Forecasting Solar Energy Generation Using Machine Learning Techniques and Hybrid Models Optimized by War SO

Fenghong Pan
School of Electric Power Engineering, Fujian Polytechnic of Water Conservancy and Electric Power, Sanming 366000, Fujian, China
E-mail: 13944238815@163.com

Keywords: solar energy, renewable energy, machine learning, Cat Boost, AdaBoost, Light GBM, War SO

Received: November 19, 2024

Due to threats caused by climate change and energy security, the attainment of adequate and sustainable energy resources is becoming of great importance. There exist promising alternatives to the traditional source, such as solar and wind. However, there are high obstacles to their penetration into a power grid because of the variability and uncertainty in renewable sources. In this regard, it becomes quite necessary to accurately forecast the models so that one can optimize energy generation and guarantee grid stability. This work studies the application of several machine learning algorithms, including Cat Boost, AdaBoost, and Light GBM, to solar energy generation forecasting. The approach has been applied based on data from two solar stations over a period of two years, where the performance of each stand-alone algorithm and a hybrid model that will be optimized with War SO optimizer is analyzed and presented. The standalone CatBoost model demonstrated superior performance, achieving an R² of 0.9106 and RMSE of 4.06 MW in the 30 MW farm. Hybrid models further improved accuracy, with the AdaBoost-War SO model reaching an R² of 0.9836 and RMSE of 1.75 MW. These results confirm the efficiency of utilizing machine learning approaches toward enhancing accuracy in renewable energy forecasting, and therefore hybrid models play an important role in energy prediction with higher accuracy.

Povzetek: Raziskava uvaja hibridne modele strojnega učenja, optimizirane z algoritmom War SO, ki kvalitetno napovedujejo proizvodnjo sončne energije.

1 Introduction

The global community is confronted with several issues concerning the sustainability and security of energy. Failure to address these challenges promptly could result in economic and political turmoil. Depletion of fossil fuel reserves and the environmental consequences of their combustion have sparked greater attention towards the exploration of alternative, sustainable energy sources. Renewable energy technologies such as solar, wind, hydropower, geothermal, and biomass have experienced substantial growth during the last few years, reflective of increasing use in international energy markets [1]. Machine learning techniques have indicated promise in addressing the variability of renewable energy sources such as solar and wind [2].

Research and development in renewable energy have attracted considerable attention lately because of the increasing need for clean and sustainable energy sources [3][4]. Renewable energy therefore comes to the front in efforts to counter greenhouse gas emissions and change due to climatic factors [5–7]. RES offers an assortment of advantages that include a reduction in reliance on foreign sources of energy, jobs, and the possibility of saving money economically [8]. However, the intrinsic variability and unpredictability associated with RES have been a significant obstacle to their wide diffusion [9][10]. For instance, solar energy generation is still very sensitive to factors affecting cloud cover and the seasonal variation of sunlight intensity [11]. All these large variations and uncertainties in renewable energy generation make their smooth integration into the power grid challenging [12].

A necessary strategy to minimize this challenge emphasizes the development of accurate forecasting models in renewable energy generation. Such models are very important in minimizing the negative effects brought about by the variability and uncertainty of the electrical grid. Traditional energy generation forecasts have, for many years, employed techniques such as statistical and physical models [13]. While statistical approaches like the autoregressive integrated moving average model have shown some promise, they are limited in terms of modeling complex nonlinear relationships and high dimensionality inherent in renewable energy signals [14]. Physical models, such as NWP and solar radiation models, play a major role in renewable energy forecasting. However, physical models face serious problems in light of complex dynamics in the Earth's atmosphere and inherent uncertainties in weather prediction. Various research and development works need to be carried out to improve their accuracy. Machine learning algorithms open up a promising direction beyond the limitations of traditional methods that have been developed for forecasting renewable energies [15][16]. First, ML algorithms are excellent at finding complex nonlinear relationships that many big datasets exhibit. For this reason, ML is suitable to handle the multidimensional nature of renewable energy data.

Second, ML algorithms can be easily modified to fit different types of input data: time series, meteorological, and geographical.

This has motivated many researchers to work on the development of machine learning algorithms in predicting solar radiation, one of the critical factors in evaluating the performance of a solar energy system [17]. Voyant et al. have reviewed several approaches to solar irradiation forecasting based on machine learning methods quite extensively. Techniques such as neural networks, support vector regression, regression trees, random forests, and gradient boosting were reviewed. Comparing many different works was challenging because of the characteristics of the diverse nature of the dataset and besides that different metrics of performance were applied. They found very similar errors of prediction overall, from which it follows that there is a huge potential for improving accuracy if hybrid models or ensemble forecasting approaches are implemented [18]. Suanpang and Jamjuntr (2024) benchmarked the Light Gradient Boosting Machine (LGBM) and K Nearest Neighbors (KNN) models for solar power generation forecasting in microgrids. Their results indicated that LGBM performed better than KNN in terms of accuracy ($R^2 = 0.84$ vs. 0.77) and error values (RMSE: 5.77 vs. 6.93; MAE: 3.93 vs. 4.34), although it took more computational power, i.e., longer training time (120 s vs. 90 s) and higher memory (500 MB vs. 300 MB). LGBM was also more consistent over periods and seasons and dealt with outliers effectively. This paper emphasizes the significance of precise prediction in enhancing solar energy utilization in microgrids and elucidates the trade-offs between computational efficiency and prediction accuracy [19]. Singh et al. (2023) introduced a robust hybrid deep learning approach for power prediction using PV, wind, and solar systems in large-scale systems. It uses preprocessing methods and K-means clustering to enhance deep learning training and eliminate noise. A GRU-based recurrent neural network yielded more accuracy than conventional approaches. Pearson coefficient analyses identified interrelations among power sources, with which hybrid renewable clusters were able to minimize forecasting errors and variability. Case studies highlighted the controllability of solar power and the model's success in boosting forecasting for mass systems[20].

Nguyen et al. (2025) identified ambient temperature and humidity as key predictors using SHAP analysis [21], while Zhu et al. (2025) demonstrated the efficacy of hybrid optimization models like HGBoost with satin bowerbird optimizers [22].

Huertas-Tato et al. (2020) blended the forecasts of four models using Support Vector Machines. They evaluated two methods for combining the forecasts and the impact of considering weather-type information in the blends. Results from evaluation at four Iberian Peninsula stations showed large performance gains due to blending, with up to 17% reduction in RRMSE for GHI (16% for DNI), and up to 15% in rMAE. Improvement was similar when evaluating regional forecast skills [23]. Four models were used by Gürel et al. (2020) in modeling solar radiation for the years 2008-2018 in Turkey. The feed-forward neural network outperformed others, followed by Holt-Winters, RSM, and empirical models [24]. Alizamir et al. (2020) estimated the performance of six machine-learning models for solar radiation forecasting at selected stations in Turkey and the USA. These authors compared different models by applying several statistical indicators and pointed out that GBT outperformed the others. GBT decreased the average RMSE by 0.26% to 19.34% for one station and by 4% to 54.8% for the other one, indicating an effective use of climatic parameters for solar radiation prediction [25]. Koo et al. developed a new methodology for estimating the monthly average daily solar radiation in China using different machine-learning techniques. Their approach was to use clustering and enhanced case-based reasoning models, which have given an average prediction accuracy of 93.23% when applied to data from 97 cities over a continuous period of 10 years. This may thus provide a very effective way of implementing solar energy systems, enabling decision-makers to determine the best locations and configurations [26]. Nath et al. (2020) discussed two machine-learning techniques for hourly solar power forecasting. Their work was focused on how to enhance energy grid integration and service quality by optimizing data preprocessing, feature selection, weather profiling, and choosing the algorithms that provide better accuracy and efficiency in the forecast of solar power, thus helping to meet global energy demands [27]. Kumar et al. (2020) suggested a short-term solar energy forecast using PI-based machine learning. The authors support the fact that their approach makes the forecast more accurate and reliable than in the case of deterministic methods, which is urgent for grid stability and reliability, considering the stochastic nature of photovoltaic power generation [28]. Jebli et al. (2021) introduced a machine and deep learning-based solar energy forecasting approach critical to increasing competitiveness for solar power plants and reducing reliance on fossil fuels. The authors conducted their research on Errachidia, Morocco, for data from 2016 to 2018, using RF and ANN models, outperforming other methods such as LR and SVR. Comparisons with Pirapora, Brazil, enhanced the quality and reproducibility of this study [29]. In this regard, Abualigah et al. (2022) reviewed all kinds of learning-based modeling for renewable power source estimation by focusing on recent deep learning and machine learning algorithms. Then they discussed the performance analysis based on the new taxonomy, challenge, and possibility for the future research direction. Based on this, the paper has highlighted that hybrid learning techniques were effective in addressing energy generation problems and thus suggested using these techniques for improvement in forecasting accuracy [30].

Nevertheless, it is noted that other well-known algorithms, i.e., XGBoost and neural networks, are widely used in the renewable energy forecasting literature. The exclusion of XGBoost is mainly due to its similarity to LightGBM, which is more computationally efficient and tailor-made for big datasets. Neural networks, including recurrent and convolutional architectures, offer significant advantages to their ability to capture temporal and spatial patterns; yet, they are computationally intensive and need big datasets. By the above study's emphasis on high-frequency but site-specific data, gradient-boosting models were preferred as they can balance accuracy, computational cost, and interpretability. Subsequent studies can be focused on using neural networks or hybrid models, for instance, the integration of gradient-boosting methods and deep learning, to leverage their respective strengths. Moreover, a finer comparative study between XGBoost and different neural network configurations can provide further clarity into their utility to solar power forecasting projects in comparable situations.

Notwithstanding improvements in solar forecasting, significant gaps exist in current methods. Most research uses low-resolution data sets, i.e., hourly or daily measurements, that are incapable of recording short-term variability important for real-time grid integration. This research bridges this gap by using high-frequency, 15-minute interval data from two solar farms to enable better modeling of dynamic conditions.

Feature selection in modern state-of-the-art (SOTA) techniques has the tendency to apply simple methods that do not consider non-linear variable interactions. The Delta Moment Independent Measure (DMIM) introduced in this paper is utilized in the identification of vital predictors like solar irradiance, and it offers improved input selection for prediction purposes. Additionally, although hybrid models have been promising, standard optimization techniques like grid search or genetic algorithms hinder their potential. The employment of the War SO optimizer in this research surpasses these constraints by improving model accuracy and computation time.

One of the most prominent limitations seen in existing work is the lack of multi-site validation, which casts doubt on results. The work demonstrates the generality of the models suggested by validation with data from two farms with capacities of 30 MW and 130 MW. Besides, the majority of SOTA work relies on a limited collection of performance measures such as RMSE or R². Nevertheless, this work applies a comprehensive evaluation framework including MAE, runtime, and convergence analysis to enable thorough inspection.

By bridging these gaps, this research establishes a new standard in solar forecasting, pushing the boundaries of high-resolution data use, robust feature engineering, hybrid optimization, and generalizable model development.

