AIMS Energy, 12(2): 350–386.
DOI: 10.3934/energy.2024017
Received: 02 January 2024
Revised: 08 February 2024
Accepted: 20 February 2024
Published: 01 March 2024

http://www.aimspress.com/journal/energy

Research article

Advancing solar energy forecasting with modified ANN and light GBM learning algorithms

Muhammad Farhan Hanif[1,3,*], Muhammad Sabir Naveed[1], Mohamed Metwaly[2], Jicang Si[1], Xiangtao Liu[1] and Jianchun Mi[1]

[1] Department of Energy & Resource Engineering, College of Engineering, Peking University, Beijing 100871, China
[2] Archaeology Department, College of Tourism and Archaeology, King Saud University, P.O. Box 2627, 12372 Riyadh, Saudi Arabia
[3] Department of Mechanical Engineering, FE&T, Bahauddin Zakariya University Multan-60000, Pakistan

* **Correspondence:** Email: farhanhanif@pku.edu.cn; Tel: +923467148083.

Abstract: In the evolving field of solar energy, precise forecasting of Solar Irradiance (SI) stands as a pivotal challenge for the optimization of photovoltaic (PV) systems. Addressing the inadequacies in current forecasting techniques, we introduced advanced machine learning models, namely the Rectified Linear Unit Activation with Adaptive Moment Estimation Neural Network (RELAD-ANN) and the Linear Support Vector Machine with Individual Parameter Features (LSIPF). These models broke new ground by striking an unprecedented balance between computational efficiency and predictive accuracy, specifically engineered to overcome common pitfalls such as overfitting and data inconsistency. The RELAD-ANN model, with its multi-layer architecture, sets a new standard in detecting the nuanced dynamics between SI and meteorological variables. By integrating sophisticated regression methods like Support Vector Regression (SVR) and Lightweight Gradient Boosting Machines (Light GBM), our results illuminated the intricate relationship between SI and its influencing factors, marking a novel contribution to the domain of solar energy forecasting. With an R^2 of 0.935, MAE of 8.20, and MAPE of 3.48%, the model outshone other models, signifying its potential for accurate and reliable SI forecasting, when compared with existing models like Multi-Layer Perceptron, Long Short-Term Memory (LSTM), Multilayer-LSTM, Gated Recurrent Unit, and 1-dimensional Convolutional Neural Network, while the LSIPF model showed limitations in its predictive ability. Light GBM emerged as a robust approach in evaluating environmental influences on SI, outperforming

the SVR model. Our findings contributed significantly to the optimization of solar energy systems and could be applied globally, offering a promising direction for renewable energy management and real-time forecasting.

Keywords: Artificial Neural Network (ANN); Support Vector Machine (SVM); Lightweight Gradient Boosting Machines (Light GBM); machine learning; Solar Irradiance (SI); solar forecasting

Acronyms: SI: Solar irradiance; PV: Photovoltaic; RELAD-ANN: Rectified linear unit activation with adaptive moment estimation neural network; LSIPF: Linear Support Vector Machine with Individual Parameter Features; SVR: Support Vector Regression; Light GBM: Lightweight gradient boosting machines; ANN: Artificial neural network; SVM: Support vector machine; MLP: Multi-Layer perceptron; LSTM: Long short-term memory; MLSTM: Multilayer long short-term memory; GRU: Gated recurrent unit; CNN: Convolutional neural network; GHI: Global horizontal irradiance; DNI: Direct normal irradiance; CSP: Concentrator solar power; RNN: Recurrent neural network; MSE: Mean squared error; MAE: Mean absolute error; MAPE: Mean absolute percentage error; R^2: Coefficient of determination

1. Introduction

The imperative for renewable energy sources has escalated in response to climate change and the depletion of fossil fuels, with the energy sector contributing significantly to global greenhouse gas emissions [1]. Amidst this transition, solar energy has been recognized as a pivotal solution, given its potential to meet growing energy demands sustainably [2–6]. The adoption of solar power varies globally, with developed nations like the UK and Germany integrating it extensively into their energy infrastructures, while developing countries face multifaceted barriers to its utilization [7–9].

Central to enhancing solar energy adoption is the improvement of SI forecasting, which plays a critical role in the reliable integration of solar power into energy systems. Accurate forecasting methods are vital for planning and optimizing solar energy deployment across different regions, each facing unique geographical and climatic challenges [10]. Forecasting methodologies span from immediate to long-term predictions, employing a range of approaches including physical, statistical, and machine learning models [11–13].

Physical models prioritize environmental data and specific characteristics of solar power plants to predict SI [14], such as global horizontal irradiance (GHI) and direct normal irradiance (DNI). These models are directly affected by the accuracy of meteorological inputs, with their reliability decreasing with imprecise weather data [15–17]. Statistical methods, on the other hand, utilize historical irradiance data to identify patterns, offering insights into solar power production despite requiring extensive site-specific data which may limit their flexibility [18–21].

Machine learning models have emerged as especially promising for SI forecasting due to their ability to process complex, large-scale datasets [22]. These models, including decision trees [23], random forests [24], and support vector machines [25], adapt to diverse forecasting scenarios, from short-term operational planning to long-term strategic development. Artificial Neural Networks (ANNs) are also utilized to forecast energy production in concentrator solar power (CSP) systems, employing different learning algorithms [26]. Recurrent neural networks (RNNs) with adaptive neural imputation modules enhance SI prediction accuracy, especially when data is incomplete [27]. ANNs combined with ripple

current correlation techniques have shown to efficiently optimize photovoltaic system performance [28]. ANNs are also utilized to forecast energy production in CSP systems, employing different learning algorithms [29]. Their advanced algorithms like LSTM [30], Gated Recurrent Unit (GRU) [31] and Convolutional Networks (CNN) [32] are capable of handling the variability and unpredictability inherent in SI, making them a cornerstone for enhancing the precision of solar energy forecasts [33]. Neural network-based metaheuristic algorithms optimize financial market analysis, improving investment decision-making [34]. Furthermore, optimizing machine parameters through a fuzzy possibility regression integrated model and an adaptive-network-based fuzzy inference system has enhanced product quality in manufacturing [35]. For environmental protection and disaster management, a decision support system leveraging machine learning computations aids in precipitation prediction in Iran [36].

Hybrid models that combine machine learning techniques with temporal and spatial data analysis further refine forecasting accuracy [37]. For instance, ConvLSTM models [38,39] integrate convolutional and LSTM units to analyze both spatial and temporal aspects of SI. The development of models like WaveNet [40] and Temporal Convolutional Network (TCN) [41], which employ causal dilated convolutions, demonstrates the advancement in temporal data processing, which is essential for accurate SI predictions [42]. Hybrid deep learning methods also assess low-carbon transportation development, identifying key factors and their interconnections [43]. Disdain their potential, these sophisticated models face challenges such as the need for extensive training data and the complexity of their architectures [44,45].

While there have been considerable strides in the domain of renewable energy forecasting, specifically SI prediction, a notable research gap persists in the integration of advanced artificial intelligence (AI) and machine learning techniques into these efforts. Current models often grapple with balancing computational efficiency against the backdrop of the inherently variable and complex solar energy data. A notable deficiency lies in the integration of cutting-edge AI methods with a deep understanding of the intricate relationship between SI and various meteorological factors. The reliance on conventional statistical or elementary machine learning algorithms by existing studies overlooks the potential that more sophisticated AI approaches could offer [46–50]. Challenges such as overfitting, data inconsistency, and insufficient analysis that links SI forecasting with crucial environmental variables (e.g., air temperature, wind speed, humidity) exacerbate these gaps. Furthermore, the testing and applicability of these models across diverse geographical landscapes remain limited, underscoring a crucial need for models that can be scaled globally for renewable energy adoption.

Addressing these gaps, we introduce the RELAD-ANN and LSIPF models as pioneering contributions to SI prediction. The RELAD-ANN model, with its multi-layer architecture and the innovative application of the Rectified Linear Unit (ReLU) activation function, represents a significant leap forward. It meticulously navigates the complexities of SI data, striking a remarkable balance between computational efficiency and forecasting accuracy, facilitated by advanced data preprocessing and AI techniques. The LSIPF model complements this by adopting a sophisticated feature selection and data representation strategy, showcasing the potential of leveraging established machine learning algorithms in novel, contextually adapted manners. Our work highlights a dedicated effort to enhance SI forecasting accuracy through the integration of sophisticated AI methodologies.

Furthermore, we provide an in-depth exploration of the effects of meteorological parameters on SI, utilizing advanced regression techniques such as SVR and Light GBM. This comprehensive analysis offers new insights into how variables like wind speed, air temperature, and humidity impact SI predictions, contributing valuable knowledge to the field of renewable energy forecasting and offering actionable guidance for the optimization of solar energy systems.

The innovation of our research lies in its comprehensive approach, which merges advanced AI modeling, thorough data analysis, and a nuanced understanding of meteorological dynamics. This integrated perspective ensures our research is positioned at the cutting edge of renewable energy forecasting innovation. We anticipate our contributions will significantly influence the development of more efficient and sustainable solar energy solutions, benefiting not only specific regions like Quetta, Pakistan but also enhancing the applicability and scalability of solar energy forecasting globally.

In essence, we aspire to:

- Elucidate the design and implementation of the RELAD-ANN and LSIPF models, specifically tailored to address the intricacies inherent in SI forecasting.
- Explicate the correlations between various parameters and SI by leveraging the potential of SVR and Light GBM.
- Empirically validate the proposed models against robust statistical benchmarks, affirming their viability for broader applications.
- Compare our models with five state-of-the-art models proposed by other researchers to validate their accuracy, demonstrating the competitive edge of our methodologies.
- Undertake comprehensive Ablation Studies, selecting five diverse cities as case studies to assess the performance of our models in varied geographical and climatic conditions, alongside performing parametric analyses to explore the impact of additional parameters on our forecasting accuracy.

2. Materials and methods

2.1. Selection of location and parameters

Selecting an optimal location for a solar power facility is crucial, given the significant initial investment and the facility's expected lifespan of 25 to 30 years. Identifying the optimal location is crucial, taking into account a myriad of criteria, including solar energy potential, duration of sunshine, solar radiation, and data accessibility for different parameters essential for deploying AI techniques.

Pakistan's energy situation is critical, consuming only 0.37% of global energy, leading to economic challenges [51]. A heavy reliance on non-renewable energy exacerbates this [52,53], while renewable energy, including solar, contributes a mere 0.3% [54]. Figure 1 illustrates the total energy supply across various fuel types from 1990 to 2020 of Pakistan [55]. Focusing on the "Wind, solar, etc." category, there is a noticeable growth in solar energy contribution over the three decades. Starting with a modest presence in the early 1990s, solar energy sees a significant rise by 2020, highlighting its increasing adoption in the energy sector [56]. Solar energy's rise since the 1990s is notable, but by 2030, the expected energy demand of 40,000 MW will be predominantly met by non-renewables (67%) [57,58]. Despite abundant solar potential, particularly in regions like Quetta with 5–7 kWh/m^2 daily solar radiation [59], its utilization is low. Enhancing solar energy use through improved forecasting and policy could diversify energy sources and aid economic growth [60,61].

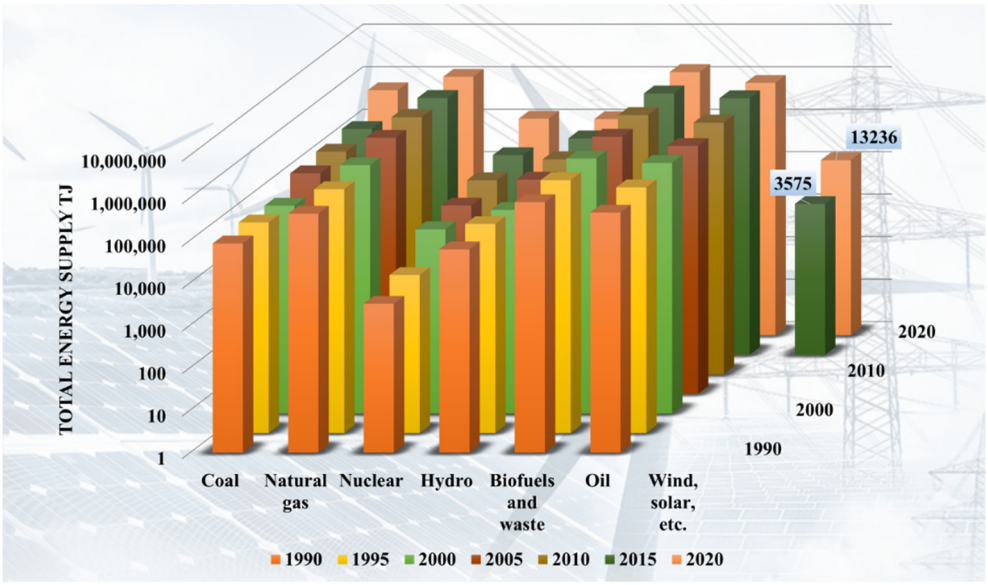

Figure 1. Total energy supply (TES) by source, Pakistan 1990–2020 [56].

In line with these criteria and parameters, we zero in on Quetta (Baluchistan, Pakistan) for its investigation. Situated at coordinates 30.195768° N, 67.017245° E, and elevated at 1586 m, Quetta is depicted in Figure 2(a) as part of the nation's GHI map [62]. Characterized by summers spanning from late-May to early-September with an average temperature hovering around 25 ℃, the city's winter, stretching from late-November to late-March, witnesses a temperature averaging at 5 ℃. The transitional seasons, autumn (from late-September to mid-November) and spring (from early-April to late-May), experience average temperatures of 16 ℃ and 15 ℃, respectively [63]. With an average humidity of 45% and wind speeds around 13 kph, Quetta stands out primarily due to two reasons: Its consistent sunshine, averaging between 8 to 8.5 hours daily, coupled with an annual DNI averaging at 2309.8 kwh/m^2. Additionally, the city boasts ample land availability for prospective solar initiatives. Such developments not only stand to augment the nation's energy supply, bolstering its economic standing, but also promise accelerated regional growth. Figure 2(b) showcases the city's monthly solar irradiations spanning 2005 to 2020 [64].This graph elucidates monthly GHI and DNI metrics between 2005 and 2020, highlighting distinct seasonal irradiation variations.

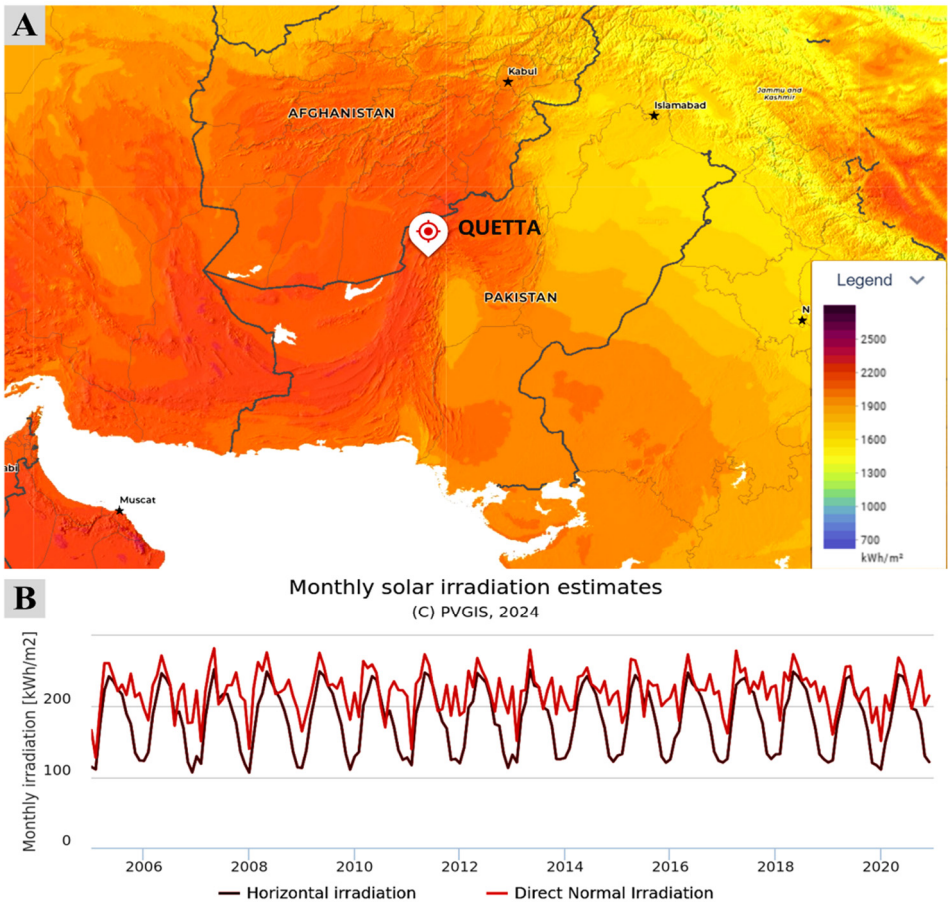

Figure 2. (a) GHI map of Pakistan showing the selected location "Quetta" [62] and (b) Monthly average radiations of Quetta from 2005 to 2020 showing GHI and DNI [64].

