessens the possibility of overfitting and enhances the model's ability to generalize. Leo Breiman initially presented the random forest method in 2001[21]. It is a well-liked and effective technique that has been used to many different fields, such as medical diagnosis, credit scoring, and picture and voice recognition. The random forest technique consists of building several decision trees, each of which is trained using a distinct subset of the data. The projections of every tree in the forest are then combined to create the final forecast. This enhances the model's accuracy and lowers variance. The particular model used in this study is referred to as RF, and it is a variation of the standard random forest technique. The RF model is a well-liked option for several applications because to its reputation for handling noisy features and high-dimensional data [21]. Tweaking hyperparameters may have a substantial influence on the performance of a model. Identifying the most favorable values may result in enhanced accuracy, precision, recall, or other pertinent metrics, contingent upon the particular objective of the model. Overfitting is the phenomenon when a model has good performance on the training data but exhibits poor performance on fresh, unseen data. Underfitting occurs when a model is too simple and fails to capture the underlying patterns. Optimizing hyperparameters aids in achieving an optimal equilibrium and mitigating the risks of overfitting or underfitting [22]. The optimization methods used in this work are Genetic algorithm (GA) [23], Moth flame optimization (MFO) [24] and Ant lion optimization (ALO) [25]. The GA optimization is a search heuristic that draws inspiration from the ideas of natural selection and genetics. This approach, often used for hyperparameter optimization in machine learning models and other applications, begins by generating an initial population of candidate solutions. Each solution corresponds to a distinct combination of hyperparameters for the machine learning model, where

individuals within the population are typically referred to as solutions [22]. The Moth Flame Optimizer represents an advanced tool known for significantly enhancing the performance of various models. The inspiration behind its development stems from the behavior of nocturnal butterflies, which exhibit a tendency to be attracted to a light source throughout the night. These insects have a natural inclination to navigate by flying toward the moon, a strategy that has proven effective for long-distance travel. However, there is a susceptibility to getting trapped when they repetitively orbit around the source of light. The specific movement pattern has been thoroughly investigated and can be employed as a highly effective optimizer across diverse domains, such as electrical and energy systems, business administration, architectural design, image processing, and medicinal applications [24]. Another model used in this method is ALO, the ALO algorithm originated from observing the hunting behavior of ant-lion larvae as they pursue ants. Initial investigations might explore additional advanced features [26]. The system, which revolves around the interactions between ant lions and their prey which are ants, aims to broaden the exploration scope. Through traps, ant lions can seize and nourish themselves with ants, thereby enhancing their overall fitness [27]. The data has been obtained from several meteorological stations equipped with modern sun radiance measuring tools. The dataset covers a defined period of time and geographical area in order to include a wide variety of solar conditions.

- The specified duration for collecting data spans a year, extending from mid-2022 to mid-2023, within the province of Qinghai. Situated in western China, northeast of the Tibetan Plateau, Qinghai's capital is Xining. Renowned for its varied topography, the province encompasses mountains, plateaus, lakes, and China's largest lake, Lake Qinghai. The data comprises six components: temperature (temp), relative humidity (RH), cloud cover (CC), wind gusts (WG), diffuse radiation (DR), and direct normal irradiance (DNI). To evaluate the magnitude and direction of a linear association between two continuous variables, the Pearson correlation is employed.
- This statistical measure yields a numerical value between -1 and 1, with 0 indicating no linear correlation, 1 representing a perfect positive linear relationship, and -1 indicating a perfect negative linear relationship. In this study, components with weak correlations to DNI are excluded based on the Pearson correlation analysis.
- Additionally, factors that showed a perfect correlation namely, 1 with DNI were also removed. The research also used a different approach called min-max normalization, often known as feature scaling or min-max scaling. It is a popular technique for preparing data for machine learning. Rescaling and normalizing the range of numerical attributes in a given dataset is the aim of this normalization procedure.

The second part of the study describes the methods and materials. The third part provides information about results and discussion. The conclusion is presented in the fourth part.

2. Material and methods

2.1.Study area

Qinghai province, which is situated in the northeastern part of the Tibetan Plateau, has great potential for solar energy harvesting and beautiful landscape. The enormous 720,000 square kilometer province of Qinghai, China, is situated at 35.7452° N latitude and 95.9956° E longitude. A key metric in solar energy research is the DNI, which measures the amount of solar radiation received per unit area by a surface that is perpendicular to the sun's beams. Precise DNI approximations are crucial for enhancing the layout and functionality of photovoltaic systems, offering important insights on the available solar power in a certain region. Qinghai's distinct topography, which includes its high altitudes and diverse terrain, greatly influences the amount of direct sunshine that reaches the surface. For precise DNI predictions, a thorough analysis of these topographical features is also necessary. Fig. 1 shows the topography of Qinghai graphically. Seasonal variations in the duration and angle of sunlight are strongly related to DNI oscillations. Understanding the availability of solar energy in detail is made easier with the use of seasonal forecasts.

Fig. 1. Map of the study area.

2.2. Data source

This dataset monitors variations for a year, from 2022 to 2023. This dataset is intended to be used to train a prediction model that can forecast DNI levels in response to diverse environmental and meteorological variables. The dataset offers a comprehensive view of the factors impacting DNI since it includes a large number of variables that are routinely gathered. Comprehending the elements listed in Table 1 facilitates comprehension of the current meteorological conditions. The dataset was enhanced by Pearson correlation analysis, which included computing the linear correlations between the variables. The outcome of this analysis is shown in Fig. 2, showcasing the results of the Pearson correlation. Following the calculation of correlation coefficients, variables exhibiting correlation coefficients ranging from -0.1 to 0.1, as well as those displaying a perfect correlation of 1, were excluded from the study. In order to guarantee that the items retained in the DNI prediction model contribute significantly, a rigorous selection method is used to exclude duplicate variables or those with weak correlations. The main objective variable is the measured DNI. The dataset is preprocessed previous to model training to achieve optimum model performance.

Table 1. Input variables, features obtained from the Pearson correlation process.

Features	Temp	RH	CC	WG	DR	DNI
Definitions	Temperature	Relative humidity	Cloud cover	Wind gusts	Diffuse radiation	Direct normal irradiance
Units	$(°C)$	$(\%)$	$(\%)$	(km/h)	(W/m^2)	(W/m^2)

2.2.1. Pearson correlation

A numerical measure known as Pearson correlation, often called Pearson's correlation coefficient or simply Pearson's r, evaluates both the direction and intensity of a linear association between two variables. It quantifies the degree and direction of a linear correlation between two continuous variables through a specific numerical value. The formula for Pearson correlation in the equation (1) is as follows:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (1)$$

In this context, X_i and Y_i represent individual data points, while $\bar{X}$ and $\bar{Y}$ denote the means of the X and Y variables, respectively. The symbol $\sum$ signifies the summation across all data points. The study examined the

relationship between DNI and various factors as presented in Fig. 2. The Pearson correlation test was employed for this analysis, and the results are depicted in Fig. 2. Factors exhibiting a strong correlation were retained as input variables, while those with weak correlations (coefficients between -0.1 and 0.1) were excluded. Dew, Rain, Snow, Press, and WS were omitted from the input variables due to their limited correlation with DNI.