Despite the huge advancements in the prediction of renewable energy, there are several challenges to limit the scalability and viability of machine learning models in the same. Most notable are the data security and privacy concerns, since the collection of sensitive operational data from solar farms is usually an essential stepping stone for model development; however, sharing the same is fraught with risks and dissuades collaboration. Also, integrating advanced machine learning models into existing energy systems, particularly legacy-based ones, is extremely challenging and must be harmonized with existing solar forecasting methods to facilitate easy implementation. Hybrid models improve forecast accuracy but often come with maintenance and integration issues with operational systems, which could deter use. These research gaps are overcome by this research using high-frequency 15-minute interval data to raise the level of granularity and predictive accuracy of the models beyond what has been possible using hourly or daily data. The use of hybrid machine learning models that have been optimized with the War Strategy Optimizer generates superior predictive capacity and computational efficacy than standard procedures. Feature robustness through strong feature selection further secures model stability against overfitting, contributing to methodological robustness. By highlighting the scalability of the hybrid models, computational efficiency, and integrability feasibility, the research translates theoretical findings toward practical realization for solar energy prediction. Problems of data privacy, system compatibility, and real-time deployments, further improving scalable and actionable models, are to be addressed in follow-up studies. Table 1 indicates the comparing the results of the discussed studies.

Table 1: Comparison of the results of the discussed studies

Study	Models Used	Metrics	Key Contributions
Suanpang & Jamjuntr (2024)[19]	LGBM, KNN	R^2 = 0.84 (LGBM), RMSE = 5.77 W (LGBM)	Benchmarking LGBM and KNN for microgrid forecasting; LGBM showed superior accuracy.
Singh et al. (2023)[20]	GRU-based hybrid deep learning model	Improved accuracy over conventional models	Use of K-means preprocessing and GRU for large-scale systems.
Nguyen et al. (2025)[21]	CatBoost, SHAP Analysis	R^2 = 0.46, RMSE = 4.748 W (CatBoost)	Identified ambient temperature and humidity as key predictors.
Zhu et al. (2025)[22]	Hybrid models (HGBoost + optimizers)	R^2 = 0.9907	Hybrid optimization with satin bowerbird optimizer.
Huertas-Tato et al. (2020)[31]	Blending ML models	Up to 17% RRMSE reduction	Blended forecasts using SVM, leveraging weather-type information.
Alizamir et al. (2020)[32]	Gradient Boosting Trees (GBT)	RMSE reduction: 0.26%–19.34%	GBT demonstrated superior

2 Methodology

The research methodology is divided into two main sections. Firstly, the data acquisition process is outlined, detailing how the relevant data was collected and sourced. Following this, the second part delves into the machine-learning algorithms utilized in the study, providing an overview of each algorithm and explaining the methods employed for their implementation in the research context. This study employs machine learning techniques to forecast energy generation from solar sources. Initially, the available data undergoes preprocessing using various methods. A crucial secondary analysis assesses the impact of input features on the output by examining correlations among parameters with the Pearson Correlation Coefficient. Following this, the dataset is split into training and testing subsets to enable accurate energy consumption prediction. Pearson Correlation Coefficient was selected as the feature selector because it is simple, interpretable, and computationally fast. The method is efficient at detecting linear associations between the input features and target variable and hence qualifies as an appropriate first-line approach to filtering out the relevant features in the structured numerical data set employed in this research. CatBoost, AdaBoost, and Light GBM algorithms are individually and collectively employed for training and prediction to improve model accuracy. The hyperparameters of these algorithms are optimized using the War SO optimizer. Performance comparison between single algorithms and hybrid models is conducted using various statistical indicators to identify the most effective approach for energy generation prediction. The focus of the present study is on CatBoost, LightGBM, and AdaBoost because of their proven performance and efficiency in renewable energy prediction tasks. The algorithms are all gradient-boosting methods with the ability to process structured datasets, prevent overfitting, and identify complex non-linear variable relationships. Specifically, CatBoost is effective in processing categorical features and removing prediction bias, while LightGBM and AdaBoost are known for scalability and iterative learning, respectively.

The entire modeling process is depicted in Fig 1.

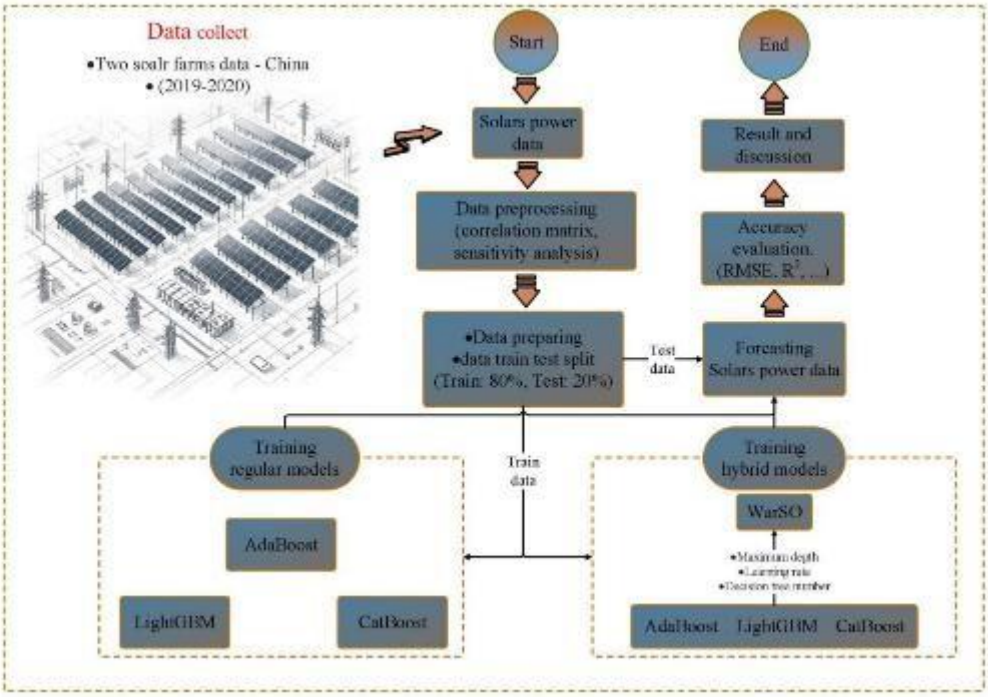

Figure 1: Flowchart diagram of the current investigation

1. Data

For this study, data collection involved the procurement of solar generation data from various on-site renewable energy stations situated across China. Specifically, information was gathered from two solar stations. Over a period spanning two years, from 2019 to 2020, data was meticulously recorded at 15-minute intervals. This dataset, comprising power generation data alongside weather-related parameters, was subsequently utilized in the Renewable Energy Generation Forecasting Competition hosted by the Chinese State Grid in 2021 [33]. Table 2 summarizes all data columns along with their respective descriptions.

Table 2: The input variables and their statistical details in the farm with a capacity of 30(MW)

	count	mean	std	min	25%	50%	75%	max
Year	20352	2019	0	2019	2019	2019	2019	2019
Day	20352	15.66037736	8.767529044	1	8	16	23	31
Month	20352	4.018867925	2.00467189	1	2	4	6	7
Hour	20352	11.5	6.92235662	0	5.75	11.5	17.25	23
Minute	20352	22.5	16.77092186	0	11.25	22.5	33.75	45
Total solar irradiance (W/m2)	20352	198.8134336	294.5791441	0	0	0	338.25	1117
Direct normal irradiance (W/m2)	20352	100.7295597	185.090418	0	0	0	112	760
Global horizontal irradiance (W/m2)	20352	69.30508058	101.8772566	0	0	0	111	656
Atmosphere (hpa)	20352	1016.013768	9.323415494	994.8	1008.3	1014.7	1024	1038.6
Relative humidity (%)	20352	58.24924332	13.15880075	14.1	50.9	61	68.6	80.5
Power (MW)	20352	5.449246788	8.258662461	0	0	0.115101	9.03832125	29.9113395

Table 3: The input variables and their statistical details in the farm with a capacity of 130(MW)

	count	mean	std	min	25%	50%	75%	max
Year	70176	2019.500684	0.500003095	2019	2019	2020	2020	2020
Day	70176	15.73871409	8.803983506	1	8	16	23	31
Month	70176	6.519835841	3.449575468	1	4	7	10	12
Hour	70176	11.5	6.922235873	0	5.75	11.5	17.25	23
Minute	70176	22.5	16.77062932	0	11.25	22.5	33.75	45
Total solar irradiance (W/m2)	70176	169.3033665	248.0776381	0	0	0	305.7575	1041.93
Direct normal irradiance (W/m2)	70176	122.1523955	178.9880244	0	0	0	220.6025	751.75
Global horizontal irradiance (W/m2)	70176	78.29928152	117.5873435	0	0	0	129.57	561.8
Air temperature (°C)	70176	13.69510759	12.03580036	-13.92	3.19	15.46	23.57	40.47
Atmosphere (hpa)	70176	861.0362624	6.147644763	844.51	856.2175	860.87	865.35	881.67
Power (MW)	70176	19.56748845	27.939605	0	0.241033	0.3269	36.8217485	109.3603

The difference in parameter scales shown in Tables 2 and 3 reflects the importance of location factors in determining solar energy output. In particular, the mean total solar irradiance measured for the 130 MW farm is higher than that of the 30 MW farm due to its larger geographic area and changing environmental conditions. These variations were adjusted for while training the models by normalizing the datasets individually for each farm, such that the models could learn to adapt to site-specific trends.

The 15-minute, high-resolution datasets also provided valuable detail for short-duration solar irradiance changes and other variables. The time resolution of the data enabled the models to detect rapid weather changes, increasing the accuracy of energy prediction. We appreciate that summary statistics in Tables 2 and 3 are unable to capture the full richness of temporal data variation. Graphical representations, such as time series plots, would be an asset in future research to better present the dynamics measured at this scale level.

2.2 Machine learning methods

This study employed advanced machine learning algorithms, including Cat Boost, AdaBoost, and Light GBM, for energy generation forecasting[34]. To improve accuracy and adaptability, a hybrid model was developed by incorporating War SO optimizers. This section provides a concise summary of the mathematical formulations and fundamental principles underlying each of these techniques.

2.2.1 Categorical gradient boosting (cat boost)

The Cat Boost [35] model is a boosting-based algorithm that constructs trees in a level-wise manner. While the overall boosting process resembles existing methods, there are notable distinctions. Instead of performing residual calculations on all training data collectively, Cat Boost selects a subset of the data for residual calculations to build a model. Subsequently, it utilizes the predicted values from this model to process the residual of subsequent data. Moreover, Cat Boost employs Random Permutation to randomly select data, thereby promoting diversity in tree creation and preventing Overfitting [36].

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n}\left[x_j^i = x_k^j\right]y_j + \alpha p}{\sum_{j=1}^{n}\left[x_j^i = x_k^j\right] + \alpha} \quad (1)$$

Here, α represents the corresponding weight, P denotes a prior value, xk = (x_k^1, ..., x_k^m) signifies the random vector of m features and yk INR den.

2.2.2 Adaptive boosting (AdaBoost)

Initially designed as a feature classification algorithm in machine learning, AdaBoost has expanded its application to regression problems [37]. Currently, it finds widespread use in load forecasting and short-term wind speed forecasting, yielding promising results. The core concept involves training multiple weak learners within the same sample space and subsequently adjusting their weights to construct a robust learner based on the prediction outcomes of each weak learner [38].

The specific steps of the AdaBoost algorithm are delineated as follows:

1. Selection of Basic Learner and Data: Initially, the weak learning algorithm and sample space (xi, yi) are determined. The sample data is denoted as group M, and the sample data is normalized with a mean of 0 and a variance of 1, where xi ∈ Rn and yi ∈ Rn.