Central to this research are four specific parameters: SI (W/m^2), surface specific humidity (dimensionless), air temperature (°C), and wind speed (kph). The requisite hourly data, spanning August 01, 2019 00:00:00 to January 30, 2021 23:00:00, sourced from National Aeronautics and Space Administration (NASA), Giovanni [65], a web application developed by the NASA Goddard Earth Science Data and Information Services Center (GES DISC), serves as a Distributed Active Archive Center (DAAC) tool. It offers a straightforward and user-friendly platform for visualizing, analyzing, and accessing Earth science remote sensing data, primarily from satellites. With Giovanni, users can work with this data directly online without the need for downloading, streamlining research and study processes involving Earth's atmospheric, oceanic, and land data.

2.2. Data pre-processing

In the pursuit of refining AI models, supplying them with rigorous parametric data for training is paramount. Our exhaustive dataset is composed of 13,176 entries, which we judiciously bifurcated into a training set and a testing set in a 70:30 ratio. The training set envelops 9,223 entries for each parameter, spanning the time frame from August 01, 2019, 00:00:00 to August 18, 2020, 21:00:00. Conversely, the testing set encapsulates the remaining 30% of the data, amounting to 3,952 entries for each parameter, covering the period from August 19, 2020, 00:00:00 to January 30, 2021, 23:00:00.

In our exploration of SI forecasting, the integrity and depth of our dataset stand as the cornerstone of our research. Recognizing the paramount importance of data quality, we have adopted a meticulous approach to data handling and preprocessing to ensure the highest degree of model accuracy and reliability. Our methodology begins with importing a comprehensive dataset that captures a wide array of meteorological variables, including air temperature, surface humidity, radiance intensity, and wind speed, collected over time. After this, the data is processed through Google Collaboratory [66], which is a platform that enables you to write and execute Python in your browser with no setup required, free access to GPUs, and easy sharing capabilities. It is designed for students, data scientists, and AI researchers alike.

Upon importing the dataset, we performed an initial analysis to understand its characteristics and identify any challenges, such as missing values [67] or outliers [68] that could potentially skew our predictions. To enhance the dataset's robustness and ensure that our models operate on the most reliable information, we employed the Imputer from sklearn [69], utilizing a median strategy for imputing missing values [70]. This step was crucial in maintaining the integrity of our dataset, allowing us to preserve the original distribution of our data while filling in gaps that might otherwise introduce bias into our forecasts. Further, recognizing the impact of variables with high variability on our model's performance, we made the strategic decision to remove the 'surface incoming short-wave flux' from our dataset. This decision was informed by its high standard deviation [71] observed during our exploratory data analysis, indicating that it could detract from the predictive accuracy of our models. To prepare our dataset for the modeling process, we applied the StandardScaler from sklearn [72], normalizing our features to ensure that they are on the same scale [73]. This step is critical in machine learning modeling, as it prevents features with larger scales from dominating the model's learning process, thereby ensuring that each variable contributes appropriately to the prediction.

This careful and deliberate approach to data preprocessing—encompassing data cleaning, imputation, and scaling—lays a solid foundation for our subsequent modeling efforts. It not only enhances the predictive performance of our RELAD-ANN and LSIPF models but also underscores our commitment to methodological rigor and precision in tackling the complexities of SI forecasting. Through this rigorous process, we aim to contribute meaningful insights and robust predictive models to the field of renewable energy forecasting, advancing our understanding of SI dynamics and their implications for solar energy utilization.

Our rigorous approach to data pre-processing, as evidenced in Figure 3, underlines our commitment to methodological precision and reinforces the foundational integrity of our predictive analysis. This meticulous data treatment not only enriches the dataset but also ensures that the models developed are based on the most reliable and comprehensive information available.

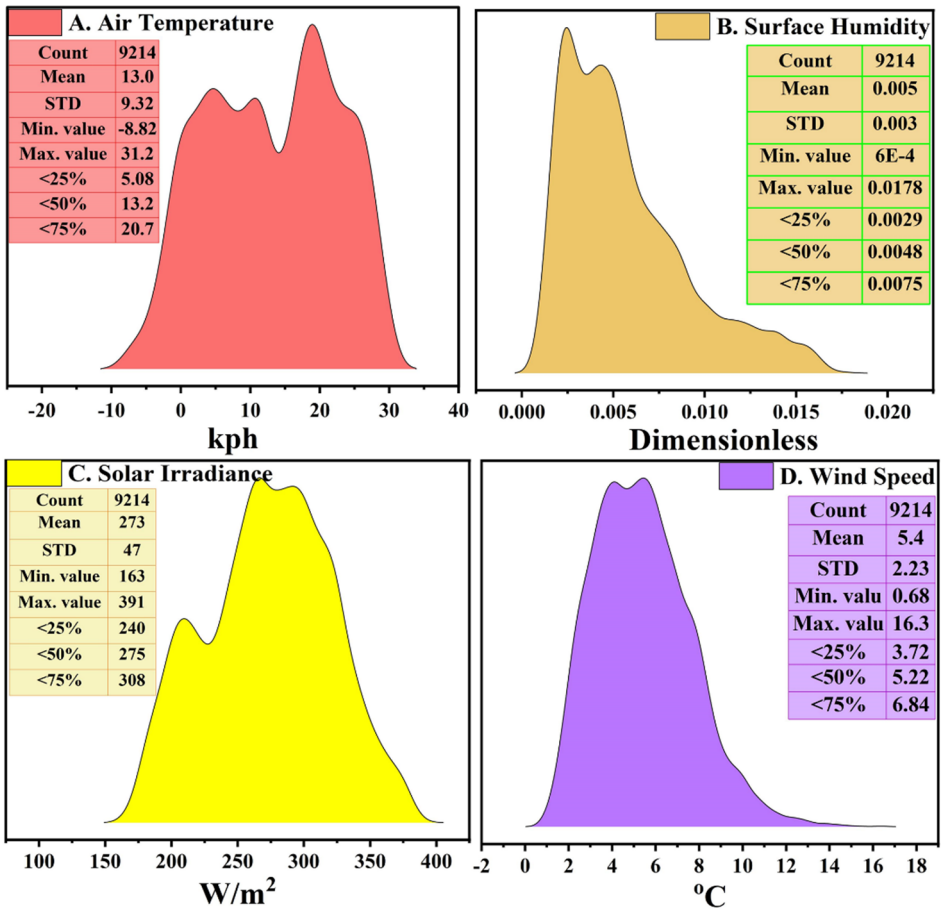

Figure 3. Probability density distribution graphs: (a) air temperature, (b) surface humidity, (c) SI, and (d) wind speed.

The air temperature, illustrated in Figure 3(a), exhibits pronounced variability with a mean of 13 ℃. This is evidenced by the broad spread of data points, ranging from a chilly –8.8 ℃ to a warm 31.2 ℃. Such variability accentuates the dynamic nature of temperature readings over the observation period. Contrarily, the surface humidity depicted in Figure 3(b) showcases a more consistent profile. Given a mean value of 0.5 (dimensionless) and a range from 0.06 to 2, the majority of the data points appear to converge near the mean, indicating limited variability. Wind speed and SI, represented in Figure 3(d) and Figure 3(c), respectively, both present moderate variabilities. The wind speed has an average reading of 5.4 kph with values fluctuating between a mild 0.7 kph to a brisk 16.4 kph. Moreover, SI, with its mean settled at 273 W/m^2, offers a range between 163 W/m^2 and 391 W/m^2, highlighting the periodic fluctuations in solar exposure. It is noteworthy that the mean values for all these parameters closely approximate their respective medians. This alignment further attests to the overall stability and reliability of the dataset, ensuring its robustness for subsequent analyses and applications.

The correlation plot presents (Figure 4) a matrix that quantitatively assesses the linear relationships between several meteorological variables, with a particular emphasis on their relevance to SI forecasting. The plot reveals a strong positive correlation (0.82) between air temperature [74] and SI is observed, which is intuitively logical as SI is a significant contributor to air temperature. This substantial correlation suggests that as air temperature increases, SI tends to increase as well, likely due to the direct heating effect of sunlight. There is a moderately strong positive correlation (0.79) between surface humidity [75] and SI. This could be interpreted as higher humidity levels being

associated with higher SI, potentially because areas with high humidity might also be regions where SI is strong, particularly in tropical and subtropical climates.

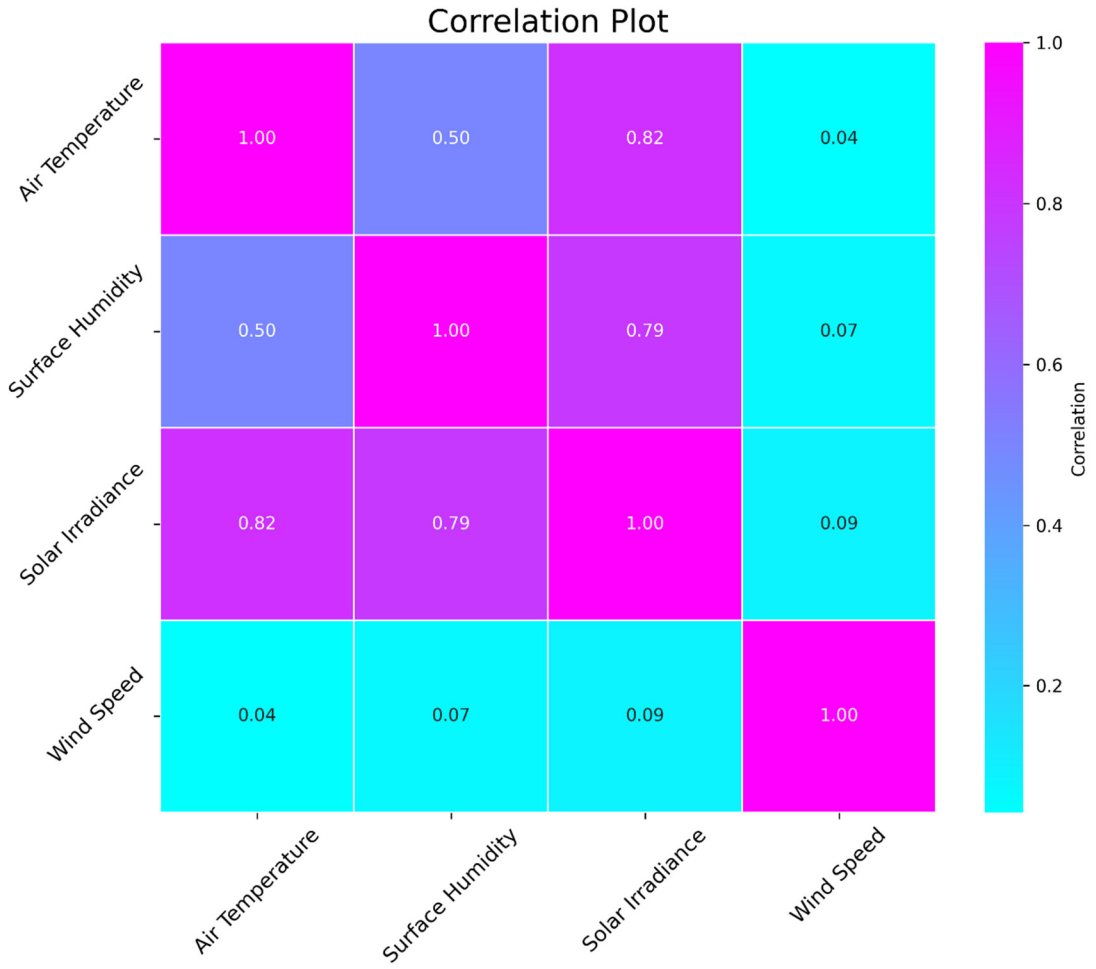

Figure 4. Correlation plot of parameters.

A positive but weaker correlation (0.09) is noted between wind speed and SI. While less pronounced, this relationship is important [29]. It can be reasoned that wind speed, by influencing the dispersal of cloud cover and humidity, can affect the amount of solar radiation that reaches the Earth's surface. Higher wind speeds might be associated with clearer skies, which would facilitate greater SI [76]. The correlation coefficient of 0.04 between wind speed and air temperature is indeed minor, yet it aligns with the physical principle of convection. As air temperature increases, it can induce wind due to the rising of warmer air and the subsequent movement of cooler air to replace it. The correlations involving wind speed, though weaker, do not diminish its importance in the context of SI forecasting. In fact, wind speed is a dynamic factor that can influence atmospheric conditions in a manner that is crucial for accurate SI prediction. For instance, wind speed affects the distribution of aerosols and cloud cover, both of which are critical for determining the amount of sunlight that penetrates the atmosphere.

In summary, while the correlation coefficients provide a snapshot of the relationships at play, they must be contextualized within the broader physical dynamics of the atmosphere. The interplay between air temperature, surface humidity, and wind speed is complex, and these variables collectively

contribute to the nuanced prediction of SI. Their combined effect on SI underscores the necessity of incorporating a multi-variable approach in SI forecasting models to capture the comprehensive picture of the factors influencing solar energy potential.

2.3. Model development for parametric forecasting

In this investigation, both the RELAD-ANN and LSIPF supervised machine learning models are harnessed for the prediction of SI as well as other paramount environmental parameters, specifically air temperature, surface humidity, and wind speed. The predictive acumen of these models is intrinsically contingent upon the quality and comprehensiveness of the training datasets.

The research, centered on Quetta, Pakistan, showcases methodologies with global application potential in SI forecasting. We introduce two distinct models: The RELAD-ANN, featuring ReLU activation and the ADAM optimizer, and the LSIPF, a Linear SVM with Individual Parameter Features. The LSIPF model is used for comparison with the RELAD-ANN, highlighting their respective strengths in SI prediction. Both models are designed for hourly forecasting of SI and its interaction with other environmental variables [77–79]. These models are not only a testament to scientific advancement in this domain but also a step beyond conventional machine learning methodologies. The models, particularly the novel RELAD-ANN approach, add a new dimension to the field of SI pattern recognition. Rooted in the empirical context of Quetta, they are poised to set new benchmarks in SI prediction, yet their fundamental scientific principles allow for broader application. They are precisely engineered to be fine-tuned to the specific climatic conditions of different regions, demonstrating their adaptability

To architect and calibrate these models, the Python programming language is selected, given its established aptitude for grappling with intricate big data quandaries [80]. Throughout the model formulation and validation stages, we leverage a suite of Python's preeminent libraries. These encompass Matplotlib, Scikit-learn, KERAS, Seaborn, Pipeline, and Pandas [81,82]. The system setup is anchored by the robust NVIDIA Tesla T4 GPU, equipped with NVIDIA's Turing architecture and is configured without display activation, indicating its dedicated use in high-performance computer operations. The system is running with a driver version of 535.104.05 and a CUDA version of 12.2, ensuring compatibility with contemporary machine learning frameworks and libraries. The Tesla T4's memory and processing capabilities are leveraged to support intensive computational tasks, underlined by its integration into the Google Compute Engine backend [66], highlighting the system's readiness for scalable and efficient data processing tasks.