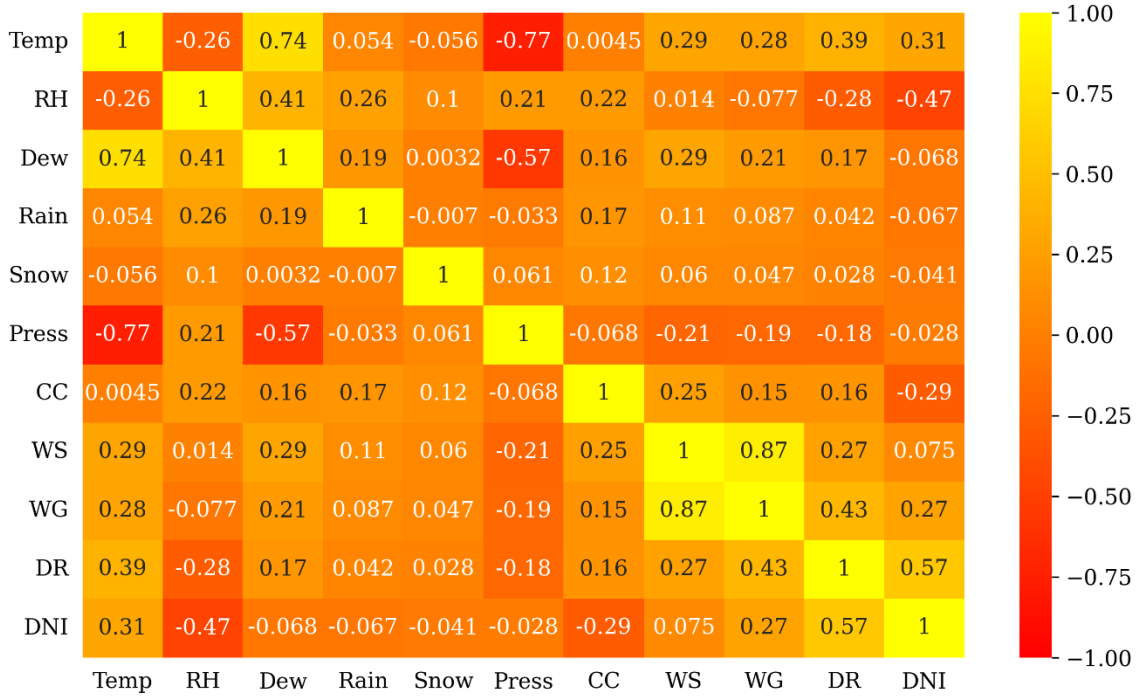

Fig. 2. Heat map illustration showing variable correlation.

Fig. 3 serves as a valuable tool for evaluating the relationship between variables. When data points on a scatterplot tightly group around a line, whether with a positive or negative slope, it signifies a correlation between the two variables.

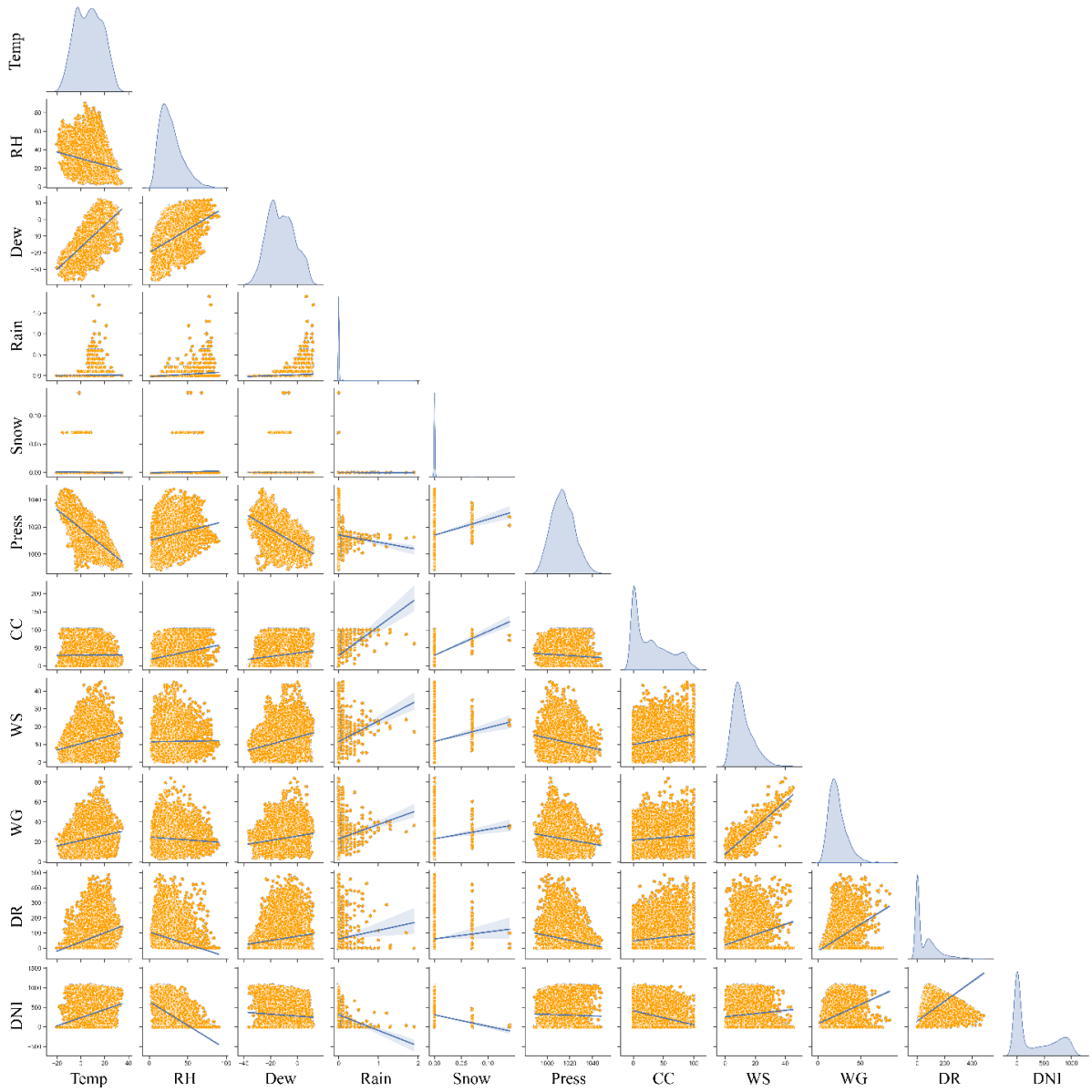

Fig. 3. Employed a Pairplot to visually represent the relationships and distribution of the data.