2. Network Initialization: Assuming a uniform sample distribution, the weight of the test data's uniform distribution, Dt(i), is set to 1/M. The neural network structure is configured based on the characteristics of the sample data, followed by the initialization of weights and thresholds for the neural network. Finally, the number of iterations is set.

(3) Weak Predictor Prediction: The t-th weak predictor undergoes training using the training data, resulting in the prediction output for the training data. Following this, the error ei and average error et of the weak learner are computed for each sample using the calculation formula.

$$e_t = \frac{1}{M}\sum_{i=1}^{M} e_i, i = 1,2,\ldots,M \quad (2)$$

(4) Computing the Weight of Weak Learner: Based on the average error et of the prediction sequence f(t), the weight of the weak learner is determined accordingly.

$$a_t = \frac{1}{2}ln(\frac{1-e_t}{e_t}) \quad (3)$$

(5) Updating Sample Weights: Adjusting the weights of the next round of training samples is based on the current weight a_t. The formula for updating sample weights can be expressed as:

$$D_t(i) = \frac{D_{t-1}(i)}{B_t} * exp[-a_t y_i f_t(x_i)] \quad (4)$$

Here, Bt represents the normalization factor, which ensures that the sum of distribution weights equals 1 while maintaining the weight proportion unchanged. ft(xi) refers to a weak predictor acquired after training the data.

(6) Strong Predictor Formation: Following t rounds of training, t sets of weak predictor functions are acquired. Subsequently, strong predictors are constructed by amalgamating these t sets of weak predictor functions, as expressed below:

$$F(x) = \sum_{t=1}^{T} a_t . f_t(x) \quad (5)$$

In this context, T symbolizes the total count of weak learners.

2.2.3 Light gradient boosting machine (Light GBM)

Light GBM [39] stands out as a boosting-based algorithm recognized for its speed and precision in forecasting, surpassing other boosting and bagging algorithms. It leverages a gradient-boosting decision tree (GBDT) framework, incorporating gradient-based one-sided sampling and exclusive feature-bundling techniques. Unlike traditional gradient boosting machine (GBM) tree splitting methods, Light GBM adopts a leaf-wise approach, which enhances accuracy through more intricate modeling, particularly advantageous for time series forecasting. This method, combined with gradient boosting decision tree (GBDT) and leaf techniques, leads to low memory usage and rapid training. Light GBM encompasses several hyperparameters, with learning rate, number of iterations, and number of leaves being crucial for forecasting accuracy. Additionally, Light GBM addresses overfitting by adjusting Col sample by tree and subsample hyperparameters. Proven effective in various time series forecasting domains such as electricity load and solar power forecasting, Light GBM's single-output forecasting demonstrates both rapidity and precision. Given the need for a fast and precise forecasting model with a single output, Light GBM is chosen for construction.

2.2.4 War strategy optimizer (War SO)

In the strategy of warfare, there are three primary factions: the King (K), the Commander (C), and the soldiers. Both the Commander and the King serve as leaders on the battlefield, overseeing the actions of the soldiers. Each soldier has an equal chance of rising to the ranks of Commander or King based on their combat effectiveness, which is measured by a cost function. However, there is a possibility that the Commander or King may face tough opposition from rival soldiers, representing a local optimum. These adversarial soldiers wield sufficient power to potentially ensnare the leaders. To avert such scenarios, the soldiers are managed in their coordinated tactics and maneuvers, guided by the status of the Commander or King [40]

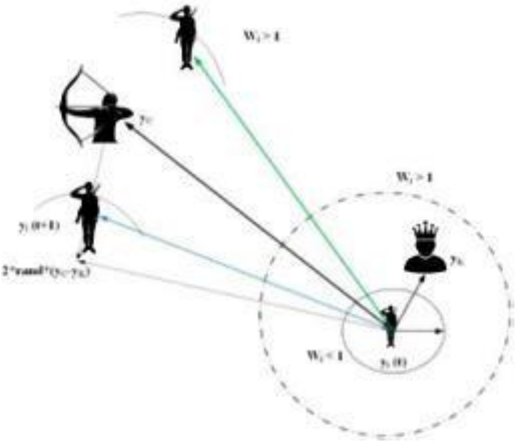

Figure 2: Renewing the attack model mechanism [41]

2.2.4.1 Attack tactic

Attack tactics are crucial components of war strategies, and this paper models two distinct policies. In the first strategy, each soldier updates their own status based on the current situation of the Commander and King. Figure (2) illustrates the procedure for updating the attack model. A favorable circumstance triggers the King to initiate a significant attack, with the soldier possessing the highest attack force or cost being appointed as the King. Initially, at the commencement of the war, all soldiers are endowed with equal weight and rank. However, their rank escalates as they effectively execute tactics. It's noteworthy that soldiers' weight and rank may be adjusted based on the success of tactics during the war's progression. As the war nears its end, the circumstances of the soldiers, Commander, and King converge as they move towards achieving the goal outlined in equation (6).

$$y_i(t+1) = y_i(t) + 2p(y_c - y_k) + rand(y_k * w_i - y_i(t)) \quad (6)$$

In this context, yi(t+1) and yi(t) denote the current and preceding statuses of the soldier, respectively. yK and yC denote the situations of the King and the Commander, while Wi represents the weight.

2.2.4.2 *Renewing weight and rank*

The renewal of each individual's situation is correlated with the status of the King, the location of the Commander, and the ranking of the soldiers. Soldiers' rankings are determined by their past performance in the war, which in turn influences the Wi factor. The ranking of each soldier signifies their proximity to achieving the goal (cost value). If the attack force (cost) in the previous situation (Fper) is significantly higher than that in the new situation (Fnew), the soldier opts to retain the previous situation, as depicted in equation (7).

$$y_i(t+1) = y_i(t) \times (F_{new} < F_{per}) + y_i(t+1) \times (F_{new} \geq F_{per}) \quad (7)$$

If soldiers successfully renew their situation, their ranking (Rai) will be upgraded, as shown in equation (8). Using this ranking, the updated weighting can be computed as described in Equation 9.

$$Ra_i = Ra_i \times (F_{new} < F_{per}) + (Ra_i + 1) \times (F_{new} \geq F_{per}) \quad (8)$$

$$w_i = w_i \times (1 - \frac{Ra_i}{Max_{iter}})^{\beta} \quad (9)$$

2.2.4.3 Defense strategy

Another approach to updating the situation involves the King, a randomly selected soldier, and the Commander's status. However, the adjustment of weight and ranking remains consistent, as illustrated in equation (7).

$$y_i(t+1) = y_i(t) + 2p(y_k - y_{rand}(t)) + rand * w_i * (y_c - y_i(t)) \quad (10)$$

Unlike the prior policy, this military strategy ventures into broader territories when incorporating the status of the randomly selected soldier. Soldiers make substantial strides in updating their situation when larger Wi values are present. Conversely, when Wi amounts are small, the opposite occurs.

2.2.4.4 Substituting the vulnerable soldier

Throughout the duration of the conflict, the weakest soldier, distinguished by the lowest value of the cost function, is singled out for replacement. This study investigates multiple strategies for substitution in such instances. The simplest approach involves replacing the weak soldier with a randomly chosen one, as determined by the formula below [equation (11)]:

$$y_w(t+1) = L_L + rand \times W_i \times (H_L - L_L) \tag{11}$$

The second approach involves substituting the weakest soldier with one in close proximity to the average of the entire army in the field, as represented by the following formula. This tactic is aimed at enhancing the convergence of the optimizer [equation (12)]:

$$y_w(t+1) = y_k - (1 - rand) \times (y_w(t) - median(y)) \tag{12}$$

2.2.4.5 Key features of the provided optimizer

The proposed optimizer possesses several important features that enhance the optimization process. Firstly, it achieves a satisfactory balance between the exploitation and exploration phases. Each individual (soldier) in this optimizer is assigned a unique weight based on their ranking. Moreover, weight adjustment only takes place if there is an enhancement in the individual's cost value during the updating phase, and this adjustment is tied to the particle's position in relation to the positions of the Commander and King. The fluctuation in weights follows a nonlinear pattern, with substantial alterations happening in the initial epochs and diminishing ones towards the conclusion, aiding in quicker convergence to the global optimum. Moreover, the situation updating involves two steps, enhancing exploration capabilities towards the global optimum. This optimizer is recognized for its simplicity, requiring fewer computations.

2.2.4.6 The stages of exploitation and exploration

The concepts of exploitation and exploration are fundamental principles in metaheuristic optimizers and are crucial for their effectiveness. The proposed optimizer maintains a balanced trade-off between these two phases. The attack tactic is representational of the exploitation side, whereby the optimizer leverages known solutions to its advantage in furthering optimization performance. Conversely, the defense tactic symbolizes exploration in allowing the optimizer to move toward newer areas of the search space and hopefully come up with far better solutions than previously obtained. This balanced approach ensures efficient optimization by leveraging both exploitation and exploration strategies.

2.2.5 Model verification and evaluation

In the present study, the effectiveness of the forecasting model is tested with several error analysis measures. These include RMSE, MAE, RAE, JSD, VAF, and R-squared. These measures test the accuracy of the model and differences in values forecasted with the model against real ones. The comprehensive evaluation will give full insight into the performance of the model and indicate if some improvement might be necessary [42]. Detailed mathematical expressions for these statistical evaluation metrics are provided in Table 4.

Table 4: Statistical evaluation indexes

Statistics	Criteria	Equation
RMSE	Root Mean Squared Error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{T}}$
MAPE	Mean Absolute Error	$\frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{n}$
VAF	Variance Accounted For	$100\% \times \frac{\sum_{i=1}^{n}(y_i - \bar{y})(f_i - \bar{f})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
JSD	Jensen Shannon Divergence	$\frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)^*$
R2	Coefficient of Determination	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
RAE	Relative Absolute Error	$\frac{[\sum_{i=1}^{n}(\hat{y}_i - y_i)^2]^{\frac{1}{2}}}{[\sum_{i=1}^{n}(y_i)^2]^{\frac{1}{2}}}$

*For more details, refer to Nielsen (2021)[43].

3 Results

This section outlines the results and analyses derived from the energy generation forecasting process. It begins with an introduction to the standalone algorithms CatBoost, LightGBM, and AdaBoost, followed by their hybrid configurations fine-tuned using the WarSO optimizer. A comprehensive array of charts and tables is provided to facilitate the assessment of the models.

Fig 3 depicts the correlation matrix created for the selected parameters in energy generation using solar energy at the first site considered with a capacity of 30 megawatts. Examination of the correlation matrix (depicted in Fig 3) indicates that total solar irradiance, direct normal irradiance, and global horizontal irradiance collectively play a substantial role. Notably, the parameter "global horizontal irradiance" exhibits the strongest correlation with the target parameter. Temperature variables display positive effects and correlations, while the impact of other parameters on the target parameter is minimal.