2.3.1. ANN model with ReLU activation and ADAM optimizer (RELAD-ANN)

The RELAD-ANN model, illustrated in Figure 5, is specifically crafted to predict SI along with other environmental parameters specially air temperature, wind speed, and surface humidity. The model is fundamentally based on a multilayer perceptron, equipped with a network of artificial neurons. Collectively, these neurons augment its computational capacity.

At the heart of RELAD-ANN's structure is a tiered arrangement of layers: A beginning input layer, a final output layer, and several hidden layers in between. In this study, the dataset encompassed an extensive collection of 40,000 data points, categorized across four distinct parameters, aggregated on an hourly basis. The architecture, consisting of three hidden layers with 512 neurons each, was judiciously chosen based on both theoretical frameworks and empirical analyses. While the Universal

Approximation Theorem asserts that a single-hidden-layer network possesses the capability to approximate any given function, it remains non-prescriptive regarding the requisite number of neurons [83–85].

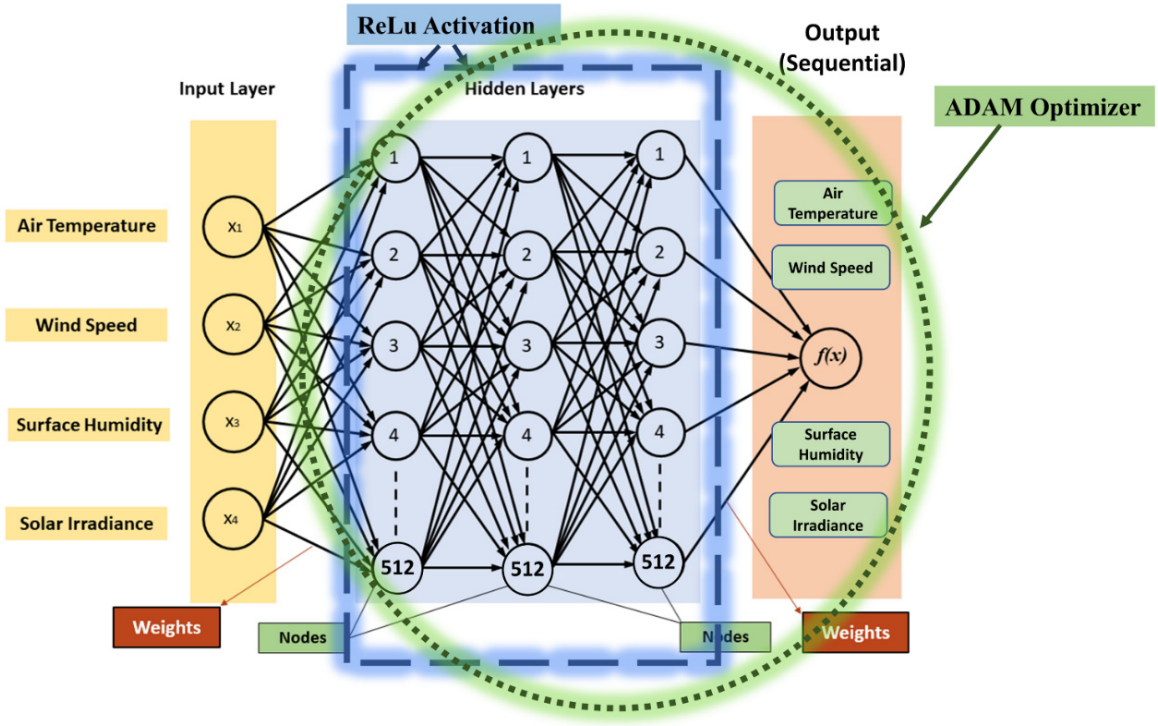

Figure 5. Illustration of RELAD-ANN.

Thus, the adopted architecture was derived from iterative experimentation, which illuminated its prowess in achieving an equilibrium between computational performance and efficiency. Furthermore, it is imperative to note that despite the model's ostensible complexity, we have integrated preventative measures, including dropout and regularization, to circumvent potential overfitting, thereby ensuring its reliable extrapolative capacity on novel data. A standout characteristic of this model is its use of the ReLU activation function in both the input and hidden layers [86]. This function plays a key role in minimizing errors, guiding the model towards highly precise forecasts. Table 1 gives detailed insight of the structured model. One of the model's strengths is its flexible number of neurons in the input and hidden layers, allowing for adaptability with varied datasets. In contrast, the neurons in the output layer are precisely set based on the output's characteristics, bringing a measure of predictability to its framework.

The training process for RELAD-ANN is thorough and systematic. The entire dataset, comprised of a notable 9223 entries, is divided into 100 epochs, each containing 10 entries. Essentially, the model processes these entries in 923 different groups, making a total of 92300 updates throughout its training phase. This repeated adjustment acts as a safeguard against mistakes, sharpening its forecasting capability.

Table 1. Hyperparameters of proposed models.

Model	Hyperparameter	Optimum value/setting
RELAD-ANN	input dimensions	4
	layers	1 input, 3 hidden, 1 output
	units per layer	input: 32, hidden: 512 (each), output: 1
	activation	relu
	optimizer	adam
	loss function	mean_squared_error (mse)
	batch size	10
	epochs	100
LSIPF	kernel	as per gridsearchcv result (linear or poly)
	c	1.0
	epsilon	0.1
	degree	3
	gamma	scale
	coef0	0
	shrinking	true
	tol	1×10^{-3}
	cache_size	200
	max_iter	–1
Light GBM	objective	'regression'
	metric	'rmse'
	learning_rate	0.1
	num_leaves	31
	verbose	–1

In the optimization of RELAD-ANN's performance, we employed the ADAM optimizer—A refined variant of the stochastic gradient descent (SGD) technique. ADAM adjusts weights considering adaptive learning rates for each parameter, adhering to the update rule in Eq 2.1;

$$\boldsymbol{\theta_t + 1 = \theta_t - \frac{\alpha . m_t}{\sqrt{\hat{v}_t} + \epsilon}} \tag{2.1}$$

where $\boldsymbol{\theta_t}$ represents the parameter vector at timestep $\boldsymbol{t}$, $\boldsymbol{\alpha}$ is the step size, $\boldsymbol{m_t}$ is the first moment estimate, $\boldsymbol{v_t}$ is the second moment estimate, and $\boldsymbol{\epsilon}$ is a small scalar used to prevent division by zero. This update rule ensures that each parameter is adjusted with an individualized learning rate, facilitating more efficient and effective convergence.

The rationale behind its adoption is the optimizer's renowned capability for adaptive learning rates for each parameter. By leveraging moment estimates of the gradients, ADAM provides a more sophisticated and efficient trajectory in the parameter space, thereby potentially accelerating convergence [87,88]. Known for its effectiveness, this optimizer significantly enhances the model's precision. Overall, the unique design, wise activation function selection, adaptable neuron setup, and state-of-the-art optimization techniques collectively make the RELAD-ANN model a standout in SI prediction.

2.3.2. Linear SVM with Individual Parameter Features (LSIPF)

In our work, we also employ an advanced LSIPF modelling technique to conceptualize our training dataset as spatial vectors. We utilize four crucial feature parameters: Air temperature, radiance

intensity, wind speed, and surface humidity, sourced from a comprehensive dataset. The primary objective of our prediction is to ascertain SI.

A salient feature of our approach is the clear demarcation that emerge between samples from distinct categories, providing an invaluable tool for the cross-validation of new data samples and enabling their efficient categorization. As illustrated in Figure 6, our model is grounded on the KERNEL linear type, a specialized mathematical function designed for transposing our training dataset into a higher-dimensional domain. Recognizing the importance of robust data preprocessing, we give a significant emphasis to feature scaling using the StandardScaler. This step ensures that our data is consistently normalized, a prerequisite for algorithms like SVM to function optimally. The SVR aims to construct a function $f(X)$ that approximates the expected outputs y_i, representing SI, within an ϵ-insensitive zone for all training data points x_i (Eq 2.2). The optimization problem that SVR solves can be formulated as:

$$\boldsymbol{min}_{\boldsymbol{w,b,\xi,\xi^*}} \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{w}\|^{\mathbf{2}} + \boldsymbol{C}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{\xi_i} + \boldsymbol{\xi^*}_{\boldsymbol{i}}) \tag{2.2}$$

This is subject to the constraints that the predicted values, $\boldsymbol{y_i}$, do not deviate from the actual values by more than the $\boldsymbol{\epsilon}$ threshold, taking into account the slack variables $\boldsymbol{\xi}$ and $\boldsymbol{\xi^*}$, which allow for flexibility in this condition (Eqs 2.3 and 2.4).

$$\boldsymbol{y_i} - \langle \boldsymbol{w},\ \boldsymbol{x_i} \rangle - \boldsymbol{b} \leq \boldsymbol{\epsilon} + \boldsymbol{\xi_i} \tag{2.3}$$

$$\langle \boldsymbol{w},\ \boldsymbol{x_i} \rangle + \boldsymbol{b} - \boldsymbol{y_i} \leq \boldsymbol{\epsilon} + \boldsymbol{\xi^*}_{\boldsymbol{i}} \tag{2.4}$$

$\boldsymbol{\xi_i \xi^*}_{\boldsymbol{i}} \geq \mathbf{0}$, for all 'i'.

here, $\boldsymbol{w}$ is the weight vector, $\boldsymbol{b}$ is the bias, $\boldsymbol{C}$ is the regularization parameter, which balances the model complexity and the degree to which deviations larger than $\boldsymbol{\epsilon}$ are tolerated.

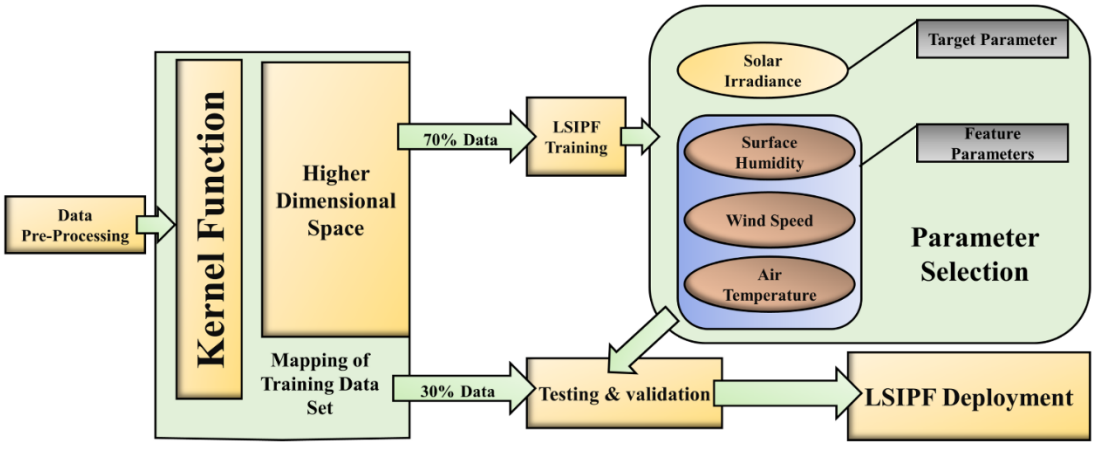

Figure 6. Illustration of LSIPF.

We harness the power of the SVM for our predictions, conducting a comprehensive search over specific kernel parameters, namely linear and poly [89,90]. Our meticulous exploration leads to the linear kernel as the superior choice, a testament to the efficacy of our methodological approach. The

kernel function, defined as $\boldsymbol{k}\left(\boldsymbol{x_i},\ \boldsymbol{x_j}\right) = \boldsymbol{\varphi}(\boldsymbol{x_i}).\boldsymbol{\varphi}\left(\boldsymbol{x_j}\right),$ represents the transformation applied to the input features. The SVR model is trained on the dataset, fitting the best hyperplane in a high-dimensional space. When making predictions for new inputs, the model utilizes the learned parameters to estimate the output $\widehat{\boldsymbol{y}}$ for a new input $\boldsymbol{X'}$, as expressed in Eq 2.5. In the presence of the kernel trick, this prediction takes the form of Eq 2.6:

$$\widehat{\boldsymbol{y}} = \langle \boldsymbol{w}, \boldsymbol{X'} \rangle + \boldsymbol{b} \tag{2.5}$$

In the kernelized version, it becomes:

$$\widehat{\boldsymbol{y}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\propto_{\boldsymbol{i}} - \propto_{\boldsymbol{i}}^{*}) \boldsymbol{K}\ (\boldsymbol{X_i}, \boldsymbol{X'}) + \boldsymbol{b} \tag{2.6}$$

here, $\propto_i$ and $\propto_i^*$ are the Lagrange multipliers obtained from solving the dual optimization problem, and $\boldsymbol{K}\ (\boldsymbol{X_i}, \boldsymbol{X'})$ is the kernel function evaluating the similarity between the support vectors and the new input $\boldsymbol{X'}$.

However, while our LSIPF model demonstrates marked proficiency, it is imperative to acknowledge its inherent limitations. Its exclusive reliance on a singular layer for data interpretation might pose challenges in certain scenarios, potentially affecting the outcome fidelity. Furthermore, the model's performance is closely linked to the quality and abundance of the training dataset; any discrepancies here could influence the predictive accuracy.

2.4. Analysing meteorological parameter influence on SI using advanced regression techniques

In the realm of solar energy generation, the intricate interplay between various meteorological parameters assumes paramount importance, especially considering the direct influence of SI on the power yield of PV installations. A profound comprehension of the consequences engendered by disparate parameters on SI can augment the precision of forecasting models, thereby catalyzing the efficacious harnessing of solar energy [91–93]. To actualize this objective, we rigorously employ two sophisticated regression methodologies: SVR and Light GBM. These models quite accurately predict the implications of three salient parameters - wind speed, air temperature, and surface humidity - on SI. Such a methodical approach not only epitomizes the vanguard of predictive modeling in renewable energy but also underscores the imperative of understanding parameter interrelationships for optimizing solar energy outcomes.

2.4.1. Support Vector Regression (SVR)

SVR is an adaptation of the SVM methodology, tailored for predictive analysis of continuous data. While SVM is typically used to classify data into distinct categories, SVR works differently. It focuses on determining an optimal fit that can predict continuous outcomes. This fit is not just about minimizing errors; it is also about ensuring that errors do not exceed a certain threshold. By setting up a boundary around our prediction line, SVR gives precedence to data points that are close to this line, ensuring a more consistent prediction quality [94–96].

In the context of our research, SI emerged as a pivotal parameter among the four we analyzed. Recognizing its significance, we employ an SVR model using a linear kernel to delve deeper into SI's relationship with the other three parameters. The choice of a linear kernel is crucial here. It allows the model to capture straightforward relationships between inputs and predicted outputs, enabling us to predict SI values with greater accuracy based on the interplay of the other parameters.

Further emphasizing the importance of this approach, employing the SVR model with a linear kernel provides us with a robust analytical tool. It offers clarity in understanding data relationships and ensures that our predictions are both accurate and consistent. This methodological choice underscores our commitment to delivering high-quality research insights, making our findings not only relevant but also trustworthy.