2.2.2. *Splitting data*

The machine learning model is trained using the training set. In order to reduce mistakes or discrepancies between its predictions and the actual results, the model modifies its parameters as it discovers patterns and linkages in the data. The testing set is put aside to assess the trained model's performance. It makes it possible to evaluate how well the model applies to fresh, untested data. The efficacy of the model is evaluated based on its capacity to provide precise forecasts on non-training data.

2.2.3. *Data normalization*

A method known as MinMax normalization is employed for data preparation to scale and adjust numerical characteristics in a dataset. It goes by various names, including feature scaling or min-max scaling. This approach involves transforming feature values to fall within a specified range, typically between 0 and 1. The utilization of MinMax normalization is aimed at preventing features with larger magnitudes from dominating the learning process, thereby ensuring equal contribution from each feature in machine learning models during analysis. The normalization formula MinMax is represented by the equation (2):

$$X_{normalized} \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

2.2.4. *The dataset's statistical outcome*

The statistical results obtained from the gathered data are shown in Table 2. Descriptive statistics are used to

delineate different characteristics of the dataset, providing a succinct overview of its aspects. The statistics provided include the Count, which represents the number of observations in the dataset, as well as the Average and Median, which serve as measures of central tendency. Skewness is a measure that reveals the degree of symmetry in the distribution of data. Standard deviation is a statistical measure that quantifies the extent to which data points deviate from the mean, indicating the dispersion of the data. Kurtosis measures the degree of peakedness or flatness of the data compared to a normal distribution. Variance quantifies the amount of data fluctuation from its mean. Maximum and minimum values represent the highest and lowest values in the dataset, respectively. By analyzing these descriptive statistics, researchers may get a more comprehensive understanding of the dataset's attributes, allowing them to make educated judgments based on the insights uncovered.

Table 2. The statistical outcome of the abstained features.

	Temp	RH	CC	WG	DR	DNI
Count	8760	8760	8760	8760	8760	8760
Mean	6.83	28.03	29.86	23.09	61.25	306.89
Std.	10.98	14.53	29.58	10.36	84.57	372.71
Min	-20.2	3	0	2.9	0	0
50%	6.85	25	22	21.2	8	27.25
Max	34.4	90	100	83.9	486	1076.6
Variance	120.6	211.13	874.72	107.34	7151.8	138913

2.3. Model description

2.3.1. Random forest

Although decision trees are often employed as simple regression and classification models, when they are used to solve complicated problems with numerous input factors, they have a tendency to over fit. The stochastic subspace approach [28] and bagging ensemble learning theory [29] were merged to create RF as demonstrated in Fig. 4, which was suggested in 2001 [21] to prevent a single decision tree from becoming unstable and prone to overfitting. RF is a technique for integrating several decision trees, where the output of each decision tree is combined to get the final conclusion. The bootstrap sampling approach is used to obtain the training samples for each base learner in the RF. Stated differently, a random subset is selected from all characteristics; the remaining samples are referred to as out-of-bag samples (OOB). Its unpredictability is evident in two ways: the feature vectors of the tree are likewise randomly produced, as are the training samples of the tree, which are drawn at random with replacement. This avoids the overfitting issue and enhances the variation between individual decision trees due to the unpredictability of training extraction. When the forest is finally built, the outcomes may be averaged using equation (3). Consequently, the final fusion model is more accurate.

$$\bar{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T(x.O_b) \quad (3)$$

Where $\bar{f}_{rf}^{B}$ is the average, B represent the trees, and the output of each tree can express as $T(x, O_b)$.

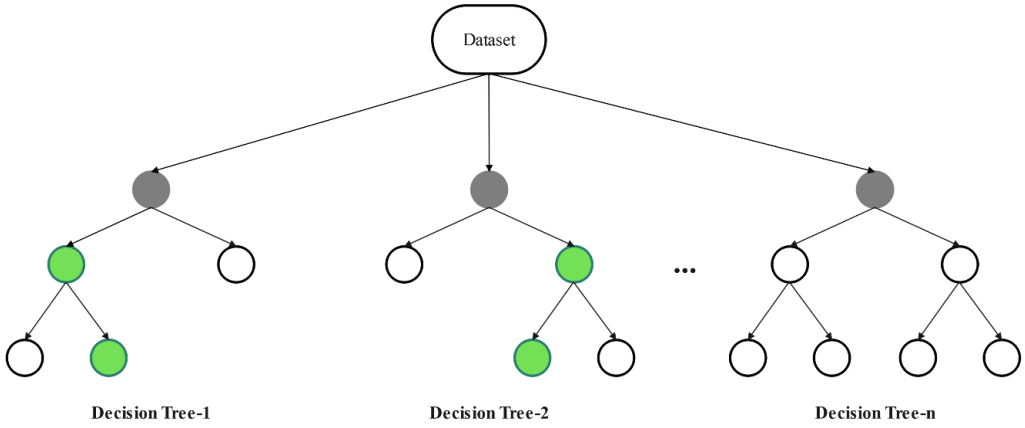

Fig. 4. Empowering Predictive Insights: Harnessing the Power of RF for Data-driven Forecasting.

2.3.2. *Genetic algorithm*

In 1971, John Holland devised the GA, a probabilistic search method aimed at addressing intricate optimization problems characterized by high complexity and unfavorable structures [23]. This approach, inspired by natural selection and genetics, emulates the mechanics of evolution to approximate solutions for optimization and search challenges. The heuristic search strategy, grounded in genetic evolution and natural selection, involves iteratively generating a population comprising individuals or chromosomes potentially serving as solutions to the optimization problem. Each participant's fitness is assessed using an objective function gauging their efficacy in problem-solving. Selection for reproduction is contingent on fitness, favoring individuals with superior fitness for procreation. The generation of offspring involves the exchange of genetic information among selectively chosen individuals, simulating biological crossover or recombination. Introducing random alterations enhances individual genetic variation. The subsequent generation is formed by crossing the newly generated offspring with some of the existing individuals, perpetuating the evolution of the population by favoring more adaptable individuals. The algorithm continues through these steps iteratively until a specified termination criterion is met or after a predefined number of generations. Genetic Algorithm Optimization excels in tackling complex, non-linear, multidimensional optimization problems that may pose challenges for conventional methods. Its efficacy lies in the exploration of a broad spectrum of potential solutions and the identification of those approaching optimality. This versatility makes it applicable across various domains, including scheduling, machine learning, engineering, and finance [30].