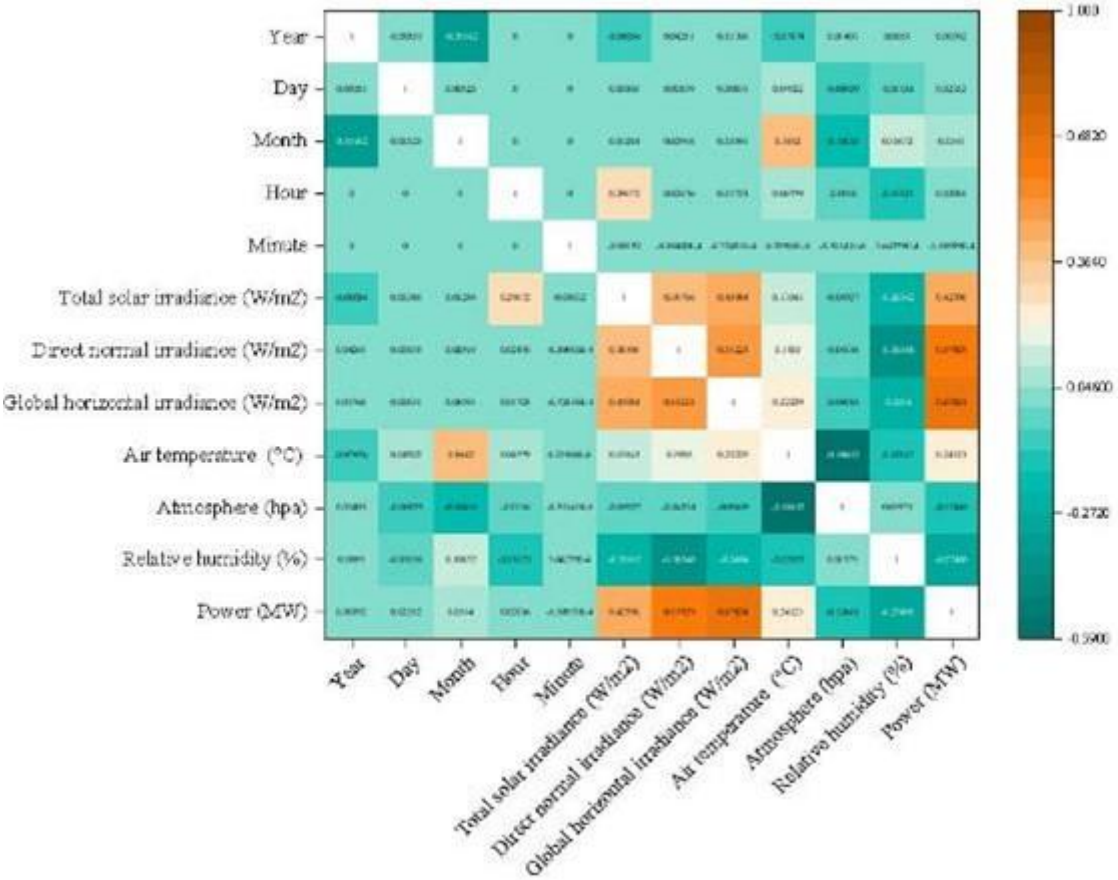

Figure 3: The correlation matrix of features in the farm with a capacity of 30(MW)

Fig 4 illustrates the correlation matrix generated for the selected parameters in solar energy generation at the second site under consideration, which has a capacity of 130 megawatts. Similar to the 30-megawatt case, three parameters, namely total solar irradiance, direct normal irradiance, and global horizontal irradiance, exhibited a very strong correlation with the target parameter. Among these, the total solar irradiance parameter demonstrated the highest correlation. Additionally, temperature and hour parameters showed positive correlations, while the remaining parameters exhibited negligible and almost neutral correlations.

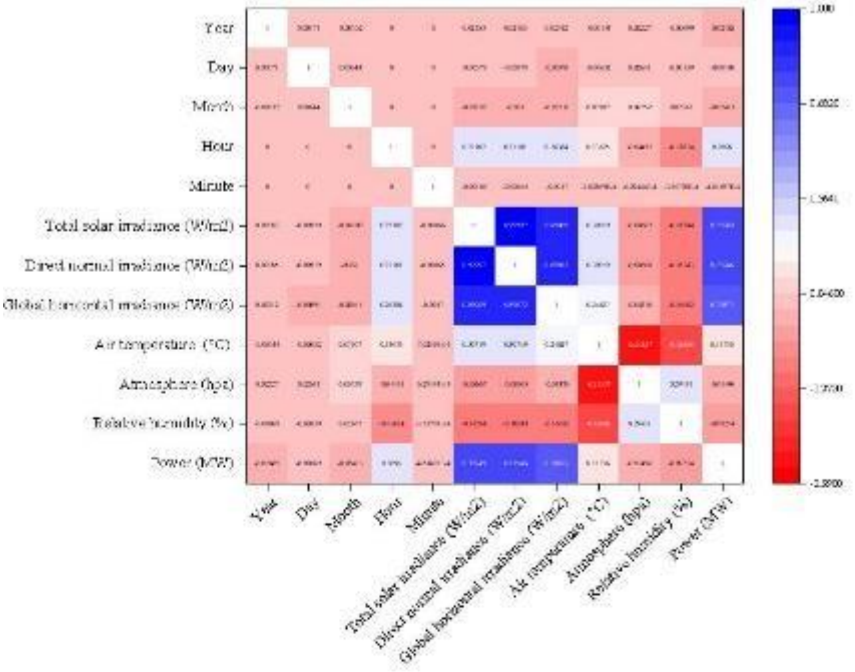

Figure 4: The correlation matrix of features in the farm with a capacity of 130(MW)

In this study, the Delta Moment independent index was used to assess the impact and sensitivity of input parameters on the output. The scaled values range between 0 and 1. Fig 5 illustrates the impact and sensitivity of input parameters at the 30-megawatt site. According to this figure, the three primary parameters, total solar irradiance, direct normal irradiance, and global horizontal irradiance, exhibited very high sensitivity. Additionally, the hour parameter also showed significant influence based on this index. Other parameters showed relatively similar sensitivity to the output.

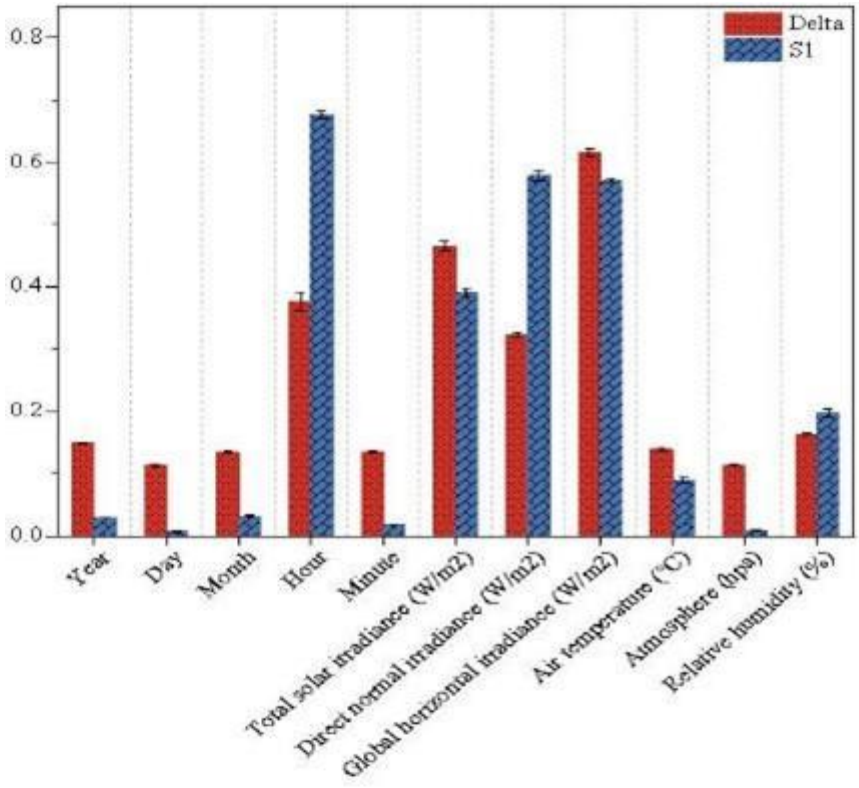

Figure 5: Sensitivity analysis of variables based on the DMIM method in the farm with a capacity of 30(MW)

Fig 6 also illustrates the sensitivity analysis of input parameters at the 130-megawatt site. In this case, the three parameters, total solar irradiance, direct normal irradiance, and global horizontal irradiance, showed higher sensitivity, with total solar irradiance exhibiting the greatest sensitivity. Similarly, the hour parameter had a significant impact and demonstrated high sensitivity at this site. Other parameters, such as temperature, humidity, atmospheric pressure, and large-scale time parameters, showed relatively similar levels of influence and sensitivity.

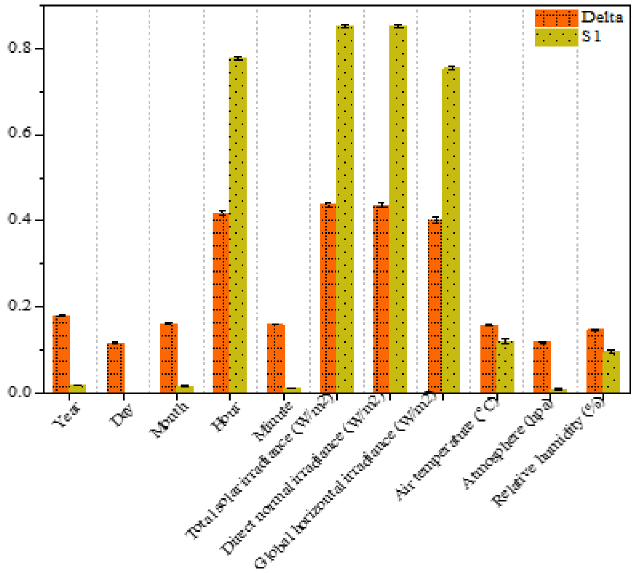

Figure 6: Sensitivity analysis of variables based on the DMIM method in the farm with a capacity of 130(MW)

Fig 7 displays the time series of observational and computational data based on single algorithms for predicting energy production in the 130-megawatt solar farm. In addition to the time series plots, scatter plots for each method are also provided. According to Figure 7, both the training and testing sections showed better overlap between observational and computational data for the Cat Boost algorithm, indicating its satisfactory performance. Conversely, the AdaBoost algorithm exhibited the poorest performance. Furthermore, based on the scatter plot, the Cat Boost algorithm demonstrated the highest correlation with observational data, with an R2 value of 0.8980, making it the most suitable algorithm.

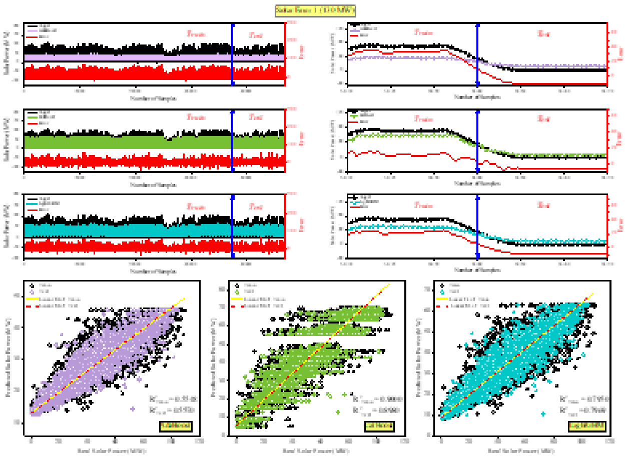

Figure 7: A detailed analysis of the outcomes from employing the AdaBoost, Light GBM, and Cat Boost models in the 130-megawatt solar farm

Fig 8 also depicts the time series of observational and computational data for the 30-megawatt farm. According to the results, the Cat Boost algorithm outperformed other algorithms in this case as well, exhibiting lower error and higher correlation with observational data. Additionally, based on the scatter plot, the Cat Boost algorithm demonstrated the best performance for energy production prediction, with an R2 value of 0.9106.