2.4.2. Lightweight Gradient Boosting Machines (Light GBM)

Within our exploration focused on understanding the factors influencing SI, we chose to employ the Light GBM regressor. This tool, available in the public domain, has consistently delivered reliable results in similar studies. To validate its efficiency for our dataset, we subjected it to a five-fold cross-validation process [97]. This technique involves dividing our data into five equal parts and, in a cyclical manner, using four parts for training and one part for testing. This process not only ensures a comprehensive assessment but also reduces any biases that might arise from the dataset's inherent randomness. The hyperparameters of the proposed model is mentioned in Table 1.

Delving into the specifics of Light GBM, it is a gradient boosting platform built on decision tree algorithms, suitable for a range of machine learning tasks, including classification and ranking. What differentiates Light GBM from other algorithms is its unique leaf-wise approach to tree splitting [98]. Instead of the traditional level-wise method, Light GBM optimizes its accuracy by minimizing potential losses through this leaf-wise method. In addition to its precision, Light GBM stands out for its speed. Aptly named "Light", it is designed to manage large datasets efficiently with minimal memory usage and even supports GPU learning [99].

In the Light GBM model, the L2 loss function measures the difference between the predicted values and the actual values. It does so by squaring the difference for each data point and then taking an average of squared differences. The L2 loss function helps guide the model during training to minimize the discrepancies between predicted and actual values.

Mathematically, the L2 loss function for a set of predictions $\hat{y}i$ and true values yi is defined as:

$$\boldsymbol{L2Loss} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} (\boldsymbol{yi} + \hat{\mathbf{y}}\boldsymbol{i})^{\mathbf{2}} \tag{2.7}$$

where:

- n is the number of data points;
- yi is the true value for the i-th data point; and
- $\hat{y}i$ is the predicted value for the i-th data point.

To conclude, the choice of Light GBM in our study is a reflection of our aim to use efficient and accurate tools. It reinforces our dedication to producing reliable results, emphasizing the significance of our findings.

2.5. Model validation

In order to rigorously validate the proposed models, this research employs an array of statistical metrics [100], namely coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE).These metrics serve a dual purpose: First, they facilitate a direct comparison between the models' predictions and the values encountered during the testing phase; and second, they offer insight into the upper limits of potential errors, thereby characterizing the models' overall performance and reliability. Furthermore, by employing these

indicators, we are better positioned to assess the respective strengths of the LSIPF and RELAD-ANN models and determine their optimal applications under specific scenarios. This systematic evaluation approach ensures the robustness of our findings, emphasizing the significance of the models' performance metrics in the broader context of the study.

2.6. Comparison with existing models

In our study, we have conducted a comprehensive comparative analysis of machine learning models for SI prediction, benchmarking our proposed models against a selection of five different algorithms as proposed by various researchers. This comparison includes a diverse array of models such as MLP [101], LSTM [102], MLSTM [103], GRU [104], and 1-dimensional CNN [105], where CNN is majorly used for image-based predictions, we used a Shallow CNN model for time series forecasting. Each of these models has been evaluated based on their efficiency and accuracy in predicting SI. It is important to note that while conducting this comparative study, we have made certain adjustments to the hyperparameters of these models to align them more closely with our data and its specific characteristics which are showcased in Table 2. This customization ensures that the comparison is not only fair but also relevant to our unique dataset and the particular requirements of our study. By doing so, we aim to provide a clear and objective assessment of how our proposed models, the RELAD-ANN and LSIPF, stand in relation to established algorithms in the field, thereby validating their effectiveness and suitability for SI prediction.

Table 2. Hyperparameters of comparative models.

Hyperparameters	MLP	LSTM	MLSTM	GRU	CNN
Number of layers	1, 3 Dense	1	2	2	1, 2 Dense
Neurons in layer 1	128	50	50	50	50
Neurons in layer 2	64		50	50	50
Filters	-	-	-	-	64
Kernel size	-	-	-	-	1
Flatten layer	-	-	-	-	used
Output layer neurons	1	1	1	1	1
Activation function	ReLU	ReLU	ReLU	ReLU	ReLU
Return sequences	-	-	True, False	True, False	-
Output layer activation	Linear	Linear	Linear	Linear	Linear
Dropout rate	50% (0.5)	-	-	-	-
Optimizer	Adam	Adam	Adam	Adam	Adam
Loss function	MSE	MSE	MSE	MSE	MSE
Batch size	32	32	32	32	32
Epochs	10	50	50	50	50
Validation split	20%	20%	20%	20%	20%

3. Results and discussion

In this section, we present our findings and discuss their implications. The section is organized into subsections that each address a specific aspect of our research. First, we examine the performance of the RELAD-ANN and LSIPF models, discussing their forecasting accuracy and computational efficiency. Second, we analyze the influence of different meteorological parameters on SI predictions,

providing insights into the significance of each predictor. Last, we compare our results with existing models, highlighting the advancements our approach offers. This structure is chosen to clearly delineate the progression of our research from model assessment to a broader evaluation within the field.

3.1. Parametric forecasting

The arena of environmental prediction has witnessed a paradigm shift with the advent of the RELAD-ANN model, meticulously detailed in section 2.3.1 and visually encapsulated in Figure 4. Embodied with both innovation and precision, this model underwent rigorous scrutiny across an array of parameters, most notably SI—A parameter of paramount significance.

Table 3 shows that the RELAD-ANN model achieves a high accuracy rate of 96.8% for SI predictions. The small mean error of 3.2% supports the model's reliability, as further depicted in Figure 7 (a–d). The model's ability to accurately capture the varying patterns of SI, particularly daily changes, is evident.

Our analysis revealed that while the predictions were generally accurate, there were identifiable patterns in the outliers, particularly during transitional times like dawn and dusk, where deviations reached up to 4.5%. These variations may be due to atmospheric conditions, instrument sensitivities, or geometric factors such as shading or the sun's angle. Table 3 further demonstrates the effectiveness of the RELAD-ANN model in predicting other environmental parameters. It shows high accuracy in forecasting surface humidity (97.2%), air temperature (95.4%), and wind speed (94.7%). These results indicate the model's ability to handle complex environmental data.

The study also discovered a notable relationship between SI and surface humidity, with a correlation coefficient of 0.78, highlighting the complex interactions within our environment. This finding underscores the RELAD-ANN model's ability to discern intricate patterns.

Table 3. Prediction details of each parameter.

Parameters		Solar irradiance	Wind speed	Air temperature	Surface humidity
Maximum actual value		391.5	16.4	31.2	0.02
Minimum actual value		150.0	0.9	–9.8	0.0005
Maximum predicted value	RELAD-ANN	373.6	8.3	27.6	0.02
	LSIPF	367.0	6.1	31.7	0.01
Minimum predicted value	RELAD-ANN	175.0	2.7	–9.3	–0.003
	LSIPF	172.0	4.3	–10.4	0.01
Maximum variance with actual	RELAD-ANN	55.4	11.6	16.2	0.007
	LSIPF	55.7	10.3	16.6	0.008
Minimum variance with actual	RELAD-ANN	0.0013	0.003	0.001	9.1×10^{-8}
	LSIPF	0.0016	8.5×10^{-5}	0.001	5.0×10^{-6}
Average variance	RELAD-ANN	8.2	1.8	2.7	0.0006
	LSIPF	12.0	1.7	3.3	0.006

The architectural design of the RELAD-ANN model is central to its success. It combines a multilayer perceptron structure with ReLU activation and the ADAM optimizer, enhancing its precision in environmental forecasting.

To encapsulate, the RELAD-ANN model demonstrates high accuracy in SI prediction, as shown by the 96.8% prediction rate in Table 3. This accuracy, along with the model's ability to analyze outliers and environmental correlations, confirms its significant role in advancing environmental forecasting. The RELAD-ANN model represents a new era in predictive modeling, characterized by its innovation, accuracy, and depth of understanding.

Within environmental forecasting, the LSIPF model, utilizing a KERNEL linear type, stands as a significant tool. It transforms the training dataset into spatial vectors to integrate key features: Air temperature, radiance intensity, wind speed, and surface humidity. The model's strength lies in its ability to predict SI.

Table 3 reveals that while the LSIPF model is effective in some areas, it shows differences compared to the RELAD-ANN model, especially in predicting surface humidity. This gap might be due to factors like consistent rainfall patterns in the studied region, which could affect the LSIPF model's ability to detect subtle humidity shifts with its single-layer data interpretation approach.

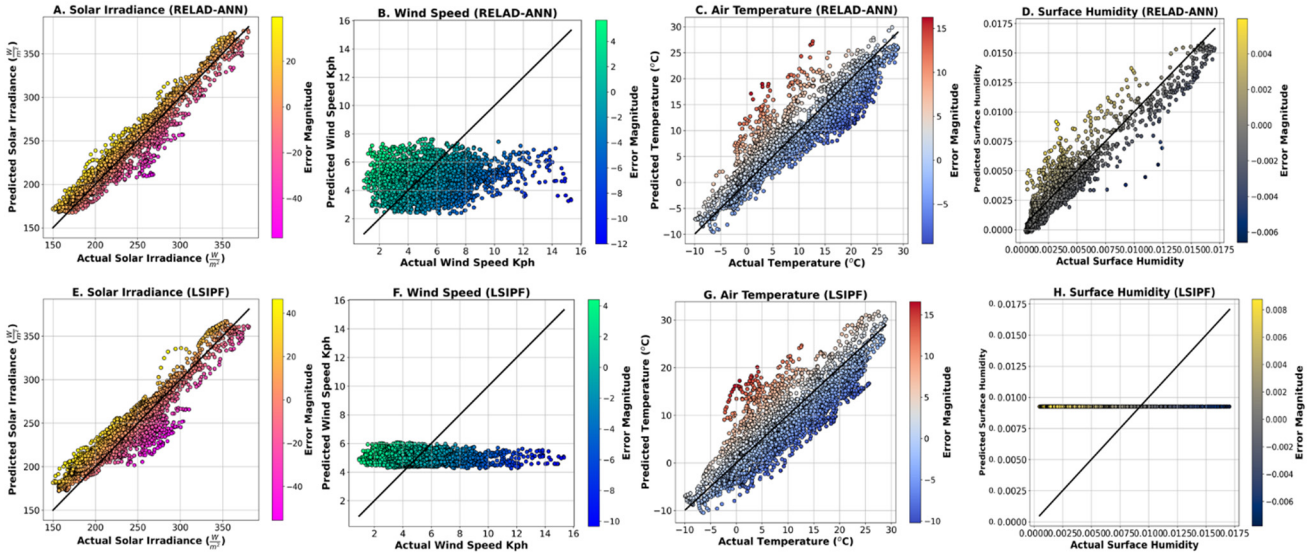

Figure 7. Predictions for testing data: (a) SI (RELAD-ANN), (b) wind speed (RELAD-ANN), (c) air temperature (RELAD-ANN), (d) surface humidity (RELAD-ANN), (e) SI (LSIPF), (f) wind speed (LSIPF), (g) air temperature (LSIPF), and (h) surface humidity (LSIPF).

However, the LSIPF model demonstrates good performance in predicting wind speed and air temperature. In terms of wind speed prediction, the LSIPF model occasionally outperforms the RELAD-ANN model. Yet, in SI prediction, as shown in Figure 7 (e–h), the LSIPF model does not match the high accuracy of the RELAD-ANN model, especially in predicting surface humidity, although it remains consistent in its predictions related to air temperature and wind speed.

A comprehensive analysis of both models, RELAD-ANN and LSIPF, indicates that each offers unique strengths. However, the RELAD-ANN model particularly excels in the domain of SI, as evidenced by Table 4. It shows a superior R^2 value of 0.936 for SI predictions, surpassing the LSIPF's R^2 value of 0.877, highlighting the distinct advantages of the RELAD-ANN model in this aspect.

Table 4. Empirical validation of proposed models.

Parameters	Model	R^2	MAPE	MAE	RMSE
Solar irradiance	LSIPF	0.877	0.053	11.95	15.09
	RELAD-ANN	0.936	0.035	8.17	10.89
Wind speed	LSIPF	0.04	0.377	1.70	2.26
	RELAD-ANN	0.23	0.389	1.78	2.33
Air temperature	LSIPF	0.73	1.8	3.31	4.2
	RELAD-ANN	0.82	1.56	2.65	3.49
Surface humidity	LSIPF	–4.86	3.64	0.01	0.01
	RELAD-ANN	0.88	0.281	0.0006	0.001

The RELAD-ANN model's performance in predicting wind speed and air temperature is on par with the LSIPF model, but it particularly excels in forecasting SI and surface humidity. The advanced architecture of RELAD-ANN's neural network enables it to effectively assimilate complex data patterns, which is essential for accurately predicting the variable nature of SI. Figure 8 presents a comparison of both models against actual data, highlighting their respective performances across different parameters. The overall predictions can be visualized on right hand side of Figure 8, while a Kernel Density Estimation (KDE) plot is presented on left hand side for a better understanding.

In the area of surface humidity prediction, an important facet of meteorological forecasting, the RELAD-ANN model clearly outperforms the LSIPF model. The LSIPF model's linear methodology struggles with the intricate nature of surface humidity, influenced by various atmospheric conditions. This is evident in Figure 8(d), where RELAD-ANN aligns closely with the actual data, whereas LSIPF shows notable variances.

While both models are competitive in predicting wind speed and air temperature, RELAD-ANN's adaptability and learning capabilities give it a distinct advantage, especially in dealing with sudden data shifts or anomalies. The LSIPF model, despite its solid foundational approach, faces challenges in accurately forecasting more complex parameters like surface humidity and SI.

The underlying strength of RELAD-ANN resides in its architectural framework. Distinct from traditional forecasting models, RELAD-ANN employs an intricate artificial neural network structure. This configuration, layered and interconnected, empowers it with an enhanced capacity for data assimilation and pattern recognition. The novelty of the RELAD-ANN model arises from its ability to dynamically adapt. It can self-learn from historical data, refine its forecasting algorithms, and consequently, deliver more accurate predictions. This, coupled with its proficiency in discerning minute data variations—A capability imperative for surface humidity predictions—Accentuates its superiority.

Conversely, the LSIPF model, though competent, is intrinsically limited by its design. Its linear nature can sometimes be insufficient in grappling with the multifaceted and interconnected variables of meteorological data. This becomes evident in its struggle to forecast surface humidity, where it manages only a meager R^2 value of approximately zero compared to 0.88 for RELAD-ANN. Such quantitative disparities highlight the stark difference in the models' capabilities. In summary, while LSIPF offers a foundational approach to prediction, RELAD-ANN, with its advanced structure and innovative mechanisms, stands out as the avant-garde in meteorological forecasting.