2.3.3. *Moth flame optimization*

The main objective of MFO is to comprehend the navigation system of moths in transverse orientation. Moths cover extensive distances during nighttime flights by maintaining a consistent angle with the sky [24]. This study focuses on the spatial configurations of moths, treated as variables, which might be remedies. Moths have the capability to adjust their position vectors for navigation in 1D, 2D, 3D, or hyper-dimensional space. MFO demonstrates computational efficiency and robustness, with the proposed method ensuring convergence. The representation of MFO is commonly articulated by using equations (4) and (5):

$$M = \begin{bmatrix} CO_{1.1} & CO_{1.2} & \cdots & \cdots & CO_{1.h} \\ CO_{2.1} & CO_{2.2} & \cdots & \cdots & CO_{2.h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CO_{a.1} & CO_{a.2} & \cdots & \cdots & CO_{n.h} \end{bmatrix} \quad (4)$$

a and h represent the number of moths and dimensions, respectively.

$$S = \begin{bmatrix} S_{1.1} & S_{1.2} & \cdots & \cdots & S_{1.h} \\ S_{2.1} & S_{2.2} & \cdots & \cdots & S_{2.h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{a.1} & S_{a.2} & \cdots & \cdots & S_{2.h} \end{bmatrix} \quad (5)$$

MFO is a global optimization process that consists of three steps.

$$MFO = (I.F.T) \quad (6)$$

Where T is the halting criterion, F is the moth's space travel, and I is a function.

$$X_i = t(C_i.S_j) \quad (7)$$

C_i represents the count of moths in the i_{th} category, S_j denotes the quantity of flames in the j_{th} category, and t is the twisting function, which can be defined by using equation (8) and (9):

$$S(C_i.S_j) = Z_i \cdot e^{bt} \cdot \cos(2\pi t) + S_j \quad (8)$$

In this context, Z_i represents the distance between a moth and a flame, b is a constant value, and t is a random number selected from the range [-1,1].

$$\mathrm{Zi} = \left|\mathrm{S_j} - \mathrm{X_i}\right| \quad (9)$$

2.3.4. *Ant lion optimizer*

A technique known as the ant lion optimizer algorithm was developed and presented in [25] to tackle a range of real-world engineering challenges. It emulates the foraging behavior of ants in their natural habitat. The literature discusses several problems that were successfully addressed using ALO [31][27][32]. In this algorithm, an antlion, equipped with large jaws, forms cone-shaped depressions in the sand, as illustrated in Fig. 5. These pits serve as traps for ants. Once the pit is created, the larva conceals itself beneath the base of the cone and awaits the arrival of ants into the depressions, as depicted in Fig. 6. Within the ALO, two distinct categories of search agents exist: ants and antlions. Notably, the superior search agents are designated as antlions, and they persist in their locations even after eliminating a particular ant.

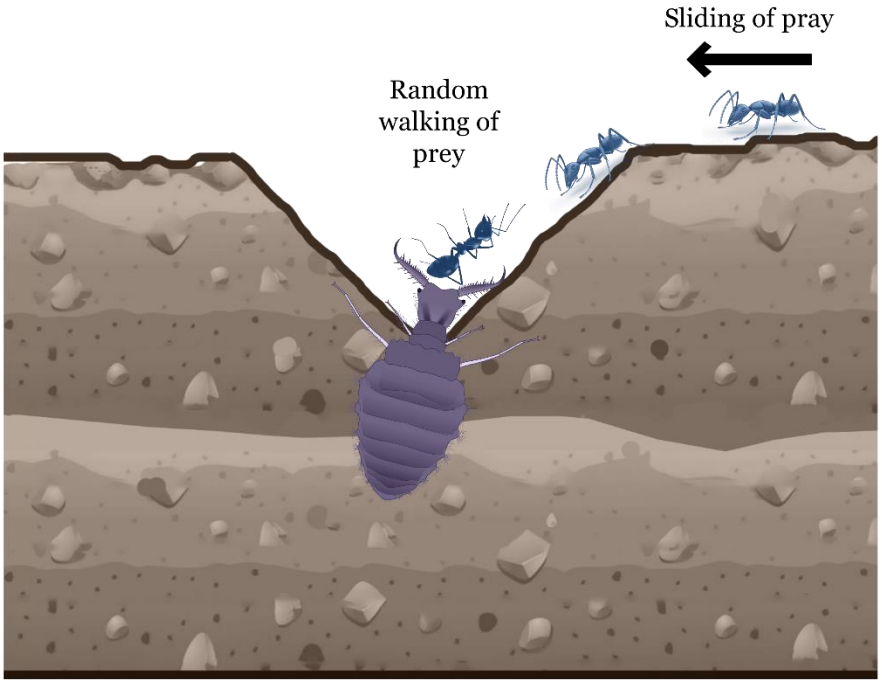

Fig. 5. Description of the hunting process of the antlion.

The identification of an ant's position can be achieved by using equation (10):

$$Ant_i^t = \frac{r_a^{it} + r_e^{it}}{2} \tag{10}$$

Here, r_a^{it} refers to the random walk around the chosen ant lion, and r_e^{it} denotes the random walk around the elite at the it^{th} iteration. The movement pattern of an Ant_i^{it} as it randomly traverses the vicinity of a presumed Antlion i_i^{it} can be expressed by using equation (11):

$$X_i^{it} = \frac{\left(X_i^{it} - A_i\right) \times \left(D_i - C_i^{it}\right)}{\left(D_i^{it} - A_i\right)} + C_i \tag{11}$$

Where A_i represents the minimum value from the random walk of the i^{th} variable, D_i represents the maximum value of the i^{th} variable. C_i^{it} is the minimum value of the i^{th} variable at the it^{th} iteration, and D_i^{it} is the maximum value of the i^{th} variable at the it^{th} iteration. The ants exhibit a stochastic movement pattern in their natural environment as they forage for food. This behavior may be replicated via simulation using equation (12):

$$X(it) = [0.\ \text{CuSu}\ (2r(it_1) - 1).\ \text{CuSu}\ (2r(it_2) - 1). \ldots .\ \text{CuSu}\ (2r(it_n) - 1)] \tag{12}$$

where: CuSu denotes the cumulative sum of the pending time $r(it)$, which is defined using equation (13):

$$r(it) = \begin{cases} 1.\ \text{if rand} > 0{,}5 \\ 0.\ \text{if rand} \leqslant 0{,}5 \end{cases} \tag{13}$$

The antlions notice a prey's intrusion in their pit. They pour sand on them and roll them down the pit. The model of this phase may be stated using equations (14) and (15):

$$c^{it} = \frac{c^{it}}{10^{\omega} \times \frac{it}{it_{\max}}} \tag{14}$$

$$d^{it} = \frac{d^{it}}{10^{\omega} \times \frac{it}{it_{\max}}} \tag{15}$$

Here, c^{it} and d^{it} represent the lower and upper bounds of the variables undergoing optimization, and ω is a constant with a fixed value determined based on the current iteration, using equation (16):

$$\omega = \begin{cases} 2.\ \text{if it} > 10\% \cdot it_{\max} \\ 3.\ \text{if it} > 50\% \cdot it_{\max} \\ 4.\ \text{if it} > 75\% \cdot it_{\max} \\ 5.\ \text{if it} > 90\% \cdot it_{\max} \\ 6.\ \text{if it} > 95\% \cdot it_{\max} \end{cases} \tag{16}$$

During the ALO optimization process, the concluding phase involves capturing prey through predatory tactics and subsequently reconstructing the trap.