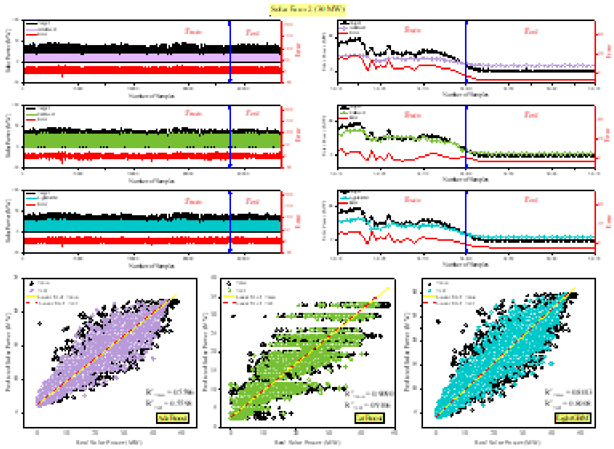

Figure 8: A detailed analysis of the outcomes from employing the AdaBoost, Light GBM, and Cat Boost models in the 130-megawatt solar farm

To conduct a comprehensive assessment of the algorithms' performance and accuracy, various statistical indicators were evaluated and compared, as presented in the preceding section. The results for these indicators are also depicted in Table 5.

Table 5: Error metrics for proposed Cat Boost, AdaBoost, and Light GBM models

Farm	**Optimizer**	**MAE (Train)**	**RMSE (Train)**	**R^2 (Train)**	**JSD(Train)**	**VAF(Train)**	**RAE(Train)**
130 MW	AdaBoost	15.575	18.732	0.555	213905.3	55.48413	0.546277
	CatBoost	6.832367	8.875713	0.900052	70437.51	90.00517	0.258847
	LightGBM	10.41079	12.71163	0.794991	125100	79.49912	0.370716
30 MW	AdaBoost	7.691782	9.115634	0.559622	30828.01	55.96231	0.54213
	CatBoost	3.356744	4.144317	0.908976	7841.534	90.89757	0.246473
	LightGBM	5.124735	6.122702	0.801327	14322.59	80.13274	0.364133
Farm	**Optimizer**	**MAE (Test)**	**RMSE (Test)**	**R^2 (Test)**	**JSD(Test)**	**VAF(Test)**	**RAE(Test)**
130 MW	AdaBoost	15.39225	18.32239	0.556952	71634.8	55.70914	0.54584
	CatBoost	6.778392	8.78999	0.898032	23920.81	89.80388	0.261861
	LightGBM	10.27857	12.40571	0.796891	42103.83	79.69671	0.369577
30 MW	AdaBoost	7.718301	9.029883	0.559829	10024.49	55.98797	0.542413
	CatBoost	3.339026	4.068872	0.910627	2533.42	91.06506	0.244412
	LightGBM	5.146532	6.075019	0.800771	4718.119	80.08289	0.364919

Hybrid models were devised to boost prediction accuracy and benchmark against individual algorithms. Employing the War SO algorithm, optimization was applied to the Cat Boost, AdaBoost, and Light GBM algorithms. Based on Fig 9, both observational and computational time series results are presented for both farms. According to Figure 8, for Farm 1, the AdaBoost-War SO hybrid model outperformed its single model counterpart, exhibiting the lowest error rate. Similarly, for the 30-megawatt Farm 2, the AdaBoost-War SO hybrid model proved suitable for prediction purposes.

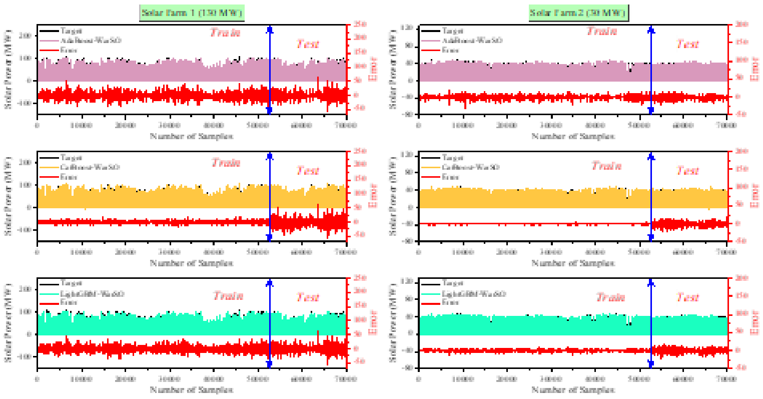

Figure 9: Evolution of Observed and Predicted Values using Hybrid Models of AdaBoost, LightGBM, and CatBoost

To comprehensively analyze and identify the most appropriate prediction algorithms, as well as evaluate their performance, scatter plots for each hybrid model are displayed in Figure 10. These plots visualize the R2 index for both the training and testing datasets. Based on Figure 10, in the first farm, the hybrid AdaBoost-War SO model achieved the highest performance with an R2 value of 0.9784, while in the second farm, a similar result was observed with the hybrid AdaBoost-War SO m

odel achieving an R2 value of 0.9836. Following these models, the hybrid Light GBM-War SO model proved to be suitable for prediction in both farms.

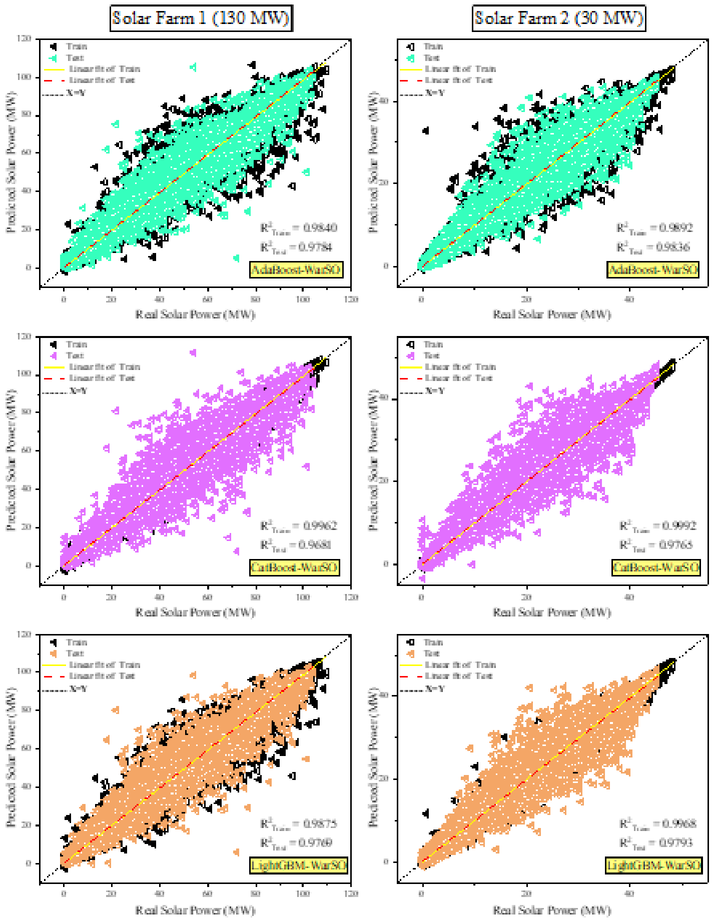

Figure 10: Scatter plot of the observation-prediction for AdaBoost, Light GBM, and Cat Boost hybrid models in Farm1 and Farm2

Fig 11 illustrates the scatter plot of errors in hybrid models for both the training and testing phases. According to this figure, during the training phase, the Cat Boost-War SO model performed the best in both farms. However, during the testing phase, although the results are close, the AdaBoost-War SO model exhibited lower error ranges in both farms, indicating its suitability for prediction purposes.

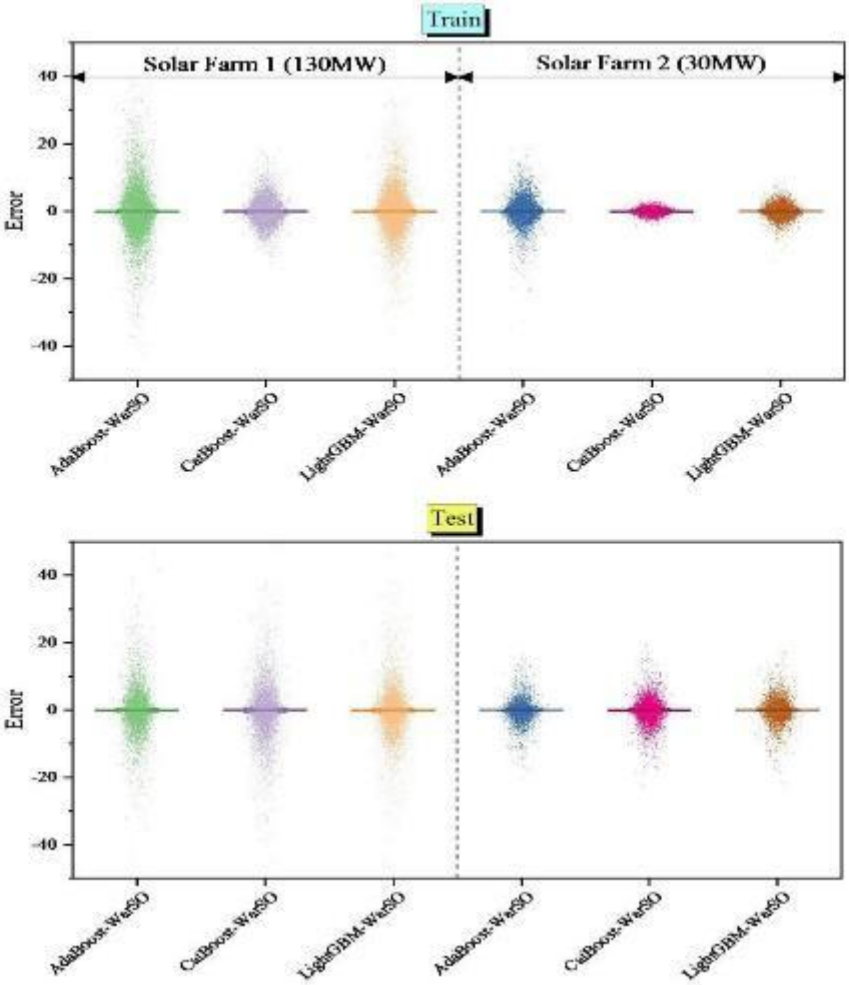

Figure 11: plots of error measurements for models during the testing and training phases in Farm1 and Farm2

Fig 12 displays the error metrics calculated for the hybrid models proposed for energy production prediction in the first farm. The calculated metrics include RMSE, R2, RAE, JSD, MAE, and VAF. Considering the two important metrics, RMSE and R2, from the RMSE plot in the testing section, it is evident that the AdaBoost-War SO hybrid model had the lowest error. Following this model, the Light GBM-War SO model proved suitable for prediction. Additionally, considering the R2 metric, it is evident from the rectangular plot in the testing section that the AdaBoost-War SO hybrid model had the highest R2. The other metrics also support this trend.