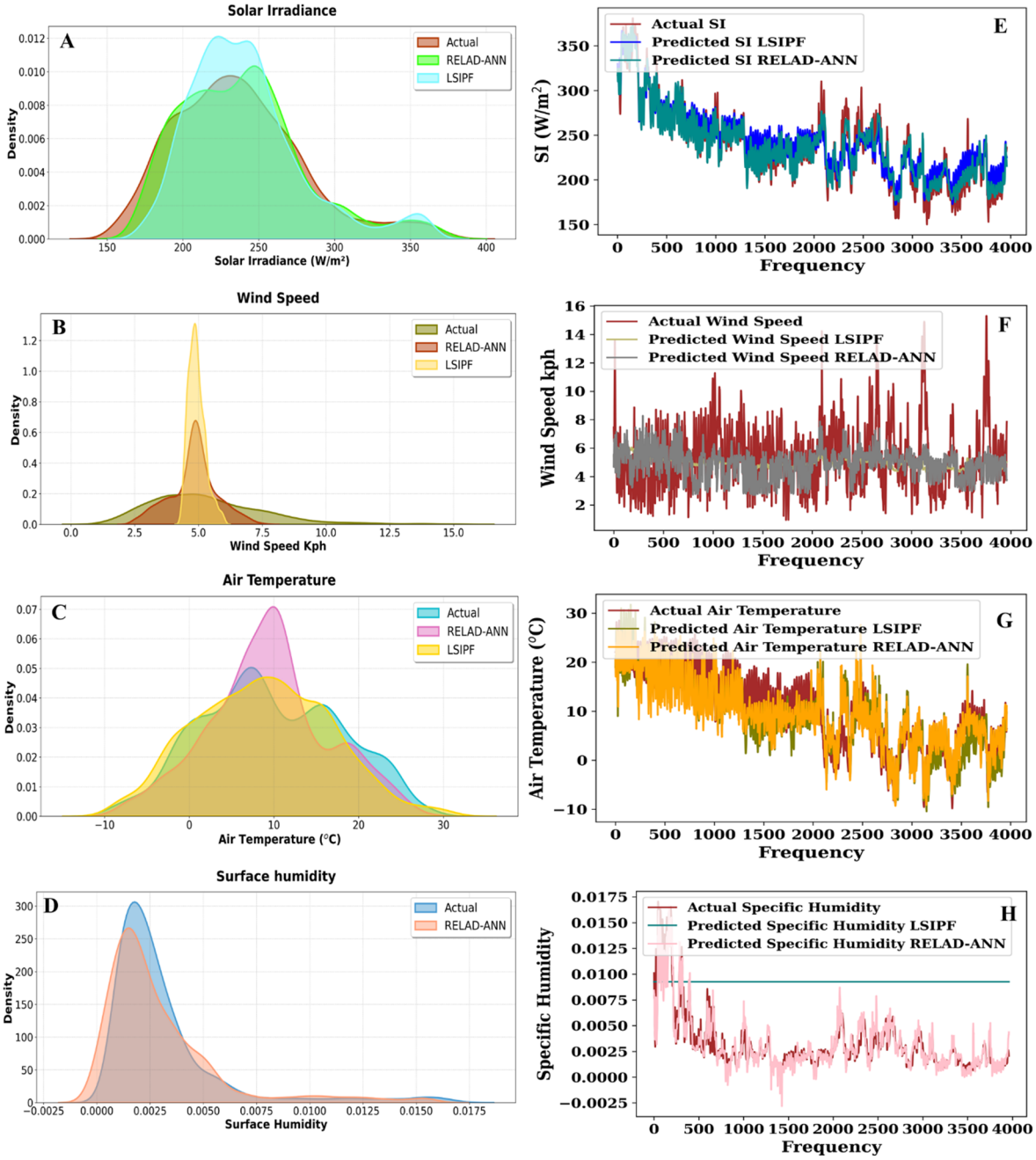

Figure 8. Comparison between RELAD-ANN and LSIPF over actual data (KDE and Frequency Plots): (a) and (e) SI, (b) and (f) Wind Speed, (c) and (g) Air temperature, (d) and (h) Surface humidity.

3.2. Meteorological parameter influence on SI

We attempt to delve into the effects of various environmental parameters, namely wind speed, surface humidity, and air temperature, on SI through the prism of the SVR and Light GBM models.

The SVR model revealed a nearly linear relationship between air temperature and SI, as shown in Figure 9(a). It indicated that an increase in air temperature is typically accompanied by a rise in SI. The model also effectively captured the relationship between wind speed and SI, particularly within

the range of 2 to 8 kph, as depicted in Figure 9(b). However, the SVR model's limitations became apparent when dealing with surface humidity data. In this aspect, as illustrated in Figure 9(c), the model struggled to provide accurate predictions, indicating its shortcomings in addressing the complexities of surface humidity.

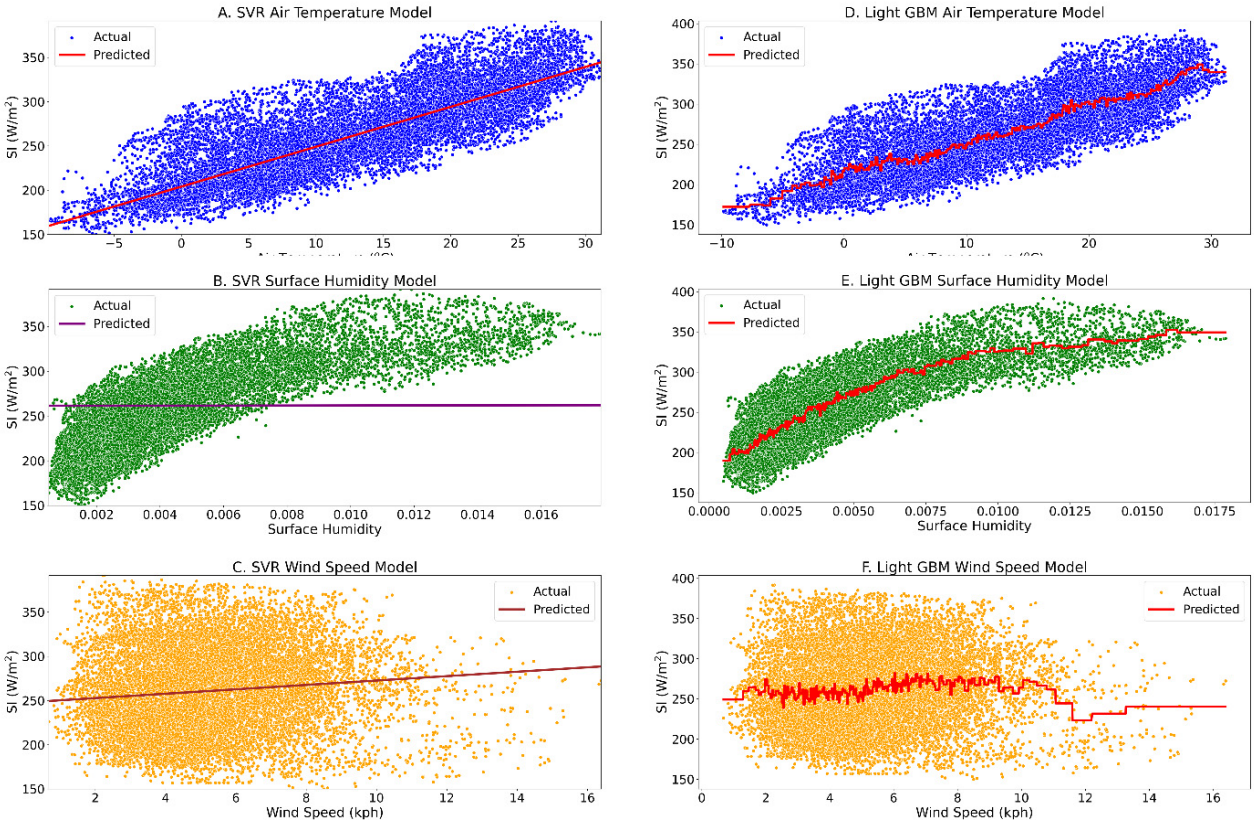

Figure 9. Impact of various parameters on SI: (a) SVR Air Temperature Model, (b) SVR Wind Speed Model, (c) SVR Surface humidity Model, (d) Light GBM Air Temperature Model, (e) Light GBM Wind Speed Model, and (f) Light GBM Surface humidity Model.

In contrast, the Light GBM model exhibited superior performance, especially with its optimization using the L2 loss function. Light GBM's predictions for SI, considering the trio of environmental parameters, are detailed across Figure 9(d–f). Notably, the model accurately predicted the highest SI at 393.8 W/m^2, correlating with an air temperature of 27.9 ℃, a wind speed of 2.3 kph, and a surface humidity of 0.01. Conversely, the lowest SI was predicted at 171.1 W/m^2, associated with an air temperature of –2.2 ℃, wind speed of 8.3 kph, and surface humidity of 0.002. Light GBM's ability to correlate air temperature and wind speed with SI was particularly effective, and it also successfully captured the crucial role of surface humidity in influencing SI, an aspect that the SVR model was less adept at.

The comparison between SVR and Light GBM models highlights their differing capabilities in analyzing the complex interaction of environmental parameters with SI. Light GBM's proficiency in handling complex datasets and its sensitivity to subtle changes in input parameters contribute to its enhanced performance. The optimization of the L2 loss function in Light GBM plays a key role in reducing discrepancies between predicted and actual data.

The radar chart in Figure 10(a) elucidates the correlation coefficients with remarkable clarity, revealing a robust positive correlation between SI and both air temperature and surface humidity, recorded at 0.811 and 0.812 respectively. In stark contrast, wind speed's correlation with SI is notably weak, registering a minimal value of 0.064. This disparity in correlation strengths is visually captured by the radar chart, which graphically conveys the varying degrees of linear association between SI and the environmental parameters in question.

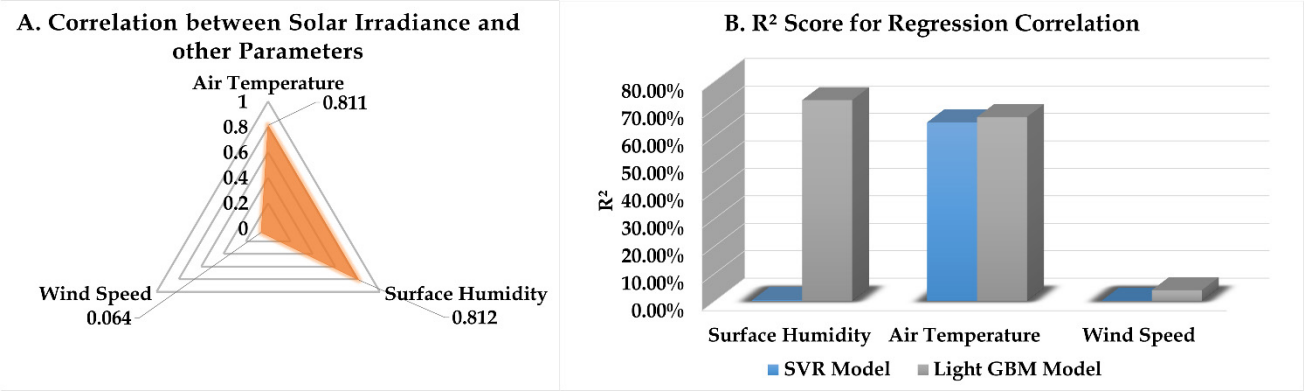

Figure 10. (a) Correlation between SI and other parameters. (b) R^2 score of regression models.

Further substantiating the efficacy of the Light GBM model, the bar chart in Figure 10(b) articulates the R^2 scores, a statistical measure reflecting the percentage of the response variable variation that is captured by the model. In the domain of air temperature, the Light GBM model exhibits a commendable R^2 score of 66.95%, marginally surpassing the SVR model's performance of 64.99%. The distinction is more pronounced in the context of surface humidity, where Light GBM achieves an R^2 score of 73.10%, indicative of a superior model fit as opposed to SVR's notably lower score, which is a mere 0.33%. The R^2 score for wind speed with Light GBM stands at 3.92%, which, while modest, is an improvement over SVR's marginal score of 0.12%.

Concluding our observations, Light GBM emerges as the more robust and versatile model for assessing the influence of environmental factors on SI. Its holistic approach, embracing the intricate interrelationships between wind speed, surface humidity, and air temperature, positions it as a superior predictive tool, overshadowing the capabilities of SVR. Our findings suggest that SI is closely associated with air temperature and surface humidity but has an almost negligible linear relationship with wind speed. This emphasizes the potential of Light GBM in solar energy forecasting, leveraging its nuanced understanding of complex, non-linear interdependencies among environmental variables.

3.3. Comparative analysis

In the comparative study of machine learning models for SI prediction, seven different algorithms were evaluated based on their efficiency and accuracy. The models included MLP, LSTM, MLSTM, GRU, CNN, and the proposed models RELAD-ANN and LSIPF. The performance of these models can be envisioned in Figure 11, the 3D-bar graph gives visual insight of all the models compared with actual data.

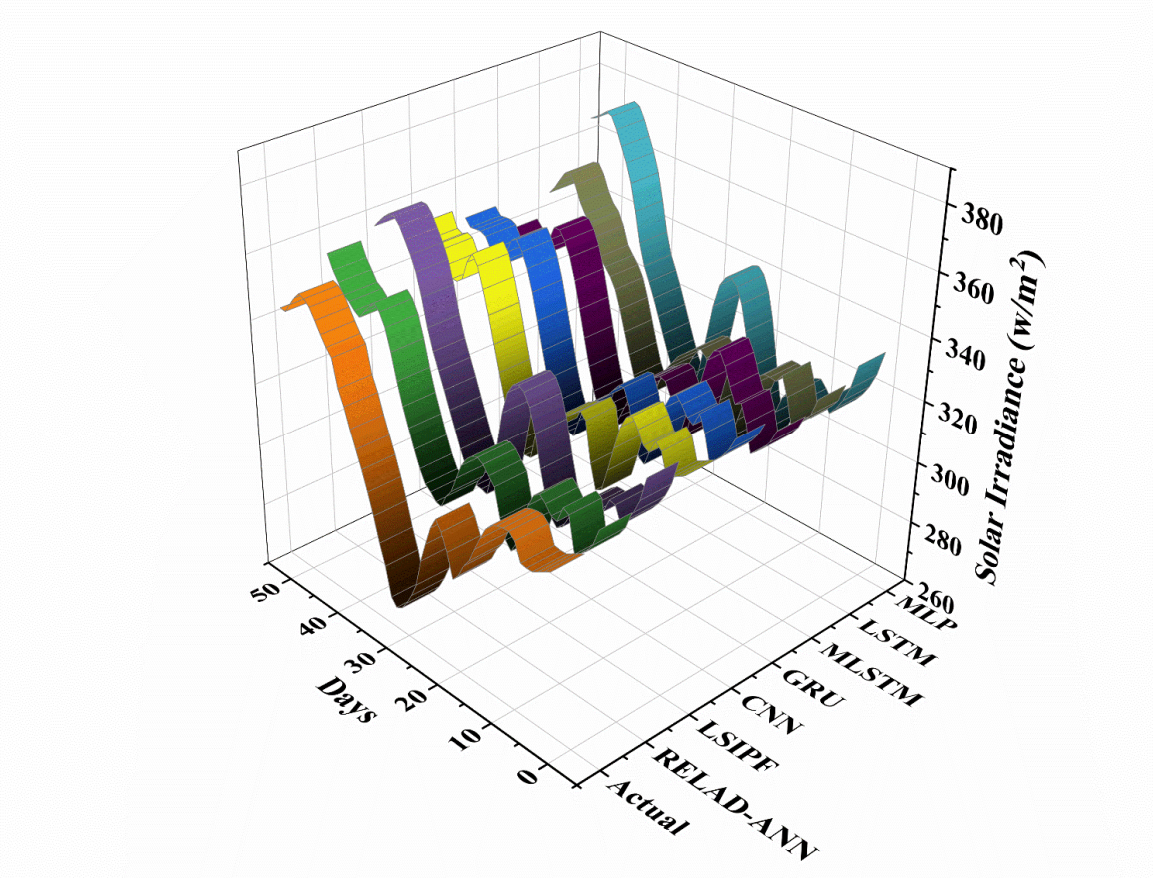

Figure 11. Visual analysis of SI prediction via different models.

In terms of computational efficiency, training times were notably similar across the models as illustrated in Table 5, with the MLP, LSTM, MLSTM, and CNN models all training within an approximate window of 4.82 to 4.86 × 10^{-5} seconds. The GRU model required a slightly longer duration, at 7.44 × 10^{-5} seconds, potentially due to the inherent complexity of its recurrent structure. Testing times followed a similar pattern, with the RELAD-ANN model emerging as the most time-efficient at 3.50 × 10^{-5} seconds, closely followed by LSTM and MLSTM at 3.65 and 3.67 × 10^{-5} seconds, respectively.

Table 5. Time comparison of various models.