This step can be obtained by using equation (17):

$$\text{Antlion}_j^{it} = \text{Ant}_i^{it};\ \text{if}\ f(\text{Ant}_i^{it}) > f(\text{Antlion}_j^{it}) \tag{17}$$

$Antlion_j^{it}$ refers to the position of the selected j^{th} antlion at the i^{th} iteration, and Ant_i^{it} represents the location of the i^{th} ant at the same iteration. Table 3 demonstrate the

setting of the hyper parameters of the RF and finding the optimal values for each parameter by ALO.

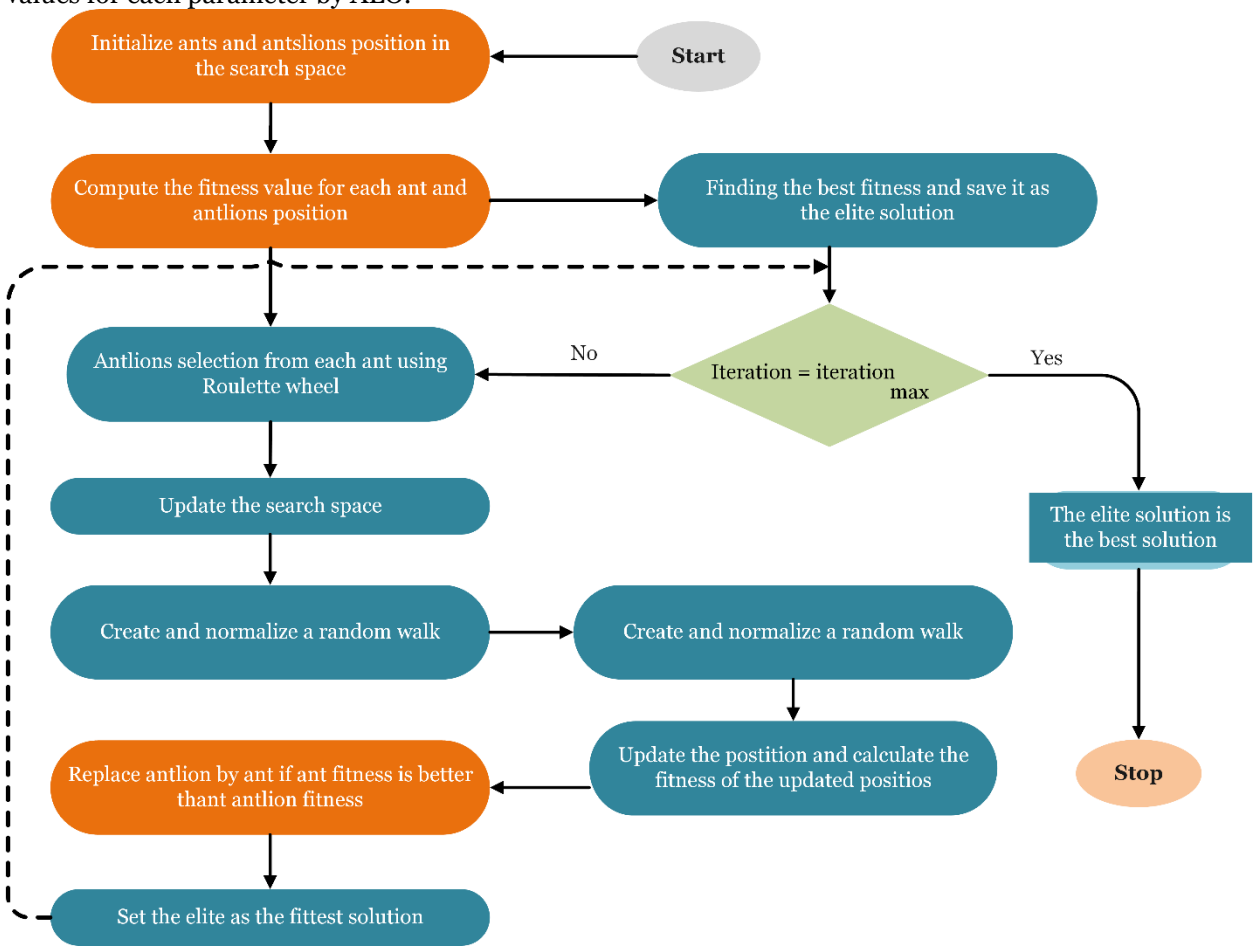

Fig. 6. The antlion optimizer flowchart.

Table 3. Setting of the hyperparameters and finding the optimal values by ALO.

Random Forest		**Best value**
Max depth	[10, 100, None]	80
Max features	[auto and sqrt]	auto
Min samples leaf	[1, 4]	2
Min samples split	[2, 10]	2
Number of estimators	[200, 2000]	500

2.4. Assessment criteria

Evaluation metrics are crucial in machine learning projects since they provide a quantifiable assessment of a model's accuracy. These metrics are used to evaluate the model's ability to predict outcomes on new and unseen data. The model's performance is assessed using four assessment metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error ($MAPE$), Mean Square Error (MSE), and R-squared (R^2) using equations (18)-(21).

$$MSE = \frac{1}{N}\sum_{k=0}^{n}\binom{n}{k}(Fi - Yi)b^2 \quad (18)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (19)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (20)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (21)$$

3. Result and Discussion

3.1.Comparative analysis

The effectiveness of the provided models was assessed using a range of common metrics, such as MAE, $MAPE$, R^2, and MSE. These metrics provide a comprehensive

assessment of the forecast precision of the models. Table 4 provides a summary of the performance metrics for four models: RF, GA-RF, MFO-RF, and ALO-RF. These models were developed and evaluated using range of obtained and prepared features for a Qinghai province, spanning from 2022 to 2023. Table 4 presented result of all models via acquired evaluation metrics.

Table 4. The presented models outcome via assessment criteria during train and testing phase.

Models/Metrics		RF	GA-RF	MFO-RF	ALO-RF
R^2	Train Set	0.971	0.984	0.988	0.994
MAPE		32.75	22.18	18.70	12.60
MAE		48.01	25.97	20.88	13.67
MSE		4159.26	2249.42	1696.00	865.24
R^2	Test Set	0.966	0.978	0.978	0.989
MAPE		39.30	28.84	31.72	28.09
MAE		46.68	36.40	36.67	31.74
MSE		3990.80	2585.19	2572.83	2183.98

Table 4 demonstrates that the ALO-RF model outperforms the other models in terms of predictive accuracy. The model's capacity to precisely depict the intricate temporal patterns and correlations found in stock price data is shown by its very low values for MAE, MAPE, and MSE. The results suggest that the ALO-RF model may be a reliable tool for detecting possible market trends and making informed investment decisions.