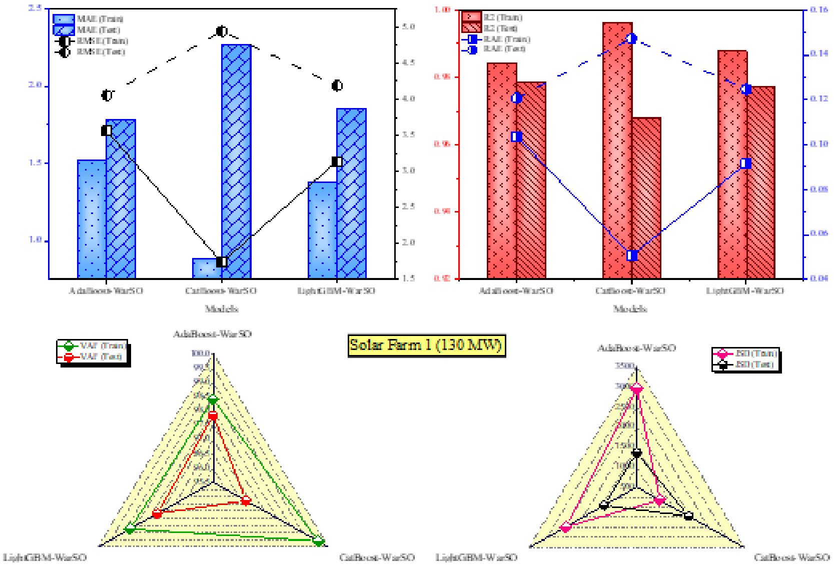

Figure 12: Performance Metrics Visualization for Proposed Models in Farm 1(130MW)

Fig 13 also illustrates the error metrics calculated for the hybrid models in the second farm with a capacity of 30 megawatts. Similar to the first farm, according to the presented metrics, the AdaBoost-War SO hybrid model has proven to be the best model for prediction in this farm as well. Details and numerical values for each of the indicators for hybrid models are presented in Table 6.

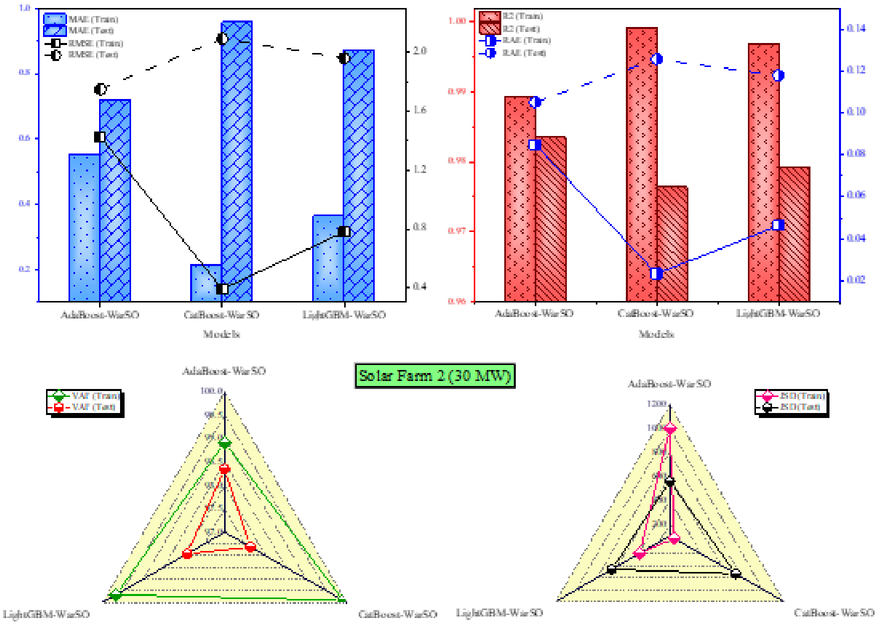

Figure 13: Performance metrics visualization for proposed models in Farm2 (30MW)

Table 6: Error metrics derived from the application of Cat Boost, AdaBoost, and Light GBM hybrid models

Optimizer	AdaBoost-War SO	Cat Boost-War SO	Light GBM-War SO	AdaBoost-War SO	Cat Boost-War SO	Light GBM-War SO
	Farm1(130 MW)			Farm2(30 MW)		
	Train					
MAE	1.51868	0.884092	1.373611	0.550378	0.214323	0.364667
RMSE	3.55647	1.735859	3.138754	1.430018	0.391776	0.779854
R2	0.983952	0.996177	0.987501	0.989162	0.999187	0.996777
JSD	2956.739	1143.396	2482.084	1001.383	143.2308	388.9295
VAF	98.39525	99.61771	98.75007	98.91624	99.91866	99.67769
RAE	0.103719	0.050624	0.091537	0.085047	0.0233	0.04638
	Test					
MAE	1.784966	2.265224	1.853122	0.71745	0.954989	0.870113
RMSE	4.053808	4.942803	4.187876	1.74971	2.094413	1.96203
R2	0.978312	0.967757	0.976854	0.983473	0.97632	0.979219
JSD	1356.879	1951.165	1413.484	555.6971	736.2619	659.5611
VAF	97.83394	96.77995	97.68843	98.35126	97.63708	97.92325
RAE	0.120766	0.14725	0.12476	0.105103	0.125809	0.117857

Fig 14 presents the runtime performance of hybrid models over 500 iterations. Based on Fig14, in the first farm, the AdaBoost-War SO hybrid model had the longest runtime with 3403 seconds, followed by the Light GBM-War SO hybrid model. Similarly, in the second farm, the AdaBoost-War SO model had the longest runtime, totaling 3960 seconds. The Cat Boost-War SO model had the shortest runtime in both farms.

Although hybrid models, in particular AdaBoost-War SO, are more accurate, their usability in the real world is diminished by high runtime expenses. As Fig 14 shows, the AdaBoost-War SO model is significantly more computationally expensive than solo models, with some instances requiring over 3960 seconds of runtime. The high computational demand stems from the ensemble learning iteratively on top of the optimization approach being used by the War Strategy Optimizer.

The extended period of operation might pose challenges in real-time operations or scenarios defined by limited computing resources. Despite the improvements in precision, justification for the use of hybrid models in critical forecast scenarios is met, their practicality in the context of sparse resources, for instance, in edge devices or small-scale microgrids, might be limited. To mitigate this trade-off, future work is invited to investigate optimization methods, including model parallelization, hardware acceleration, or pruning strategies, to minimize runtime without compromising accuracy. Furthermore, combining hybrid models with distributed computing platforms can increase their scalability for large-scale deployment.

This analysis highlights the significance of striking a balance between model performance and computational efficiency, such that hybrid models are still effective and feasible for a broad variety of solar energy forecasting applications.

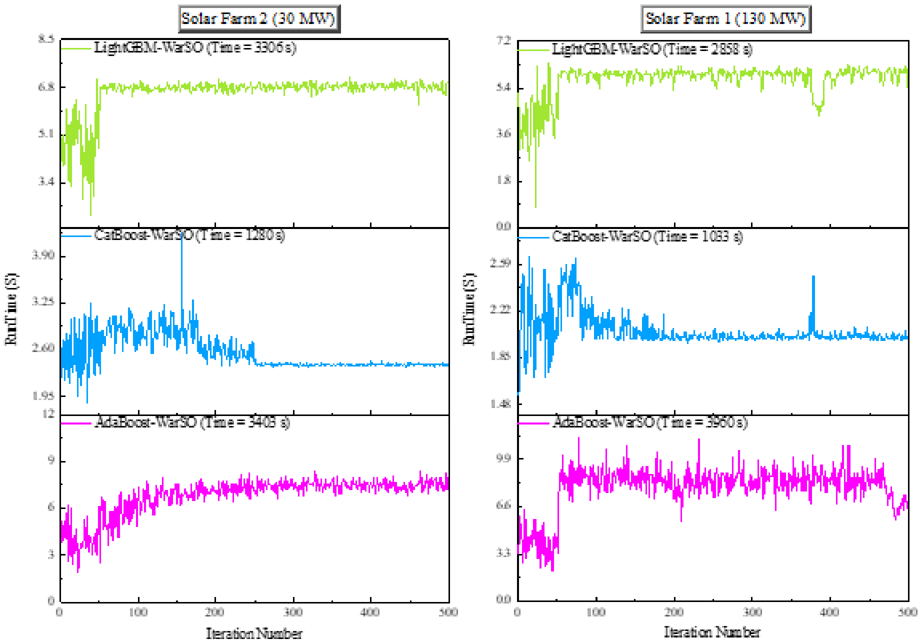

Figure 14: Comparison of runtime for various hybrid models in both Farm1 and Farm2

Fig 15 illustrates the convergence chart for the hybrid models, using the Mean Squared Error (MSE) index as the convergence metric with a set number of iterations at 300. Based on Figure 15, the values for the first farm exhibit higher MSE, whereas for the second farm, these values are lower. In the first farm, the hybrid AdaBoost-War SO model has the lowest MSE. Similarly, in the second farm, as expected, the hybrid AdaBoost-War SO model has the lowest MSE.

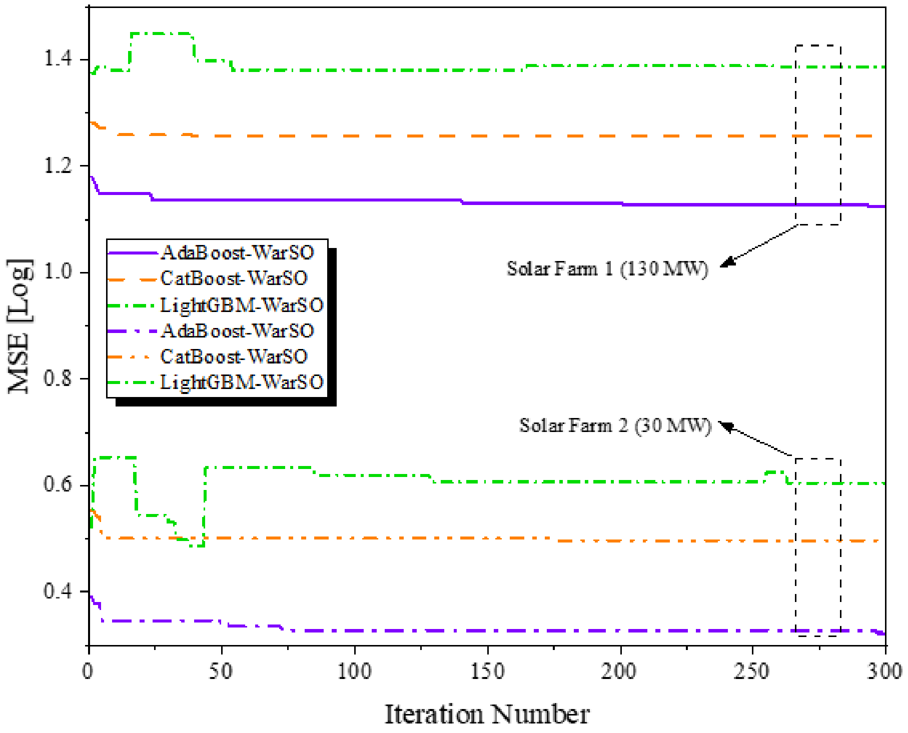

Figure 15: The convergence plots of the Cat Boost, AdaBoost, and Light GBM hybrid model

4 Discussion

The outcome of this study exhibits significant improvement in solar power prediction over the SOTA benchmarks due mainly to the incorporation of machine learning methods in addition to the War SO optimization algorithm. Out of the proposed methods, the AdaBoost-War SO model stood out, with $R^2 = 0.9836$ and root mean square error (RMSE) = 1.75 MW for the 30 MW solar power plant. This performance surpassed the performance of single algorithms like CatBoost ($R^2 = 0.9106$, RMSE = 4.06 MW) and even other combinations, i.e., CatBoost-War SO and LightGBM-War SO.