	MLP	LSTM	MLSTM	GRU	CNN	RELAD-ANN	LSIPF
Training time (sec. 10^{-5})	4.82	4.41	4.36	7.44	4.86	6.1	5.91
Testing time (sec. 10^{-5})	4.81	3.65	3.67	4.98	5.29	3.5	4.46

The RELAD-ANN model emerged as the standout performer in the accuracy assessment as depicted in Table 6, recording the most favorable statistical indicators with an MAE of 8.20, MAPE of 3.48%, and an R^2 of 0.935. These metrics underscore its proficiency in closely tracking the observed SI values with minimal deviation. In contrast, the LSIPF model, managed to capture a broad trend in the dataset, as reflected by an R^2 of 0.876, but its precision was less convincing, with an MAE of 11.96 and the highest RMSE among the models at 15.09. This suggests that while LSIPF could grasp the

general variance within the data, it struggled with exact predictions, possibly due to suboptimal kernel and regularization parameter choices inherent to SVR.

Table 6. Statical validation of compared models.

	MLP	LSTM	MLSTM	GRU	CNN	RELAD-ANN	LSIPF
MAE	11.89	8.32	8.22	8.92	9.49	8.17	11.96
MAPE	5.43	3.60	3.53	3.92	4.17	0.035	5.33
RMSE	14.98	11.09	11.20	11.44	11.98	10.89	15.09
R^2	0.878	0.933	0.932	0.929	0.922	0.936	0.876

Examining other models, the LSTM and MLSTM models achieved commendable accuracy with R^2 values of 0.933 and 0.932, respectively, signifying their robustness in fitting the data. Their MAE and MAPE values were competitive at around 8.32 and 3.60% for LSTM, and 8.22 and 3.53% for MLSTM, reinforcing their reliability in predictions closely behind the leading RELAD-ANN model. The GRU model, another variant within the recurrent neural networks, also showcased strong modeling capability with an R^2 of 0.929, although its error metrics, with an MAE of 8.92 and MAPE of 3.92%, were marginally higher than the LSTM variants. The CNN model, typically known for image processing tasks, was adapted for time series forecasting, yielding an R^2 of 0.922; while this was lower than the recurrent models, it indicated a high level of fit to the data variance. However, its error rates, with an MAE of 9.49 and RMSE of 11.98, pointed to a less precise forecasting ability compared to the leading models.

In this collective assessment, the GRU model, while slightly behind its LSTM counterparts in precision, was nonetheless effective, with the CNN model demonstrating that convolutional architectures can transition beyond their conventional image-focused domain to provide substantial time series forecasting capabilities. However, the MLP model, often a baseline in neural network comparisons, was observed to have the highest errors, with an MAE of 11.89 and an RMSE of 14.98, suggesting that more complex network architectures could be more suitable for this specific forecasting task.

The applicability of these models to other locations would depend on the data's consistency with the environment where the models were originally trained. If the underlying patterns of SI and its influencing factors are similar, models that performed well in this study, particularly the RELAD-ANN, may continue to exhibit robust predictive capabilities. However, if the new locations present different environmental characteristics or data distributions, reevaluation and potential recalibration of the models would be necessary to maintain predictive accuracy.

Despite the RELAD-ANN model's evident predictive prowess, it does carry limitations common to artificial neural networks. Its requirement for ample data to discern the underlying patterns, the potential for overfitting, and its inherent "black box" nature pose challenges to its interpretability and transferability. These constraints necessitate cautious application, particularly in scenarios where model transparency and the ability to generalize are paramount.

In conclusion, the RELAD-ANN model proved to be the most effective for SI prediction within the confinement of this study, balancing time efficiency with statistical accuracy. However, the generalization of this model to different geographies or temporal ranges calls for a tailored approach, ensuring that the model's strengths are leveraged without overlooking the contextual dynamics of the new data environment.

4. Ablation studies

We conduct a two-fold investigation to enhance and validate our SI forecasting models. Initially, we perform a Feature Analysis on RELAD-ANN model as it outperformed LSIPF in Quetta case study, methodically evaluating the impact of various meteorological parameters to identify the most influential features for accurate forecasting. This step helps in optimizing the model's input features for better performance. Subsequently, we undertake Global Validation, applying both models (LSIPF & RELAD-ANN) across diverse geographic locations to test their adaptability and effectiveness in different climatic conditions. This dual approach allows us to refine the models' configuration for enhanced accuracy and assess its universal applicability, ensuring their effectiveness and reliability for SI predictions worldwide.

4.1. Feature analysis

In Feature Analysis, we evaluate the impact of various feature sets [106] on the predictive performance of the RELAD-ANN model for SI forecasting in Quetta. We meticulously analyze seven different feature combinations, with the goal of determining the optimal feature set that balances performance with model computational efficiency.

As presented in Table 7, Feature Set 1, which encompassed a broad spectrum of six variables, not only achieved an R^2 value of 0.934 but also maintained favorable error statistics, with a MAE of 8.05, a MSE of 115.67, and a RMSE of 10.75. The exclusion of 'Surface Albedo' in Feature Set 2 offered a nuanced improvement, reflected by a superior R^2 value of 0.944, and further reductions in MAE to 7.37, MSE to 103.98, and RMSE to 10.19, subtly hinting at the limited impact of this variable on the model's efficiency. Figure 12 presents a clear visualization of different Feature sets predictions over actual SI values.

Table 7. Statistical metrics of feature analysis.

Feature set	Feature combination	R^2	MAE	MSE	RMSE
Feature set 1	Air temperature, wind speed, surface humidity, surface skin temperature, total surface precipitation, surface albedo	0.934	8.05	115.67	10.75
Feature set 2	Air temperature, wind speed, surface humidity, surface skin temperature, total surface precipitation	0.944	7.37	103.98	10.19
Feature set 3	Air temperature, surface humidity, surface skin temperature, total surface precipitation, surface albedo	0.941	7.76	108.82	10.43
Feature set 4	Surface skin temperature, total surface precipitation, surface albedo	0.60	21.13	734.29	27.09
Feature set 5	Air temperature, wind speed, surface humidity, surface skin temperature	0.929	8.21	130.63	11.43
Feature set 6	Surface humidity, surface skin temperature, total surface precipitation, surface albedo	0.893	10.85	196.89	14.03
Feature set 7	Air temperature, wind speed, surface humidity	0.921	9.84	146.64	12.10

Further pruning of 'wind speed' from Feature Set 2, resulting in Feature Set 3, led to a marginal decrease in R^2 to 0.941, alongside a slight increase in error metrics (MAE of 7.76, MSE of 108.82,

RMSE of 10.43), indicating the non-critical yet valuable role of 'wind speed'. The significant curtailment to three variables in Feature Set 4 precipitated a notable plunge in model performance, with an R^2 of only 0.60, and elevated errors (MAE of 21.13, MSE of 734.29, RMSE of 27.09), underscoring the importance of the omitted predictors.

Conversely, the removal of 'Total surface Precipitation' in Feature Set 5 reduced the model's explanatory power to an R^2 of 0.929, accompanied by a rise in errors (MAE of 8.21, MSE of 130.63, RMSE of 11.43), suggesting a moderate but tangible impact of this variable. The exclusion of both 'air temperature' and 'wind speed' in Feature Set 6 resulted in a further decreased R^2 of 0.893 and increased errors (MAE of 10.85, MSE of 196.89, RMSE of 14.03), reaffirming the pivotal role of these features in accurate SI prediction.

Ultimately, Feature Set 7, which amalgamated 'air temperature', 'wind speed', and 'surface humidity', achieved an optimal trade-off between model simplicity and predictive accuracy. With an R^2 of 0.921, an MAE of 9.84, an MSE of 146.64, and an RMSE of 12.10, it did not reach the highest R^2 but successfully balanced essential features and model performance. This strategic choice of feature set aligns with our goal to develop a model that offers precision without overcomplication, demonstrating a conscientious approach towards efficient and effective SI forecasting. This streamlined feature set not only fosters a clearer understanding of the model's workings, facilitating broader acceptance and application, but also underscores the model's Computational Efficiency. The fewer the features, the swifter the computation, a boon for real-time forecasting and scalability to larger datasets.

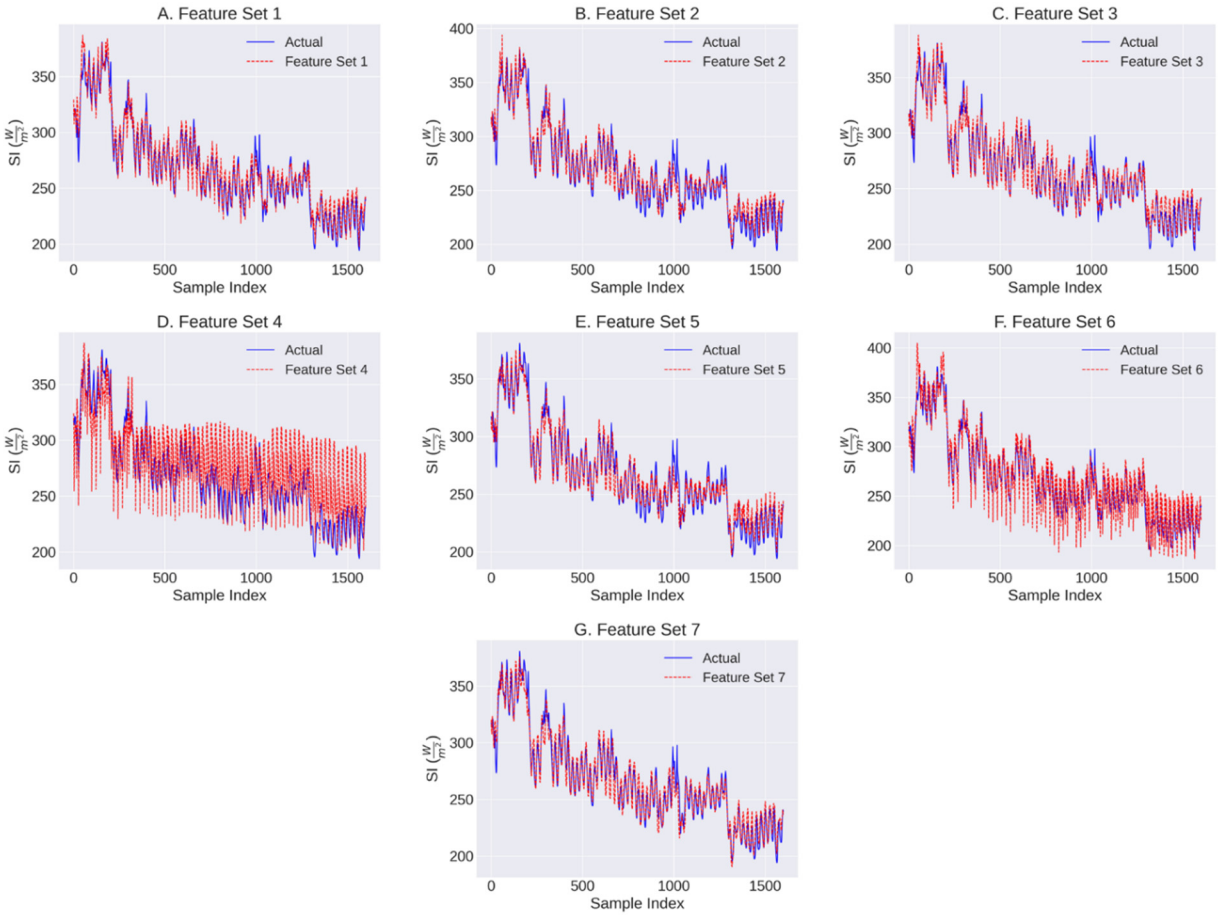

Figure 12. SI predictions of different feature sets.

Our adherence to the Parsimony Principle [107], commonly known as Occam's Razor, further justifies our selection of Feature Set 7. By opting for a model that performs comparably to more complex counterparts with lesser complexity, we embody the principle that simplicity is preferred when predictive performance is not compromised. Thus, despite other feature sets presenting marginal gains in accuracy, Feature Set 7 is identified as the optimal choice. It strikes a prudent balance between high predictive capability and model parsimony, yielding a solution that is not only scientifically robust but also pragmatically suited to end-user deployment and interpretation.

4.2. Global validation of proposed SI models

The comprehensive analysis across the five chosen global locations (Table 8) provides an insightful evaluation of the predictive capabilities of the LSIPF and RELAD-ANN models for SI forecasting. The detailed data outlined in Table 9 allows for a thorough comparison in terms of several key performance indicators.

In Sana'a (Figure 13a), both models showcase strong predictive abilities with the LSIPF model yielding an R^2 value of 0.884, indicative of a high degree of variance explained. However, the RELAD-ANN model, with an almost identical R^2 of 0.882, suggests a slight edge in precision with a marginally lower MAE, although it does register a slightly higher RMSE. The near-identical MAPE for both models implies that the percentage errors relative to the actual values are similar, pointing towards comparable forecasting reliability from a percentage error standpoint.

Table 8. Solar potential of understudy locations.

City	Country	Terrain elevation (m)	GHI/year (kWh/m^2)	DNI/year (kWh/m^2)	Air temperature (°C)
Sana'a [108]	Yemen	2252	2348.4	2322.0	19.1
Kabul [109]	Afghanistan	1802	2036.7	2196.4	13.3
Denver [110]	USA	1636	1736.6	2163.9	10.2
Dushanbe [111]	Tajikistan	834	1734.2	1690.1	15.2
Granada [112]	Spain	689	1882.7	2191.5	16.6

Kabul's (Figure 13b) results differentiate the models more clearly. The RELAD-ANN model not only achieves a higher R^2 value of 0.879 compared to LSIPF's 0.807 but also records a significantly lower MAPE and MAE. The reduced error metrics of the RELAD-ANN model, with a notable difference in RMSE, highlight its superior performance in this location, pointing to its robustness in more variable climates.

Denver's (Figure 13c) assessment further underscores the superior adaptability of the RELAD-ANN model. The LSIPF's lower R^2 and higher error metrics suggest a weaker fit for the data, while the RELAD-ANN model, with an R^2 of 0.805 and notably lower MAE and RMSE values, shows an enhanced ability to model the SI accurately, despite the complexities presented by the location's varied climate. In Dushanbe (Figure 13d), the RELAD-ANN model continues to outperform LSIPF, with a higher R^2 value of 0.795 and substantially better error metrics. The lower MAPE of 0.020 and the reduced MAE and RMSE underscore the RELAD-ANN model's efficacy in capturing the nuanced SI patterns, reaffirming its strength in diverse geographical conditions.

Finally, in Granada (Figure 13e), while the RELAD-ANN model does not exhibit as stark a contrast in R^2 value compared to LSIPF as in other locations, it maintains a higher value of 0.700. The error metrics, particularly the MAE and RMSE, are moderately better than those of LSIPF, suggesting

that even in environments with moderate SI variability, the RELAD-ANN model consistently provides a more accurate prediction.

Table 9. Key Performance Indicators for ablation study locations.

