Applied Soft Computing 112 (2021) 107768

Contents lists available at ScienceDirect

Applied Soft Computing

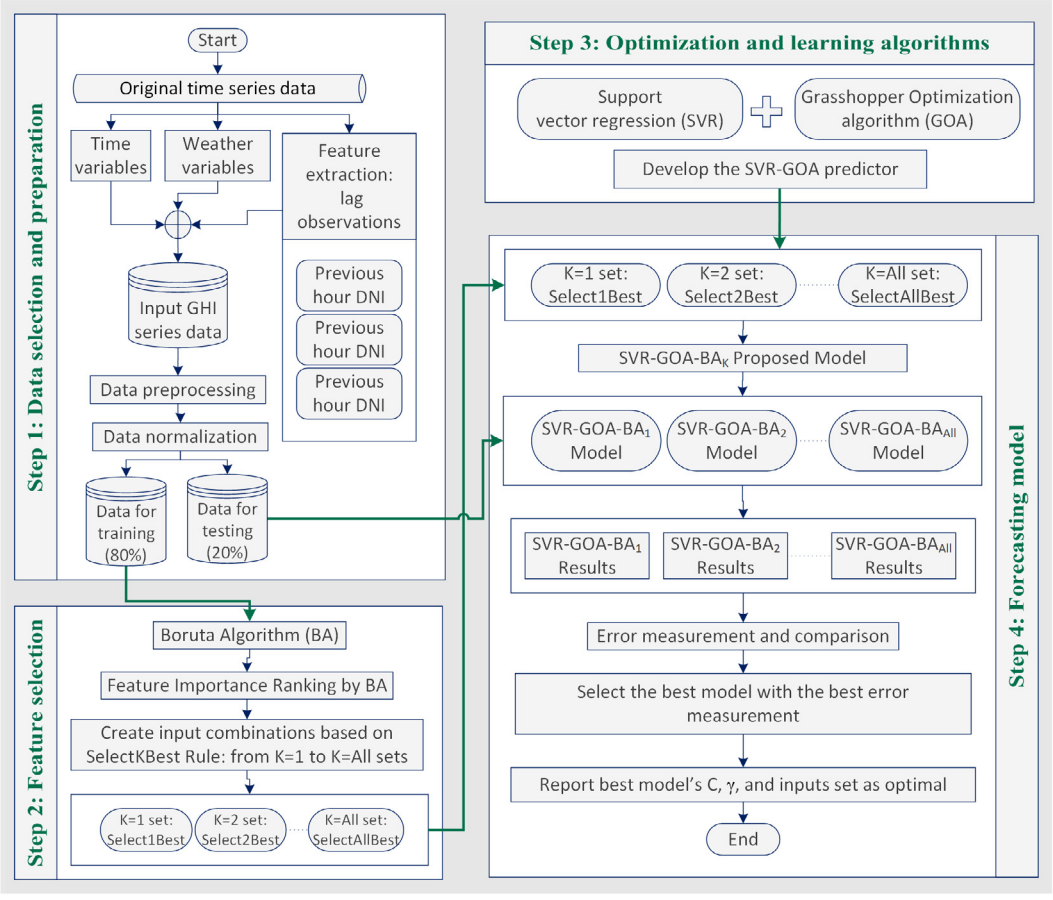
journal homepage: www.elsevier.com/locate/asoc

Global solar radiation prediction: Application of novel hybrid data-driven model

Massoud Alrashidi [a,b,*], Musaed Alrashidi [a,c], Saifur Rahman [a]

[a] Bradley Department of Electrical and Computer Engineering, Advanced Research Institute, Virginia Tech, VA 22203, USA
[b] Department of Electrical Engineering, College of Engineering, Qassim University, Onaizah 56452, Saudi Arabia
[c] Department of Electrical Engineering, College of Engineering, Qassim University, Buraidah 51452, Saudi Arabia

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 11 November 2020
Received in revised form 29 June 2021