4.1 Reason for high performance of AdaBoost-War SO

The iterative boosting mechanism of AdaBoost allows it to correct errors yielded by weak learners, thereby allowing it to effectively model the nonlinear relationships that are inherent in solar energy data. The model flexibility remains supplemented by the War SO optimizer that optimally trades off exploration of new parameter spaces and exploitation of already known optimum solutions. Through this dual capability, the hybrid model is assured of converging to an improved global optimum than traditional optimization methods like grid search or particle swarm optimization (PSO).

The AdaBoost-War SO hybrid model that combined both of them possessed lower prediction variance, especially when testing, which means better generalization power. Although CatBoost and LightGBM can be used as individual models, their precision was restricted due to the absence of an external optimization platform for dynamic fine-tuning of the hyperparameters.

4.2 Importance of high-sensitivity parameters

The sensitivity analysis revealed that global horizontal irradiance, direct normal irradiance, and total solar irradiance were the most significant parameters used to predict solar energy production. These findings have very practical applications:

4.2.1 Improved feature selection

Models can reduce computational requirements and, simultaneously, increase accuracy by concentrating on the most sensitive parameters. By giving priority to these features, noise caused by unimportant variables is reduced.

4.2.2 Instant predictions

Continuous real-time monitoring of high-sensitivity parameters is essential for forecasting system operation. Improved sensors must favor quality solar irradiance measurement for input data for forecasting.

4.2.3 Site-specific calibration

The high sensitivity of irradiance parameters makes location-specific model calibration highly necessary, as irradiance behavior varies greatly with climate and geography. Region-specific models provide more accurate energy forecasting.

4.2.4 Supporting grid stability

Accurate prediction based on high-sensitivity parameters enhances the integration of solar power into the electricity grid. Reducing prediction errors allows grid managers to balance demand and supply, thus ensuring stability and avoiding outages.

4.2.5 Resource and risk management

Sensitivity analysis outputs may be used to inform resource planning, e.g., investing in high-end measurement technologies, and formulating risk reduction strategies for solar power plants. Understanding the major drivers of variability can allow for contingency planning to address outages due to weather.

4.3 Dataset and benchmark comparisons

Hybrid models, specifically the AdaBoost-War SO model, always performed better than single models and other hybrid combinations in accuracy metrics. For example, the AdaBoost-War SO model had an R^2 of 0.9836 and RMSE of 1.75 MW for the 30 MW solar farm and did better than the research of previous scholars such as Suanpang and Jamjuntr (2024), where LGBM had an R^2 of 0.84 and RMSE of 5.77 W. Likewise, in comparison with Singh et al. (2023), whose hybrid model using GRU enhanced accuracy for large systems, the current research demonstrated improved generalization on multi-site datasets.

The application of the War SO optimizer was significant in enhancing the performance of the AdaBoost-War SO model. Through its provision of a trade-off between exploration and exploitation, War SO allowed for efficient hyperparameter adjustment, thus avoiding local optima—a limitation that is usually faced with typical optimization methods like grid search or genetic algorithms used in state-of-the-art models. This further strengthened the capability of the hybrid models to efficiently capture the non-linear relationships in the data.

4.4 Optimization and computational efficiency

The second significant contribution of this study is its consideration of runtime and convergence. Even though the AdaBoost-War SO model was the most precise, its runtime was comparatively higher because both AdaBoost's iterative boosting and War SO's optimization are slow processes. However, this is warranted due to the substantial improvements in predictive accuracy and trustworthiness. Convergence analysis indicated that War SO significantly lowered the possibility of trapping in local optima, especially in high-dimensional parameter spaces, hence making it an apt option for hybrid model optimization for renewable energy forecasting.

4.5 Computational cost: Trade-Offs between accuracy and runtime

Increased accuracy of the hybrid models presented here, AdaBoost-War SO, comes with increased computational costs, thus a compromise between accuracy and computation time. The AdaBoost-War SO model was more accurate, achieving R^2 of 0.9836 and RMSE of 1.75 MW for the 30 MW farm but also had the highest processing time of approximately 3,960 seconds, as shown in Fig 14. Its high computational complexity is due to both the iterative approach of AdaBoost with training multiple weak learners and dynamic adjustment of their weights and the optimization approach of War SO that balances exploration and exploitation via successive iterations. While the enhanced accuracy significantly reduces prediction error and enables great generalization on diverse datasets, heightened runtime is a scalability problem in large-scale or real-time applications, for instance, energy network integration. Nevertheless, other hybrid models, for instance, CatBoost-War SO with an R^2 score of 0.9763 and considerably lower runtime, offer an acceptable trade-off and therefore are viable where computational efficiency matters. To mitigate the computational expense of AdaBoost-War SO, parallelization, distributed computing, and dynamic model selection can be used. Such methods can achieve a balance between accuracy and execution time, enabling hybrid models for specific needs in forecasting. Whereas AdaBoost-War SO is appropriate for applications where precision is paramount, more speedy options can be adequate for less resource-intensive applications, indicating a compromise between efficiency and performance.

4.6 Broader implications

The study brings into focus the potential of hybrid machine learning architectures augmented by innovative algorithms such as War SO. The accurate forecasting of solar energy generation, as a function of high sensitivity parameters and efficient optimization methods, is of particular importance to power grid reliability, resource planning, and power system integration of renewable energies. This work sets the new standard for predicting solar energy by overcoming key limitations in current best-practice methods, including low data resolution, sparse sensitivity testing, and the lack of hybrid optimization.

The fluctuation in the performance measures, i.e., RMSE and MAE, from training to test data indicates possible overfitting in certain of the models. For example, models such as CatBoost performed best during training (e.g., $R^2 = 0.608$, RMSE = 4.478 W, MAE = 3.367 W) but significantly declined during test ($R^2 = 0.46$, RMSE = 4.748 W, MAE = 3.583 W). This gap indicates that while the model was able to find patterns in the training set, it was struggling to generalize to novel, unseen data.

4.7 Comparative performance across farms

The performance of the models was extremely inconsistent between the 130 MW and 30 MW solar farms, showing the role of farm capacity and characteristics of data in model performance. For 130 MW, the models were subjected to higher variability of important parameters such as solar irradiance and temperature, seemingly due to the higher geographical spread of the farm. This greater exposure to more variable microclimatic conditions brought more noise into the data, and therefore it was more challenging to gain precise predictions. Conversely, the smaller 30 MW farm provided more consistent conditions, so there was less variance and the models could operate better.

Performance variations can also be accounted for by dataset-specific factors. The 130 MW dataset included higher variability in solar irradiance, which negatively affected the potential of the models to generalize well. The 30 MW farm dataset included more uniform patterns, and these translated into higher accuracy results for most of the models. These findings suggest that site-specific factors such as farm size, local weather, and dataset variability play important roles in the effectiveness of forecasting models.

To address these issues, model site-specific calibration is required. Normalization of the dataset per agricultural field aided the models in conforming to particular patterns with lesser effort; additional advances can be achieved by incorporating additional features, including wind speed and cloud cover, to better reflect environmental variation. In addition, the creation of hybrid approaches that combine localized tuning with generalized prediction capability promises improved scaling up of such models to farms of varied sizes and conditions.

This research places importance on how one should consider the specific nature of every farm while forecasting solar power and the need for subsequent research with models being tested under different geographic and operational conditions. These results add to the body of knowledge regarding solar farm capacity and dataset attributes and how they influence model performance and consequently enable the design of more precise and adaptive forecasting models.

4.8 Limitation

Overfitting is a result of many different causes, including model complexity that is too high, lack of diversity in the training data, or weak regularization. Combating it is important for maintaining the reliability and stability of forecast models in actual usage. Cross-validation, early stopping, and hyperparameter tuning are some of the methods that can reduce overfitting by avoiding the model from over-focusing on noise or irrelevant patterns in the training data.

Future research activities can explore the use of simpler models or hybrid approaches that balance predictive power with the ability to generalize. Expanding the dataset to cover a wider variety of diverse and representative samples, such as data from various geographic regions or seasonal differences, could also help increase model performance and reduce the danger of overfitting. In addition, the use of methods like dropout or L2 regularization in models such as CatBoost and LightGBM could potentially increase their generalizability to different datasets.

Through the elimination of such constraints, future research can make predictive models perform stably on training and testing datasets, thus encouraging their application in volatile and uncertain solar energy conditions.

PCC was able to capture significant features, i.e., solar irradiance, temperature, and humidity, but its focus on linear relationships could have overlooked non-linear relationships that would be useful for model performance. More sophisticated approaches, e.g., mutual information or machine learning model-based feature importance, would provide a more nuanced picture of feature importance, particularly for variables with complex interactions. Also, the exclusion of potentially important meteorological variables such as wind speed and cloud cover might have limited the model's ability to capture environmental heterogeneity to some degree. For example, solar irradiance is highly influenced by wind speed and cloud cover under specific conditions, which could potentially affect prediction under varying weather conditions. These shortcomings can be improved in future studies by including more variables and using strict imputation protocols, which would enable higher generalizability and predictive validity of the proposed models.

4.9 Comparison of solar energy forecasting models

Table 7 provides a concise comparison of key performance metrics across different studies, highlighting the effectiveness and computational considerations of various machine learning approaches in solar energy forecasting. Based on the comparison, the study method, which employs the AdaBoost-War SO hybrid model, demonstrates superior performance in solar energy forecasting.

Table 7: Comparison of solar energy forecasting models

Aspect	**Best Model R²**	**Best Model RMSE**
This study (AdaBoost-War SO)	0.9836	1.75 MW
Nguyen et al. (2025) (CatBoost)	0.608 (Training), 0.46 (Testing)	4.478 W (Training), 4.748 W (Testing)
Suanpang and Jamjuntr (2024) (LightGBM)	0.84	5.77

5 Conclusion

This study demonstrates the ability of hybrid machine learning models, optimized by the War SO algorithm, to improve solar power prediction accuracy. Using high-resolution data that was recorded every 15 minutes and advanced feature selection techniques, such as the Delta Moment Independent Measure (DMIM), the models achieved improved performance compared to their separate models. The recognition of solar irradiance as the largest contributing factor aligns with earlier research; yet, using DMIM in this study is a more rigorous sensitivity analysis, therefore further enriching knowledge on its effects on energy output.

The findings indicate the promising prospect of improving the accuracy of forecasts, but broader implications for large-scale renewables integration and grid stability necessitate further investigation. The findings have specific significance to standalone energy management use cases, such as solar energy installation operation optimization and policy guidance for storage and grid balancing. Future studies should extend on these findings through experiments with testing the models across different geographical and climatic locations and assessing their implementation in actual-time grid management systems.