Parameters	Model	R^2	MAPE	MAE	RMSE
Sana'a	LSIPF	0.884	0.033	8.75	11.23
	RELAD-ANN	0.882	0.033	8.63	11.29
Kabul	LSIPF	0.807	0.031	7.06	10.14
	RELAD-ANN	0.879	0.020	4.52	8.02
Denver	LSIPF	0.615	0.055	13.07	15.68
	RELAD-ANN	0.805	0.039	9.12	11.14
Dushanbe	LSIPF	0.634	0.039	9.74	12.93
	RELAD-ANN	0.795	0.020	4.90	9.67
Granada	LSIPF	0.610	0.031	9.14	11.60
	RELAD-ANN	0.700	0.029	8.58	10.15

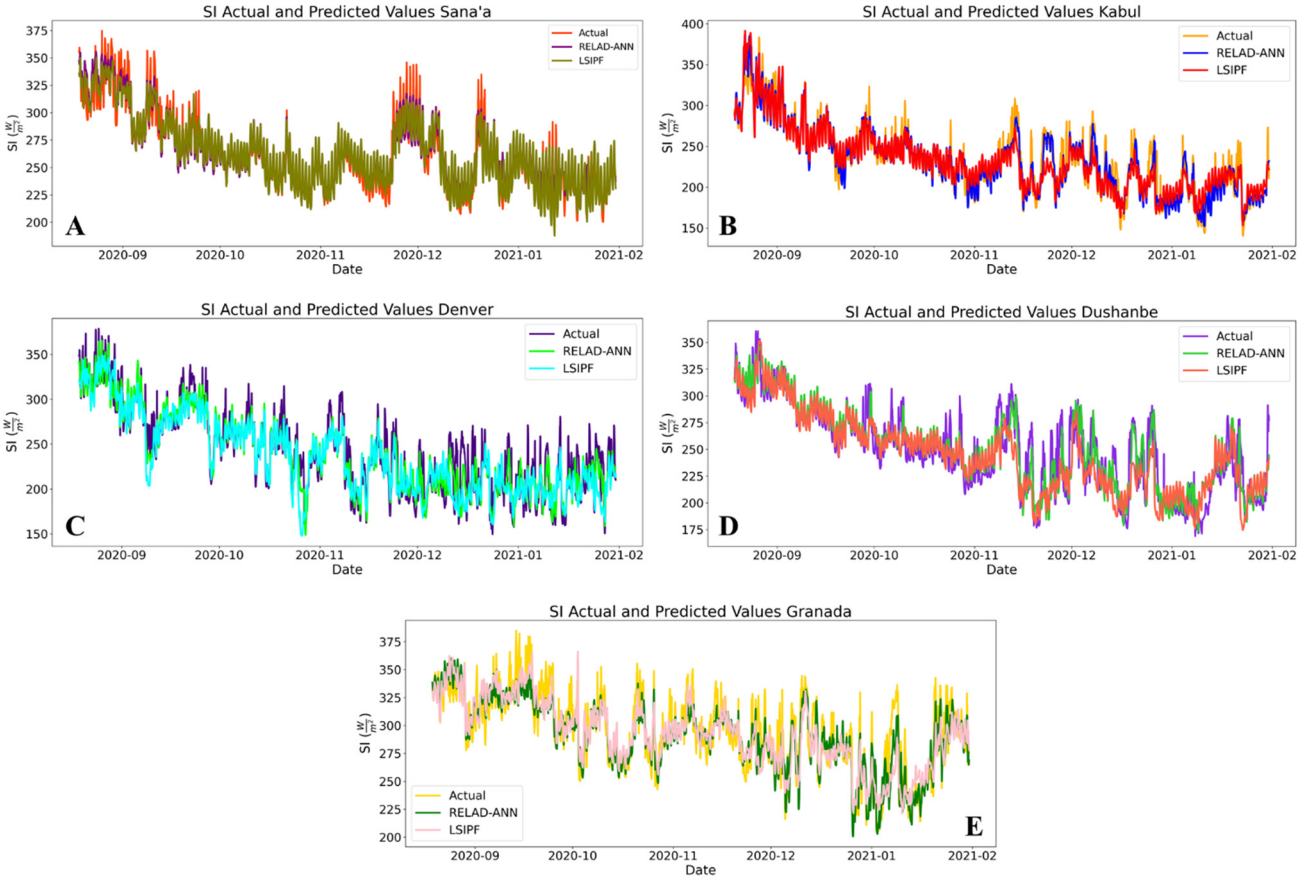

Figure 13. Proposed models' SI predictions over various geographical terrains.

Overall, the RELAD-ANN model consistently presents as the more robust and precise option for SI forecasting across the diverse range of global locations tested. It's generally higher R^2 values and lower error metrics across all sites demonstrate a reliable model performance, ensuring its suitability for deployment in various geographical and climatic conditions. This consistent performance, coupled with its simplicity, computational efficiency, and strong interpretability, solidifies the RELAD-ANN model as the preferred choice for SI forecasting.

5. Conclusions

We investigated the solar potential of Quetta (a city of Pakistan) and the dependency of SI on other parameters. Towards this end, two ML models RELAD-ANN and LSIPF have been generated using Python. To compare the two models, various parametric predictions have been made and validated through various statistical indicators. Moreover, two regression models SVR and Light GBM have been structured to check the effects of other parameters of SI forecasting. Based on the results reported in Section 3, several conclusions can be made below.

- The RELAD-ANN model outperforms the LSIPF in SI forecasting, excelling in predicting surface humidity and air temperature, despite some difficulty with high-speed wind occurrences. In contrast, the LSIPF shows precision in wind speed and air temperature but struggles with surface humidity, underscoring RELAD-ANN's broader effectiveness.
- Light GBM outperforms SVR with R^2 scores of 66.95% for air temperature and 73.10% for surface humidity, affirming its strength in capturing the pivotal environmental influences on SI, while wind speed remains a negligible predictor with an R^2 score of 3.92%.
- The study highlights strong correlations of SI with air temperature (0.811) and surface humidity (0.812), while wind speed shows a minimal correlation (0.064), indicating its lesser predictive significance.
- The RELAD-ANN model excelled in forecasting SI with superior accuracy, evidenced by an MAE of 8.20, a MAPE of 3.48%, and an R^2 of 0.935, while the LSIPF, MLP, LSTM, MLSTM, GRU, and CNN models displayed varying levels of predictive performance, with R^2 values ranging from 0.876 to 0.933 and MAEs between 8.22 and 11.96.
- Feature Analysis identifies 'air temperature', 'wind speed', and 'surface humidity' as key drivers, enhancing the RELAD-ANN model's forecasting accuracy to an R-squared value of over 0.92.
- Global Validation substantiates the model's efficacy in a variety of climatic conditions, achieving an average R^2 value exceeding 0.8 across multiple global locations, thereby affirming its universal applicability and consistent predictive performance.

In inference, we introduce a robust approach to predict SI and investigate its interdependencies with other parameters. The present models, i.e., RELAD-ANN and Light GBM regressor, offer accurate and reliable predictions of SI and its intertwined factors. This research holds significant value in the global landscape of renewable energy planning and management, equipping stakeholders with vital insights for optimizing solar energy harnessing. As a scalable and adaptable study, there is potential for its methodologies to be applied across varied geographic contexts and be enhanced by integrating additional parameters. Furthermore, the suggested models hold promise for real-time forecasting, paving the way for improved renewable energy system management worldwide.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the Research Supporting Project number (RSP2024R89), King Saud University, Riyadh, Saudi Arabia for funding this work.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization, M.F.H.; methodology, M.F.H.; software, M.F.H.; validation, M.S.N., X.L and J.S.; formal analysis, M.F.H.; investigation, M.F.H.; resources, M.M.; data curation, M.S.N.; writing—original draft preparation, M.F.H.; writing—review and editing, M.S.N, M.M, J.S, X.L and J.M.; visualization, M.F.H.; supervision, J.M.; project administration, J.M. All authors have read and agreed to the published version of the manuscript.

References

1. Guan Y, Lu H, Jiang Y, et al. (2021) Changes in global climate heterogeneity under the 21st century global warming. *Ecol Indic* 130: 108075. https://doi.org/10.1016/j.ecolind.2021.108075
2. Sohani A, Shahverdian MH, Sayyaadi H, et al. (2021) Energy and exergy analyses on seasonal comparative evaluation of water flow cooling for improving the performance of monocrystalline PV module in hot-arid climate. *Sustainability* 13: 6084. https://doi.org/10.3390/su13116084
3. Sahebi HK, Hoseinzadeh S, Ghadamian H, et al. (2021) Techno-economic analysis and new design of a photovoltaic power plant by a direct radiation amplification system. *Sustainability* 13: 11493. https://doi.org/10.3390/su132011493
4. Hoseinzadeh S, Ghasemi MH, Heyns S (2020) Application of hybrid systems in solution of low power generation at hot seasons for micro hydro systems. *Renewable Energy* 160: 323–332. https://doi.org/10.1016/j.renene.2020.06.149
5. Makkiabadi M, Hoseinzadeh S, Mohammadi M, et al. (2020) Energy feasibility of hybrid PV/wind systems with electricity generation assessment under Iran environment. *Appl Sol Energy* 56: 517–525. https://doi.org/10.3103/s0003701x20060079
6. Hannan MA, Al-Shetwi AQ, Ker PJ, et al. (2021) Impact of renewable energy utilization and artificial intelligence in achieving sustainable development goals. *Energy Rep* 7: 5359–5373. https://doi.org/10.1016/j.egyr.2021.08.172
7. Rafique MM, Rehman S (2017) National energy scenario of Pakistan—Current status, future alternatives, and institutional infrastructure: An overview. *Renewable Sustainable Energy Rev* 69: 156–167. https://doi.org/10.1016/j.rser.2016.11.057
8. Pikus M, Wąs J (2023) Using deep neural network methods for forecasting energy productivity based on comparison of simulation and DNN results for central Poland—Swietokrzyskie Voivodeship. *Energies* 16: 6632. https://doi.org/10.3390/en16186632
9. Rafique MM, Bahaidarah HMS, Anwar MK (2019) Enabling private sector investment in off-grid electrification for cleaner production: Optimum designing and achievable rate of unit electricity. *J Clean Prod* 206: 508–523. https://doi.org/10.1016/j.jclepro.2018.09.123

10. Sørensen ML, Nystrup P, Bjerregård MB, et al. (2023) Recent developments in multivariate wind and solar power forecasting. *Wiley Interdiscip Rev Energy Environ,* 12. https://doi.org/10.1002/wene.465
11. Wang H, Liu Y, Zhou B, et al. (2020) Taxonomy research of artificial intelligence for deterministic solar power forecasting. *Energy Convers Manage* 214: 112909. https://doi.org/10.1016/j.enconman.2020.112909
12. Sobri S, Koohi-Kamali S, Rahim NA (2018) Solar photovoltaic generation forecasting methods: A review. *Energy Convers Manage* 156: 459–497. https://doi.org/10.1016/j.enconman.2017.11.019
13. Mokarram M, Mokarram MJ, Gitizadeh M, et al. (2020) A novel optimal placing of solar farms utilizing multi-criteria decision-making (MCDA) and feature selection. *J Clean Prod* 261: 121098. https://doi.org/10.1016/j.jclepro.2020.121098
14. Cesar LB, Silva RAE, Callejo MÁM, et al. (2022) Review on Spatio-temporal solar forecasting methods driven by in Situ measurements or their combination with satellite and numerical weather prediction (NWP) estimates. *Energies* 15: 4341. https://doi.org/10.3390/EN15124341
15. Miller SD, Rogers MA, Haynes JM, et al. (2018) Short-term solar irradiance forecasting via satellite/model coupling. *Sol Energy* 168: 102–117. https://doi.org/10.1016/j.solener.2017.11.049
16. Hao Y, Tian C (2019) A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting. *Appl Energy* 238: 368–383. https://doi.org/10.1016/j.apenergy.2019.01.063
17. Murata A, Ohtake H, Oozeki T (2018) Modeling of uncertainty of solar irradiance forecasts on numerical weather predictions with the estimation of multiple confidence intervals. *Renewable Energy* 117: 193–201. https://doi.org/10.1016/j.renene.2017.10.043
18. Munkhammar J, van der Meer D, Widén J (2019) Probabilistic forecasting of high-resolution clear-sky index time-series using a Markov-chain mixture distribution model. *Sol Energy* 184: 688–695. https://doi.org/10.1016/j.solener.2019.04.014
19. Halabi LM, Mekhilef S, Hossain M (2018) Performance evaluation of hybrid adaptive neuro-fuzzy inference system models for predicting monthly global solar radiation. *Appl Energy* 213: 247–261. https://doi.org/10.1016/j.apenergy.2018.01.035
20. Dong J, Olama MM, Kuruganti T, et al. (2020) Novel stochastic methods to predict short-term solar radiation and photovoltaic power. *Renewable Energy* 145: 333–346. https://doi.org/10.1016/j.renene.2019.05.073
21. Ahmad T, Zhang D, Huang C (2021) Methodological framework for short-and medium-term energy, solar and wind power forecasting with stochastic-based machine learning approach to monetary and energy policy applications. *Energy* 231: 120911. https://doi.org/10.1016/j.energy.2021.120911
22. Ağbulut Ü, Gürel AE, Biçen Y (2021) Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable Sustainable Energy Rev* 135: 110114. https://doi.org/10.1016/j.rser.2020.110114
23. Jumin E, Basaruddin FB, Yusoff YBM, et al. (2021) Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia. *Environ Sci Pollut Res* 28: 26571–26583. https://doi.org/10.1007/s11356-021-12435-6

24. Benali L, Notton G, Fouilloy A, et al. (2019) Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable Energy* 132: 871–884. https://doi.org/10.1016/j.renene.2018.08.044
25. Zendehboudi A, Baseer MA, Saidur R (2018) Application of support vector machine models for forecasting solar and wind energy resources: A review. *J Clean Prod* 199: 272–285. https://doi.org/10.1016/j.jclepro.2018.07.164
26. André PS, Dias LMS, Correia SFH, et al. (2024) Artificial neural networks for predicting optical conversion efficiency in luminescent solar concentrators. *Sol Energy* 268: 112290. https://doi.org/10.1016/j.solener.2023.112290
27. Girimurugan R, Selvaraju P, Jeevanandam P, et al. (2023) Application of deep learning to the prediction of solar irradiance through missing data. *Int J Photoenergy* 2023: 4717110. https://doi.org/10.1155/2023/4717110
28. Noman AM, Khan H, Sher HA, et al. (2023) Scaled conjugate gradient artificial neural network-based ripple current correlation MPPT algorithms for PV system. *Int J Photoenergy* 2023: 8891052. https://doi.org/10.1155/2023/8891052
29. Ricci L, Papurello D (2023) A prediction model for energy production in a solar concentrator using artificial neural networks. *Int J Energy Res* 2023: 9196506. https://doi.org/10.1155/2023/9196506
30. Konstantinou M, Peratikou S, Charalambides AG (2021) Solar photovoltaic forecasting of power output using LSTM networks. *Atmosphere* 12: 124. https://doi.org/10.3390/atmos12010124
31. Pan C, Tan J, Feng D (2021) Prediction intervals estimation of solar generation based on gated recurrent unit and kernel density estimation. *Neurocomputing* 453: 552–562. https://doi.org/10.1016/j.neucom.2020.10.027
32. Feng C, Zhang J, Zhang W, et al. (2022) Convolutional neural networks for intra-hour solar forecasting based on sky image sequences. *Appl Energy* 310: 118438. https://doi.org/10.1016/j.apenergy.2021.118438
33. Colak HE, Memisoglu T, Gercek Y (2020) Optimal site selection for solar photovoltaic (PV) power plants using GIS and AHP: A case study of Malatya Province, Turkey. *Renewable Energy* 149: 565–576. https://doi.org/10.1016/j.renene.2019.12.078
34. Mousapour Mamoudan M, Ostadi A, Pourkhodabakhsh N, et al. (2023) Hybrid neural network-based metaheuristics for prediction of financial markets: A case study on global gold market. *J Comput Des Eng* 10: 1110–1125. https://doi.org/10.1093/jcde/qwad039
35. Gholizadeh H, Fathollahi-Fard AM, Fazlollahtabar H, et al. (2022) Fuzzy data-driven scenario-based robust data envelopment analysis for prediction and optimization of an electrical discharge machine's parameters. *Expert Syst Appl* 193: 116419. https://doi.org/10.1016/j.eswa.2021.116419
36. Ghazikhani A, Babaeian I, Gheibi M, et al. (2022) A smart post-processing system for forecasting the climate precipitation based on machine learning computations. *Sustainability* 14: 6624. https://doi.org/10.3390/su14116624
37. Han Z, Zhao J, Leung H, et al. (2021) A review of deep learning models for time series prediction. *IEEE Sens J* 21: 7833–7848. https://doi.org/10.1109/jsen.2019.2923982
38. Ghimire S, Deo RC, Raj N, et al. (2019) Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Appl Energy* 253: 113541. https://doi.org/10.1016/j.apenergy.2019.113541