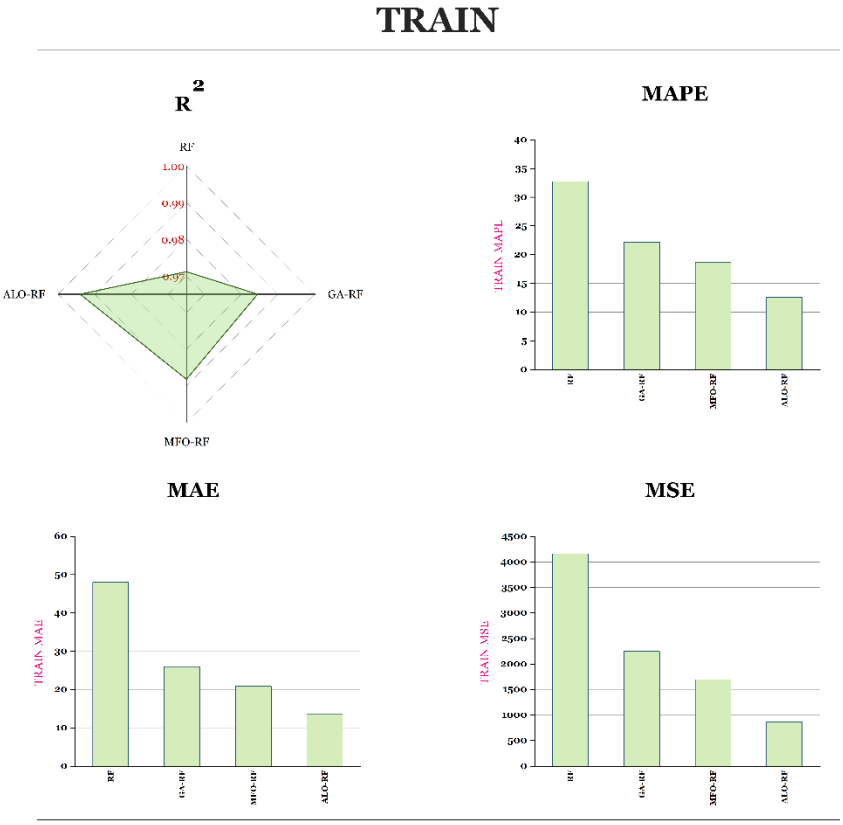

Fig. 7. The illustration of the comparison between present models during train.

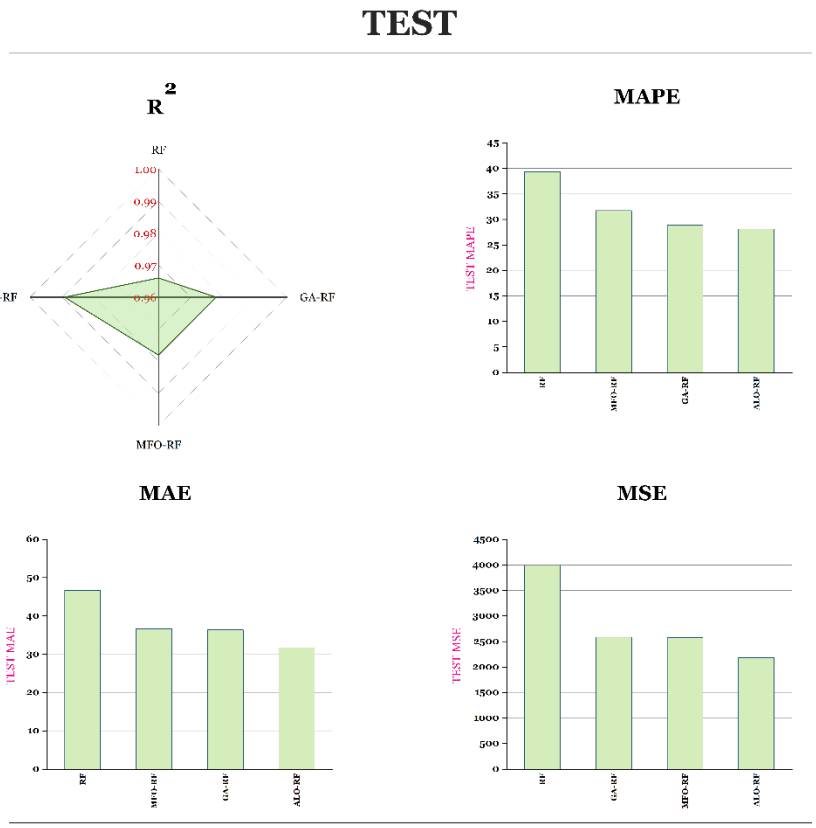

Fig. 8. The illustration of the comparison between present models during test.

Upon comparing the performance of the four models listed in Table 3, it is evident that the ALO approach outperformed the MFO and GA techniques in optimizing the hyperparameters of the provided model. The RF, GA-RF, MFO-RF, and ALO-RF results, with corresponding values of 0.966, 0.978, 0.978 and 0.989, provide evidence of the enhanced performance of the model. The assessment results of the created models are shown in Fig. 7 and Fig. 8. It is obvious from the figures that ALO-RF outperforms all other models in terms of all evaluation criteria. The results indicate that the optimized model has improved the accuracy of the forecast. The ALO-RF model, optimized using the ALO approach, has a R^2 evaluation criteria score of 0.989. This result illustrates that optimization has a favorable influence on prediction in comparison to the non-optimized RF model. The RF achieved a performance metric of 0.95 without using the optimization strategy. The created models are compared in Fig. 9 and Fig. 10.

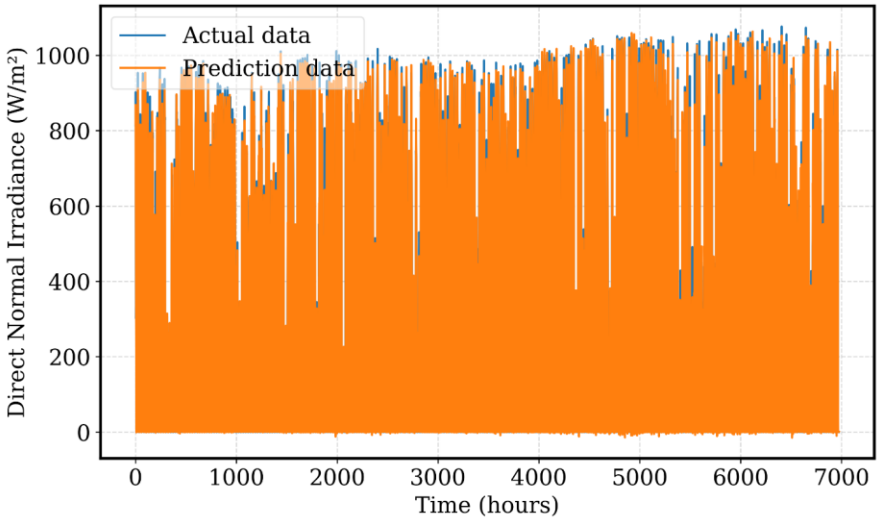

Fig. 9. Visual comparison between real data and data predicted by the ALO-RF.