This study deals with the critical issues of solar forecasting, thus making renewable energy systems more efficient and reliable. Nevertheless, the findings are presented cautiously under the limitations of the study, outlining the short-term practical applications and laying the ground for further development in the field of renewable energy forecasting.

Abbreviation

AdaBoost	Adaptive Boosting	NWS	National Weather Service
ANN	Artificial Neural Network	P	Prior value
ARIMA	Autoregressive integrated moving average	PI	Prediction intervals
Bt	The normalization factor	PV	Photovoltaic
C	Commander	R2	Coefficient of Determination
Cat Boost	Categorical Gradient Boosting	Rai	The ranking
DL	Deep Learning	RAE	Relative Absolute Error
Dt(i)	Uniform sample distribution	RES	Renewable energy sources
et	Average error	RF	Random forest
Fper	The previous situation	RMSE	Root means square error
Fnew	The new situation	SVM	Support vector machines
ft(xi)	A weak predictor		
GBT	Gradient boosting tree	SVR	Support Vector Regression
JSD	Jensen Shannon Divergence	VAF	Variance Accounted For
K	The King	War SO	War Strategy Optimizer
LACE	Levelized Avoided Cost of Electricity	α	The corresponding weight
LCOE	Levelized Cost of Electricity	xk	The random vector
Light GBM	Light Gradient Boosting Machine	yK and yC	The situations of the King and The Commander
LR	Linear regression		
MADSR	Monthly average daily solar radiation		
MAE	Mean Absolute Error		
ML	Machine Learning		
NWP	Numerical weather prediction		

Acknowledgements

I would like to take this opportunity to acknowledge that there are no individuals or organizations that require acknowledgment for their contributions to this work.

Competing interests

The authors declare no competing interests.

Authorship contribution statement

Fenghong Pan: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Data availability

Data can be shared upon request.

Declarations

Not applicable.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author statement

All the authors have read and approved the manuscript. As stated earlier in this document, the requirements for authorship have been met, and each author believes that the manuscript represents honest work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

[1] S. Koohi-Fayegh, M.A. Rosen, A review of renewable energy options, applications, facilitating technologies and recent developments, European Journal of Sustainable Development Research 4 (2020) em0138.

[2] K.S. Perera, Z. Aung, W.L. Woon, Machine learning techniques for supporting renewable energy generation and integration: a survey, in: Data Analytics for Renewable Energy Integration: Second ECML PKDD Workshop, DARE 2014, Nancy, France, September 19, 2014, Revised Selected Papers 2, Springer, 2014: pp. 81–96.

[3] D. Gielen, F. Boshell, D. Saygin, M.D. Bazilian, N. Wagner, R. Gorini, The role of renewable energy in the global energy transformation, Energy Strategy Reviews 24 (2019) 38–50.

[4] W. Strielkowski, L. Civín, E. Tarkhanova, M. Tvaronavičienė, Y. Petrenko, Renewable energy in the sustainable development of electrical power sector: A review, Energies (Basel) 14 (2021) 8240.

[5] G.A. Tiruye, A.T. Besha, Y.S. Mekonnen, N.E. Benti, G.A. Gebreslase, R.A. Tufa, Opportunities and challenges of renewable energy production in Ethiopia, Sustainability 13 (2021) 10381.

[6] N.E. Benti, T.A. Woldegiyorgis, C.A. Geffe, G.S. Gurmesa, M.D. Chaka, Y.S. Mekonnen, Overview of geothermal resources utilization in Ethiopia: Potentials, opportunities, and challenges, Sci Afr 19 (2023) e01562.

[7] N.E. Benti, A.B. Aneseyee, C.A. Geffe, T.A. Woldegiyorgis, G.S. Gurmesa, M. Bibiso, A.A. Asfaw, A.W. Milki, Y.S. Mekonnen, Biodiesel production in Ethiopia: Current status and future prospects, Sci Afr 19 (2023) e01531.

[8] N.E. Benti, Y.S. Mekonnen, A.A. Asfaw, combining green energy technologies to electrify rural community of Wollega, Western Ethiopia, Sci Afr 19 (2023) e01467.

[9] C.R. Kumar, M.A. Majid, Renewable energy for sustainable development in India: Current status, future prospects, challenges, employment, and investment opportunities, TIDEE: TERI Information Digest on Energy and Environment 21 (2022) 33.

[10] P. Denholm, D.J. Arent, S.F. Baldwin, D.E. Bilello, G.L. Brinkman, J.M. Cochran, W.J. Cole, B. Frew, V. Gevorgian, J. Heeter, The challenges of achieving a 100% renewable electricity system in the United States, Joule 5 (2021) 1331–1352.

[11] E. Alhamer, A. Grigsby, R. Mulford, The Influence of Seasonal Cloud Cover, Ambient Temperature and Seasonal Variations in Daylight Hours on the Optimal PV Panel Tilt Angle in the United States, Energies (Basel) 15 (2022) 7516.

[12] S. Impram, S.V. Nese, B. Oral, Challenges of renewable energy penetration on power system flexibility: A survey, Energy Strategy Reviews 31 (2020) 100539.

[13] I. Ghalehkhondabi, E. Ardjmand, G.R. Weckman, W.A. Young, An overview of energy demand forecasting methods published in 2005–2015, Energy Systems 8 (2017) 411–447.

[14] A. Krechowicz, M. Krechowicz, K. Poczeta, Machine learning approaches to predict electricity production from renewable energy sources, Energies (Basel) 15 (2022) 9146.

[15] Y.-Y. Hong, T.R.A. Satriani, Day-ahead spatiotemporal wind speed forecasting using robust design-based deep learning neural network, Energy 209 (2020) 118441.

[16] X. Zhao, J. Liu, D. Yu, J. Chang, One-day-ahead probabilistic wind speed forecast based on optimized numerical weather prediction data, Energy Convers Manag 164 (2018) 560–569.

[17] J. Fan, L. Wu, F. Zhang, H. Cai, W. Zeng, X. Wang, H. Zou, Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China, Renewable and Sustainable Energy Reviews 100 (2019) 186–212.

[18] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renew Energy 105 (2017) 569–582.

[19] P. Suanpang, P. Jamjuntr, Machine learning models for solar power generation forecasting in microgrid application implications for smart cities, Sustainability 16 (2024) 6087.

[20] S. Singh, V. Subburaj, K. Sivakumar, R. Anil Kumar, M.S. Muthuramam, R. Rastogi, V. Ratansing Patil, A. Rajaram, Optimum Power Forecasting Technique for Hybrid Renewable Energy Systems Using Deep Learning, Electric Power Components and Systems (2024) 1–18.

[21] H.N. Nguyen, Q.T. Tran, C.T. Ngo, D.D. Nguyen, V.Q. Tran, Solar energy prediction through machine learning models: A comparative analysis of

regressor algorithms, PLoS One 20 (2025) e0315955.

[22] C. Zhu, M. Wang, M. Guo, J. Deng, Q. Du, W. Wei, Y. Zhang, Innovative approaches to solar energy forecasting: unveiling the power of hybrid models and machine learning algorithms for photovoltaic power optimization, J Supercomput 81 (2025) 20.

[23] J. Huertas-Tato, R. Aler, I.M. Galván, F.J. Rodríguez-Benítez, C. Arbizu-Barrena, D. Pozo-Vázquez, A short-term solar radiation forecasting system for the Iberian Peninsula. Part 2: Model blending approaches based on machine learning, Solar Energy 195 (2020) 685–696.

[24] A.E. Gürel, Ü. Ağbulut, Y. Biçen, Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation, J Clean Prod 277 (2020) 122353.

[25] M. Alizamir, S. Kim, O. Kisi, M. Zounemat-Kermani, A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions, Energy 197 (2020) 117239.

[26] C. Koo, W. Li, S.H. Cha, S. Zhang, A novel estimation approach for the solar radiation potential with its complex spatial pattern via machine-learning techniques, Renew Energy 133 (2019) 575–592.

[27] N.C. Nath, W. Sae-Tang, C. Pirak, Machine learning-based solar power energy forecasting, Journal of the Society of Automotive Engineers Malaysia 4 (2020) 307–322.

[28] D.S. Kumar, W. Teo, N. Koh, A. Sharma, W.L. Woo, A Machine Learning Framework for Prediction Interval based Technique for Short-Term Solar Energy Forecast, in: 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), IEEE, 2020: pp. 406–409.

[29] I. Jebli, F.-Z. Belouadha, M.I. Kabbaj, A. Tilioua, Prediction of solar energy guided by pearson correlation using machine learning, Energy 224 (2021) 120109.

[30] L. Abualigah, R.A. Zitar, K.H. Almotairi, A.M. Hussein, M. Abd Elaziz, M.R. Nikoo, A.H. Gandomi, Wind, solar, and photovoltaic renewable energy systems with and without energy storage optimization: A survey of advanced machine learning and deep learning techniques, Energies (Basel) 15 (2022) 578.

[31] J. Huertas-Tato, R. Aler, I.M. Galván, F.J. Rodríguez-Benítez, C. Arbizu-Barrena, D. Pozo-Vázquez, A short-term solar radiation forecasting system for the Iberian Peninsula. Part 2: Model blending approaches based on machine learning, Solar Energy 195 (2020) 685–696.

[32] M. Alizamir, S. Kim, O. Kisi, M. Zounemat-Kermani, A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions, Energy 197 (2020) 117239.

[33] Y. Chen, J. Xu, Solar and wind power data from the Chinese state grid renewable energy generation forecasting competition, Sci Data 9 (2022) 577.

[34] M.A. Oladipupo, P.C. Obuzor, B.J. Bamgbade, A.E. Adeniyi, K.M. Olagunju, S.A. Ajagbe, An automated python script for data cleaning and labeling using machine learning technique, Informatica 47 (2023).

[35] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, ArXiv Preprint ArXiv:1810.11363 (2018).

[36] J. Fan, X. Wang, F. Zhang, X. Ma, L. Wu, predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data, J Clean Prod 248 (2020) 119264.

[37] E.K. Ampomah, Z. Qin, G. Nyame, F.E. Botchey, Stock market decision support modeling with tree-based AdaBoost ensemble machine learning models, Informatica 44 (2021).

[38] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J Comput Syst Sci 55 (1997) 119–139.

[39] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Adv Neural Inf Process Syst 30 (2017).

[40] T.S.L. V Ayyarao, N.S.S. Ramakrishna, R.M. Elavarasan, N. Polumahanthi, M. Rambabu, G. Saini, B. Khan, B. Alatas, War strategy optimization algorithm: a new effective metaheuristic algorithm for global optimization, IEEE Access 10 (2022) 25073–25105.

[41] H. Hu, S. Gong, B. Taheri, Energy demand forecasting using convolutional neural network and modified war strategy optimization algorithm, Heliyon (2024).

[42] H. Khajavi, A. Rastgoo, Improving the prediction of heating energy consumed at residential buildings using a combination of support vector regression and meta-heuristic algorithms, Energy 272 (2023) 127069.

[43] C. Pasion, T. Wagner, C. Koschnick, S. Schuldt, J. Williams, K. Hallinan, Machine learning modeling of horizontal photovoltaics using weather and location data, Energies (Basel) 13 (2020) 2570.