39. Zang H, Liu L, Sun L, et al. (2020) Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. *Renewable Energy* 160: 26–41. https://doi.org/10.1016/j.renene.2020.05.150
40. Rathore N, Rathore P, Basak A, et al. (2021) Multi Scale Graph Wavenet for wind speed forecasting. *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 4047–4053. https://doi.org/10.1109/bigdata52589.2021.9671624
41. Shaikh AK, Nazir A, Khalique N, et al. (2023) A new approach to seasonal energy consumption forecasting using temporal convolutional networks. *Results Eng* 19: 101296. https://doi.org/10.1016/j.rineng.2023.101296
42. Qu J, Qian Z, Pei Y (2021) Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. *Energy* 232: 120996. https://doi.org/10.1016/j.energy.2021.120996
43. Zhan C, Zhang X, Yuan J, et al. (2024) A hybrid approach for low-carbon transportation system analysis: integrating CRITIC-DEMATEL and deep learning features. *Int J Environ Sci Technol* 21: 791–804. https://doi.org/10.1007/s13762-023-04995-6
44. Kumari P, Toshniwal D (2021) Deep learning models for solar irradiance forecasting: A comprehensive review. *J Clean Prod* 318: 128566. https://doi.org/10.1016/j.jclepro.2021.128566
45. Akram MW, Li G, Jin Y, et al. (2019) CNN based automatic detection of photovoltaic cell defects in electroluminescence images. *Energy* 189: 116319. https://doi.org/10.1016/j.energy.2019.116319
46. Halton C (2023) Predictive analytics: Definition, model types, and uses, 2021. *Investopedia* Available from: https://www.investopedia.com/terms/p/predictive-analytics.asp#:~:text=the most common predictive models,deep learning methods and technologies.
47. Manju S, Sandeep M (2019) Prediction and performance assessment of global solar radiation in Indian cities: A comparison of satellite and surface measured data. *J Clean Prod* 230: 116–128. https://doi.org/10.1016/j.jclepro.2019.05.108
48. Ahmad S, Parvez M, Khan TA, et al. (2022) A hybrid approach using AHP-TOPSIS methods for ranking of soft computing techniques based on their attributes for prediction of solar radiation. *Environ Challenges* 9: 100634. https://doi.org/10.1016/j.envc.2022.100634
49. Ağbulut Ü, Gürel AE, Biçen Y (2021) Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable Sustainable Energy Rev* 135: 110114. https://doi.org/10.1016/j.rser.2020.110114
50. Islam S, Roy NK (2023) Renewable's integration into power systems through intelligent techniques: Implementation procedures, key features, and performance evaluation. *Energy Rep* 9: 6063–6087. https://doi.org/10.1016/j.egyr.2023.05.063
51. Farooqui SZ (2014) Prospects of renewables penetration in the energy mix of Pakistan. *Renewable Sustainable Energy Rev* 29: 693–700. https://doi.org/10.1016/j.rser.2013.08.083
52. Government of Pakistan FD (2022) Pakistan Economic Survey 2021-22. Available from: https://www.finance.gov.pk/survey_2022.html.
53. Đukanović M, Kašćelan L, Vuković S, et al. (2023) A machine learning approach for time series forecasting with application to debt risk of the Montenegrin electricity industry. *Energy Rep* 9: 362–369. https://doi.org/10.1016/j.egyr.2023.05.240

54. Irfan M, Zhao ZY, Mukeshimana MC, et al. (2019) Wind energy development in South Asia: Status, potential and policies. *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET),* Sukkur, Pakistan, 1–6. https://doi.org/10.1109/icomet.2019.8673484
55. Energy system of Asia Pacific. *International Energy Agency.* Available from: https://www.iea.org/regions/asia-pacific.
56. Climate change. *International Energy Agency.* Available from: https://www.iea.org/.
57. Rafique MM, Rehman S (2017) National energy scenario of Pakistan—Current status, future alternatives, and institutional infrastructure: An overview. *Renewable Sustainable Energy Rev.* 69: 156–167. https://doi.org/10.1016/j.rser.2016.11.057
58. Awan U, Knight I (2020) Domestic sector energy demand and prediction models for Punjab Pakistan. *J Building Eng* 32: 101790. https://doi.org/10.1016/j.jobe.2020.101790
59. Muhammad F, Waleed Raza M, Khan S, et al. (2017) Different solar potential co-ordinates of Pakistan. *Innovative Energy Res* 6: 1–8. https://doi.org/10.4172/2576-1463.1000173
60. Farooq M, Shakoor A (2013) Severe energy crises and solar thermal energy as a viable option for Pakistan. *J Renewable Sustainable Energy* 5: 013104. https://doi.org/10.1063/1.4772637
61. Shabbir N, Usman M, Jawad M, et al. (2020) Economic analysis and impact on national grid by domestic photovoltaic system installations in Pakistan. *Renewable Energy* 153: 509–521. https://doi.org/10.1016/j.renene.2020.01.114
62. Global solar atlas. Available from: https://globalsolaratlas.info/map.
63. CDPC, *Department PM Climate Records Quetta*. Available from: https://cdpc.pmd.gov.pk/.
64. JRC Photovoltaic Geographical Information System (PVGIS)—European Commission. Available from: https://re.jrc.ec.europa.eu/pvg_tools/en/#PVP/.
65. Earthdata. *NASA*. Available from: https://www.earthdata.nasa.gov/.
66. Welcome to Colaboratory—Google Colaboratory. Available from: https://colab.research.google.com/.
67. Emmanuel T, Maupong T, Mpoeleng D, et al. (2021) A survey on missing data in machine learning. *J Big Data* 8: 1–37. https://doi.org/10.1186/S40537-021-00516-9
68. Ackerman S, Farchi E, Raz O, et al. (2020) Detection of data drift and outliers affecting machine learning model performance over time. *arXiv* In: JSM Proceedings, Nonparametric Statistics Section, 20202. Philadelphia, PA: American Statistical Association, 144–160. https://doi.org/10.48550/arXiv.2012.09258
69. Khadka N (2019) General machine learning practices using Python. Available from: https://www.theseus.fi/bitstream/handle/10024/226305/Khadka_Nibesh.pdf?sequence=2.
70. Pereira Barata A, Takes FW, Van Den Herik HJ, et al. (2019) Imputation methods outperform missing-indicator for data missing completely at random. *2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China,* 407–414. https://doi.org/10.1109/ICDMW.2019.00066
71. Wu P, Zhang Q, Wang G, et al. (2023) Dynamic feature selection combining standard deviation and interaction information. *Int J Mach Learn Cyber* 14: 1407–1426. https://doi.org/10.1007/S13042-022-01706-4
72. Begum S, Meraj S, Shetty BS (2023) Successful data mining: With dimension reduction. *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics*, 11–22. https://doi.org/10.2991/978-94-6463-136-4_3

73. Li B, Wu F, Lim S-N, et al. (2021) On feature normalization and data augmentation. *IEEE/CVF Conference on Computer Vision and Pattern,* 12383–12392. https://doi.org/10.48550/arXiv.2002.11102
74. Ramirez-Vergara J, Bosman LB, Leon-Salas WD, et al. (2021) Ambient temperature and solar irradiance forecasting prediction horizon sensitivity analysis. *Machine Learning Appl* 6: 100128. https://doi.org/10.1016/j.mlwa.2021.100128
75. Verbois H, Huva R, Rusydi A, et al. (2018) Solar irradiance forecasting in the tropics using numerical weather prediction and statistical learning. *Sol Energy* 162: 265–277. https://doi.org/10.1016/j.solener.2018.01.007
76. Ssekulima EB, Anwar MB, Al Hinai A, et al. (2016) Wind speed and solar irradiance forecasting techniques for enhanced renewable energy integration with the grid: A review. *IET Renewable Power Generation* 10: 885–989. https://doi.org/10.1049/iet-rpg.2015.0477
77. Kumar N, Sinha UK, Sharma SP, et al. (2017) Prediction of daily global solar radiation using Neural Networks with improved gain factors and RBF Networks. *Int J Renewable Energy Res* 7: 1235–1244. https://doi.org/10.20508/ijrer.v7i3.5988.g7156
78. Siva Krishna Rao KDV, Premalatha M, Naveen C (2018) Models for forecasting monthly mean daily global solar radiation from in-situ measurements: Application in Tropical Climate, India. *Urban Clim* 24: 921–939. https://doi.org/10.1016/j.uclim.2017.11.004
79. Yıldırım HB, Çelik Ö, Teke A, et al. (2018) Estimating daily Global solar radiation with graphical user interface in Eastern Mediterranean region of Turkey. *Renewable Sustainable Energy Rev* 82: 1528–1537. https://doi.org/10.1016/j.rser.2017.06.030
80. Mohaideen Abdul Kadhar K, Anand G (2021) Basics of Python programming. *Data Sci Raspberry Pi*, 13–47. https://doi.org/10.1007/978-1-4842-6825-4_2
81. Gholizadeh S (2022) Top popular Python libraries in research. *J Robot Auto Res* 3: 142–145. http://dx.doi.org/10.33140/jrar.03.02.02
82. Stančin I, Jović A (2019) An overview and comparison of free Python libraries for data mining and big data analysis. *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 977–982. https://doi.org/10.23919/mipro.2019.8757088
83. Voigtlaender F (2023) The universal approximation theorem for complex-valued neural networks. *Appl Comput Harmon Anal* 64: 33–61. https://doi.org/10.1016/j.acha.2022.12.002
84. Winkler DA, Le TC (2017) Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. *Mol Inf,* 36. https://doi.org/10.1002/minf.201600118
85. Lu Y, Lu J (2020) A universal approximation theorem of Deep Neural Networks for expressing probability distributions. *arXiv.* https://doi.org/10.48550/arXiv.2004.08867
86. Dubey SR, Singh SK, Chaudhuri BB (2022) Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* 503: 92–108. https://doi.org/10.1016/j.neucom.2022.06.111
87. Tato A, Nkambou R (2018) Improving Adam optimizer. *ICLR 2018 Workshop*. Available from: https://openreview.net/forum?id=HJfpZq1DM.
88. Toh SC, Lai SH, Mirzaei M, et al. (2023) Sequential data processing for IMERG satellite rainfall comparison and improvement using LSTM and ADAM optimizer. *Appl Sci* 13: 7237. https://doi.org/10.3390/app13127237

89. Amose J, Manimegalai P, Narmatha C, et al. (2022) Comparative performance analysis of Kernel functions in Support Vector Machines in the diagnosis of pneumonia using lung sounds. *Proceedings of 2022 2nd International Conference on Computing and Information Technology, ICCIT 2022*, 320–324. https://doi.org/10.1109/iccit52419.2022.9711608
90. Karyawati AE, Wijaya KDY, Supriana IW, et al. (2023) A comparison of different Kernel functions of SVM classification method for spam detection. *JITK* 8: 91–97. https://doi.org/10.33480/jitk.v8i2.2463
91. Munir MA, Khattak A, Imran K, et al. (2019) Solar PV generation forecast model based on the most effective weather parameters. *1st International Conference on Electrical, Communication and Computer Engineering, ICECCE 2019,* 24–25. https://doi.org/10.1109/icecce47252.2019.8940664
92. Wang F, Mi Z, Su S, et al. (2012) Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. *Energies* 5: 1355–1370. https://doi.org/10.3390/en5051355
93. Kashyap Y, Bansal A, Sao AK (2015) Solar radiation forecasting with multiple parameters neural networks. *Renewable Sustainable Energy Rev* 49: 825–835. https://doi.org/10.1016/j.rser.2015.04.077
94. Sayad S. Support Vector Machine-Regression (SVR). *An introduction to data science.* Available from: http://www.saedsayad.com/support_vector_machine_reg.htm.
95. Lu Y, Roychowdhury V (2008) Parallel randomized sampling for support vector machine (SVM) and support vector regression (SVR). *Knowl Inf Syst* 14: 233–247. https://doi.org/10.1007/s10115-007-0082-6
96. Kleynhans T, Montanaro M, Gerace A, et al. (2017) Predicting top-of-atmosphere thermal radiance using MERRA-2 atmospheric data with deep learning. *Remote Sens* 9: 1133. https://doi.org/10.3390/rs9111133
97. Obviously AI: data science without code (2022) *Obviously AI Inc*. Available from: https://app.obviously.ai/predict.
98. Data Science, what is Light GBM? Available from: https://datascience.eu/machine-learning/1-what-is-light-gbm/.
99. Mandot P (2017) What is LightGBM, how to implement it? How to fine tune the parameters? *Medium*. Available from: https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc.
100. Gueymard CA (2014) A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. *Renewable Sustainable Energy Rev* 39: 1024–1034. https://doi.org/10.1016/j.rser.2014.07.117
101. De Paiva GM, Pimentel SP, Alvarenga BP, et al. (2020) Multiple site intraday solar irradiance forecasting by machine learning algorithms: MGGP and MLP neural networks. *Energies* 13: 3005. https://doi.org/10.3390/en13113005
102. Yildirim A, Bilgili M, Ozbek A (2023) One-hour-ahead solar radiation forecasting by MLP, LSTM, and ANFIS approaches. *Meteorology Atmospheric Physics,* 135. https://doi.org/10.1007/s00703-022-00946-x
103. Huang X, Li Q, Tai Y, et al. (2021) Hybrid deep neural model for hourly solar irradiance forecasting. *Renewable Energy* 171: 1041–1060. https://doi.org/10.1016/j.renene.2021.02.161

104. Mellit A, Pavan AM, Lughi V (2021) Deep learning neural networks for short-term photovoltaic power forecasting. *Renewable Energy* 172: 276–288. https://doi.org/10.1016/j.renene.2021.02.166
105. Bhatt A, Ongsakul W, Nimal Madhu M, et al. (2022) Sliding window approach with first-order differencing for very short-term solar irradiance forecasting using deep learning models. *Sustainable Energy Technol Assess,* 50. https://doi.org/10.1016/j.seta.2021.101864
106. Wang J, Zhong H, Lai X, et al. (2019) Exploring key weather factors from analytical modeling toward improved solar power forecasting. *IEEE Trans Smart Grid* 10: 1417–1427. https://doi.org/10.1109/tsg.2017.2766022
107. Basak SC, Vracko MG (2020) Parsimony principle and its proper use/application in computer-assisted drug design and QSAR. *Curr Comput Aided Drug Des* 16: 1–5. https://doi.org/10.2174/1573409916012001061228 54
108. Almekhlafi MAA (2018) Justification of the advisability of using solar energy for the example of the Yemen republic. *National University of Civil Defence of Ukraine*, 41–50. Available from: http://repositsc.nuczu.edu.ua/handle/123456789/7224.
109. Naseri M, Hussaini MS, Iqbal MW, et al. (2021) Spatial modeling of solar photovoltaic power plant in Kabul, Afghanistan. *J Mt Sci* 18: 3291–3305. https://doi.org/10.1007/S11629-021-7035-5
110. Elizabeth Michael N, Hasan S, Al-Durra A, et al. (2022) Short-term solar irradiance forecasting based on a novel Bayesian optimized deep long short-term memory neural network. *Appl Energy* 324: 119727. https://doi.org/10.1016/j.apenergy.2022.119727
111. Safaraliev MK, Odinaev IN, Ahyoev JS, et al. (2020) Energy potential estimation of the region's solar radiation using a solar tracker. *Appl Sol Energy* 56: 270–275. https://doi.org/10.3103/s0003701x20040118
112. Rodríguez-Benítez FJ, Arbizu-Barrena C, Huertas-Tato J, et al. (2020) A short-term solar radiation forecasting system for the Iberian Peninsula. Part 1: Models description and performance assessment. *Sol Energy* 195: 396–412. https://doi.org/10.1016/j.solener.2019.11.028

AIMS Press