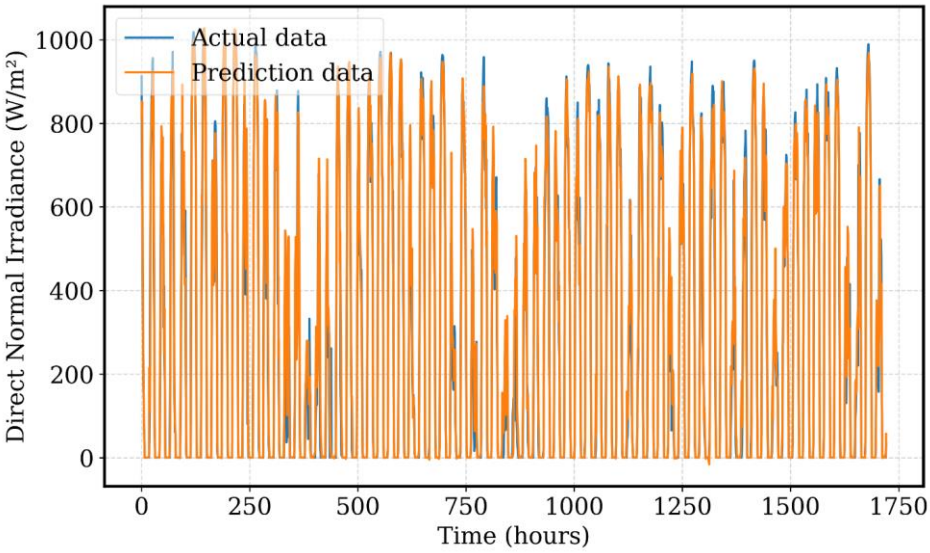

Fig. 10. Visual comparison between real data and data predicted by the ALO-RF during test.

Realistically, a comprehensive solar energy forecasting Method might be implemented in response to the study's identified limitations and suggestions for future research. In order to address the demand for streamlined solar power generation and grid integration, this Method would utilize sophisticated machine learning algorithms and integrate real-time data to deliver precise forecasts of DNI across diverse geographical areas. By extending the geographical coverage beyond Qinghai Province and integrating more extended time periods, the Method would be capable of accommodating a wide range of locations and capturing fluctuations in seasons, thereby guaranteeing the provision of dependable and complex forecasts. By integrating supplementary meteorological variables, a comprehensive comprehension of solar radiation patterns could be achieved, whereas the application of ensemble techniques could augment the accuracy of predictions. Constant refinement of forecasts and adaptation to changing environmental conditions would be made possible through the incorporation of real-time data and model updating mechanisms. In essence, the implementation of this solar energy forecasting Method would grant grid operators and renewable energy developers the ability to maximize the production of solar power, diminish dependence on non-renewable energy resources, and make a positive contribution towards a sustainable energy landscape.

4. Conclusion

The expanding global population is causing an increase in the use of non-renewable natural resources. More effective ways to handle the energy dilemma must be developed since non-renewable energy sources are limited and will soon run out. Adopting a plan in line with sustainable development, the use of renewable resources becomes apparent as a critical step in guaranteeing a durable and resilient energy future. Given that the sun is the ultimate source of all energy, developing the capacity to forecast radiation levels in various regions ahead of time may provide a workable answer to the global energy dilemma. The conversion of radiation into energy has the potential to be far more efficient by taking advantage of these expected outcomes. This work has been presented using a variety of machine learning techniques, including RF and optimized RF with GA, MFO, and ALO. The criteria demonstrate ALO's advantage over other used techniques. The variables were gathered between 2022 and 2023; the data came from the Chinese province of Qinghai. Before being utilized as input data for the model, the acquired data underwent a number of phases of processing. The procedures utilized to prepare the data included the Pearson correlation test, data normalization, and splitting the data into training and testing sets. The overall findings of this study are as follows:

- The outcome demonstrates that hybrid models outperform models without non-optimized approaches in producing more satisfactory result. In comparison the hybrid ALO-RF in this study performed better than other models with outcomes of 0.989 respectively.

Developers in the renewable energy industry stand to gain greatly from the use of this advanced model to predict the sun's DNI in Qinghai Province. With improved solar panel efficiency, pollution in the environment might be decreased as a result of this concept.

The research is limited to Qinghai Province in terms of its geographic scope, which may restrict the applicability of

the results to other areas characterized by potentially distinct atmospheric conditions. In addition, the temporal scope is restricted to the period between June 2022 and July 2023, which may result in the omission of long-term patterns or seasonal fluctuations. Dependence on machine learning models necessitates a consideration of the training data and model assumptions, potentially compromising the accuracy of predictions due to concerns regarding data quality and availability. In dynamic environmental conditions, assumptions of stationary relationships between predictors and direct normal irradiance may not hold. Further investigation may aim to broaden the geographical range of the model's applicability in order to better suit diverse regions. By integrating longer-term datasets, one could account for seasonal and inter-annual fluctuations, thereby enhancing the precision of predictions. An examination of ensemble learning methodologies may enhance the resilience of the model, whereas the incorporation of supplementary meteorological variables may yield a more holistic comprehension of direct normal irradiance. The implementation of real-time data integration and model updating mechanisms has the potential to improve the accuracy of forecasts for shorter time periods. As a final step, the performance of the proposed model would be assessed in comparison to industry benchmarks and independent datasets through rigorous validation and benchmarking.

REFERENCES

[1] P. Guo, Y. Zhai, X. Xu, and J. Li, "Assessment of levelized cost of electricity for a 10-MW solar chimney power plant in Yinchuan China," *Energy Convers Manag*, vol. 152, pp. 176–185, 2017, doi: https://doi.org/10.1016/j.enconman.2017.09.055.

[2] L. Ju, Z. Tan, J. Yuan, Q. Tan, H. Li, and F. Dong, "A bi-level stochastic scheduling optimization model for a virtual power plant connected to a wind–photovoltaic–energy storage system considering the uncertainty and demand response," *Appl Energy*, vol. 171, pp. 184–199, 2016, doi: https://doi.org/10.1016/j.apenergy.2016.03.020.

[3] H.-Y. Cheng, C.-C. Yu, and C.-L. Lin, "Day-ahead to week-ahead solar irradiance prediction using convolutional long short-term memory networks," *Renew Energy*, vol. 179, pp. 2300–2308, 2021, doi: https://doi.org/10.1016/j.renene.2021.08.038.

[4] Y. Feng, W. Hao, H. Li, N. Cui, D. Gong, and L. Gao, "Machine learning models to quantify and map daily global solar radiation and photovoltaic power," *Renewable and Sustainable Energy Reviews*, vol. 118, p. 109393, 2020, doi: https://doi.org/10.1016/j.rser.2019.109393.

[5] Y. El Mghouchi, E. Chham, M. S. Krikiz, T. Ajzoul, and A. El Bouardi, "On the prediction of the daily global solar radiation intensity on south-facing plane surfaces inclined at varying angles," *Energy Convers Manag*, vol. 120, pp. 397–411, 2016, doi: https://doi.org/10.1016/j.enconman.2016.05.005.

[6] Y. Feng, D. Gong, Q. Zhang, S. Jiang, L. Zhao, and N. Cui, "Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation," *Energy Convers Manag*, vol. 198, p. 111780, 2019, doi: https://doi.org/10.1016/j.enconman.2019.111780.

[7] E. F. Alsina, M. Bortolini, M. Gamberi, and A. Regattieri, "Artificial neural network optimisation for monthly average daily global solar radiation prediction," *Energy Convers Manag*, vol. 120, pp. 320–329, 2016, doi: https://doi.org/10.1016/j.enconman.2016.04.101.

[8] W. De Soto, S. A. Klein, and W. A. Beckman, "Improvement and validation of a model for photovoltaic array performance," *Solar Energy*, vol. 80, no. 1, pp. 78–88, 2006, doi: https://doi.org/10.1016/j.solener.2005.06.010.

[9] A. Dolara, S. Leva, and G. Manzolini, "Comparison of different physical models for PV power output prediction," *Solar Energy*, vol. 119, pp. 83–99, 2015, doi: https://doi.org/10.1016/j.solener.2015.06.017.

[10] N. Premalatha and A. Valan Arasu, "Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms," *Journal of Applied Research and Technology*, vol. 14, no. 3, pp. 206–214, 2016, doi: https://doi.org/10.1016/j.jart.2016.05.001.

[11] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, 2018, doi: https://doi.org/10.1016/j.energy.2018.01.177.

[12] S. Srivastava and S. Lessmann, "A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data," *Solar Energy*, vol. 162, pp. 232–247, 2018, doi: https://doi.org/10.1016/j.solener.2018.01.005.

[13] M. Ding, L. Wang, and R. Bi, "An ANN-based Approach for Forecasting the Power Output of Photovoltaic System," *Procedia Environ Sci*, vol. 11, pp. 1308–1315, 2011, doi: https://doi.org/10.1016/j.proenv.2011.12.196.

[14] G. Cervone, L. Clemente-Harding, S. Alessandrini, and L. Delle Monache, "Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble," *Renew Energy*, vol. 108, pp. 274–286, 2017, doi: https://doi.org/10.1016/j.renene.2017.02.052.

[15] A. Mellit, S. Sağlam, and S. A. Kalogirou, "Artificial neural network-based model for estimating the produced power of a photovoltaic module," *Renew Energy*, vol. 60, pp. 71–78, 2013, doi: https://doi.org/10.1016/j.renene.2013.04.011.

[16] J. Qu, Z. Qian, and Y. Pei, "Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern," *Energy*, vol. 232, p. 120996, 2021, doi: https://doi.org/10.1016/j.energy.2021.120996.

[17] Z. Tan, H. Yu, R. Lu, R. Zhu, and S. Han, "Non-locally coded Fourier-transform ghost imaging," *Opt Express*, vol. 27, no. 3, pp. 2937–2948, 2019.
[18] A. Seifi, M. Ehteram, and M. Dehghani, "A robust integrated Bayesian multi-model uncertainty estimation framework (IBMUEF) for quantifying the uncertainty of hybrid meta-heuristic in global horizontal irradiation predictions," *Energy Convers Manag*, vol. 241, p. 114292, 2021, doi: https://doi.org/10.1016/j.enconman.2021.114292.
[19] M. A. Hassan, L. Al-Ghussain, A. D. Ahmad, A. M. Abubaker, and A. Khalil, "Aggregated independent forecasters of half-hourly global horizontal irradiance," *Renew Energy*, vol. 181, pp. 365–383, 2022, doi: https://doi.org/10.1016/j.renene.2021.09.060.
[20] Y. Yang, W. Hong, and S. Li, "Deep ensemble learning based probabilistic load forecasting in smart grids," *Energy*, vol. 189, p. 116324, 2019, doi: https://doi.org/10.1016/j.energy.2019.116324.
[21] L. Breiman, "Random forests," *Mach Learn*, vol. 45, pp. 5–32, 2001.
[22] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.
[23] B. Mohan and J. Badra, "A novel automated SuperLearner using a genetic algorithm-based hyperparameter optimization," *Advances in Engineering Software*, vol. 175, no. September 2022, p. 103358, 2023, doi: 10.1016/j.advengsoft.2022.103358.
[24] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowl Based Syst*, vol. 89, pp. 228–249, 2015, doi: https://doi.org/10.1016/j.knosys.2015.07.006.
[25] S. Mirjalili, "The ant lion optimizer," *Advances in engineering software*, vol. 83, pp. 80–98, 2015.
[26] S. Mirjalili, "Ant Lion Optimization," *Advances in engineering software*, vol. 83, pp. 80–98, 2015.
[27] A. A. Heidari, H. Faris, S. Mirjalili, I. Aljarah, and M. Mafarja, "Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks," *Nature-Inspired Optimizers: Theories, Literature Reviews and Applications*, pp. 23–46, 2020.
[28] F. Orte, J. Mira, M. J. Sánchez, and P. Solana, "A random forest-based model for crypto asset forecasts in futures markets with out-of-sample prediction," *Res Int Bus Finance*, vol. 64, no. August 2022, 2023, doi: 10.1016/j.ribaf.2022.101829.
[29] L. Breiman, "Using iterated bagging to debias regressions," *Mach Learn*, vol. 45, pp. 261–277, 2001.
[30] J. H. Holland, "Genetic algorithms," *Sci Am*, vol. 267, no. 1, pp. 66–73, 1992.
[31] L. Abualigah, M. Shehab, M. Alshinwan, S. Mirjalili, and M. A. Elaziz, "Ant lion optimizer: a comprehensive survey of its variants and applications," *Archives of Computational Methods in Engineering*, vol. 28, pp. 1397–1416, 2021.
[32] H. Abderazek, A. R. Yildiz, and S. Mirjalili, "Comparison of recent optimization algorithms for design optimization of a cam-follower mechanism," *Knowl Based Syst*, vol. 191, p. 105237, 2020, doi: https://doi.org/10.1016/j.knosys.2019.105237.