ABSTRACT

One of the significant prerequisites for harvesting solar energy is precise global solar radiation (GHI) forecasts. However, variability and uncertainty are inherent characteristics of solar radiation. It is challenging to show better generalization using current data analysis approaches. Thus, this

* Corresponding author at: Bradley Department of Electrical and Computer Engineering, Advanced Research Institute, Virginia Tech, VA 22203, USA.
E-mail address: rashidi1@vt.edu (M. Alrashidi).

https://doi.org/10.1016/j.asoc.2021.107768

Accepted 27 July 2021
Available online 31 July 2021

Keywords:
Boruta algorithm
Support vector regression
Grasshopper optimization algorithm
Hyperparameter
Global solar radiation prediction
Feature selection

research presents a new intelligence framework by hybridizing Support Vector Regression (SVR) with the Grasshopper Optimization Algorithm (GOA) and the Boruta-based feature selection algorithm (BA) for forecasting GHI values at different sites of Saudi Arabia. Interestingly, the most significant distinction that differentiates this proposed prediction model (SVR-GOA-BA_K) from other models is that the GOA is automatically employed to search for optimal SVR's hyperparameters. In contrast, these hyperparameters are chosen randomly and manually in conventional models. Consequently, the contribution helps save time, reduce cost, and avoid the possibility of models' overfitting or underfitting caused by random and manual selection. A diversity of statistical measures has justified the proposed model's effectiveness and superiority. In terms of mean absolute percentage error (MAPE), the proposed model outperformed the standalone SVR models by 32.15–39.69% at different study sites. In tuning the SVR's parameters, GOA outperforms popular optimization algorithms. All the simulation test results demonstrate the superiority of the proposed model. Hence, the proposed approach provides a foundation for precise solar radiation forecasting, which can aid in the growth of renewable-energy-based technologies.

1. Introduction

The increasing quest for alternative energy sources away from fossil fuels results from four trends associated with fossils: (a) their depletion; (b) limited resources, which lead to rising prices; (c) the environmental dilemma triggered by greenhouse gases; and (d) the emergence of renewable, ecological, sustainable, or natural consumer culture [1]. Electricity suppliers can adopt the notion of sustainable supplies by the employment of green renewables, particularly solar energy, which is nevertheless distinguished by a high degree of uncertainty in availability and production. However, sudden variations in solar electricity output are among the terrible impacts of momentary changes in climatic circumstances. Indeed, renewables' stochastic and intermittent existence could impede their efficient usage by power suppliers [2]. It could also stifle the growth of renewable energy technologies like photovoltaics, wind turbines, and concentrating solar power plants in the future [3]. Also, using solar-based electricity in a power grid remains troublesome at elevated scales for various nuanced but now well-established reasons [4]. The supply does not match the demand. Changes to power grid operations are expected to manage solar power variability and unpredictability, increasing demand for ancillary services and energy balancing in general [5]. Therefore, the costly expenditures associated with such adjustments and requirements earnestly impact renewables' economic viability.

Multiple feasible options can help mitigate technical and practical problems triggered by the short-term uncertainty, until seven days ahead, in the solar power supply. For example, raising the level of demand-side engagement, raising the volume of collaboration to manage allocations, introducing smart grids, and installing more adaptable – and often more expensive – energy storage technologies [6]. However, precise solar irradiance forecasting is among the most functional and cost-effective strategies for penetrating higher solar power levels and making optimal decisions on the planning of renewable energy projects [7]. Balancing agencies should use such predictions to run power grids more reliably and effectively. Forecasts for the future, on the other hand, need more complex models. Due to the non-linearity and difficulty of modeling the solar radiation series, this is considered a difficult challenge [8]. Hence, solar radiation prediction has been a hot subject in energy research, and several new methods have been proposed to strengthen the solar radiation predictive modeling literature.

Information on global solar radiation (GHI) at any location is essential for various requirements, including climatology, hydrology, public health, and clean energy utilization [9]. The latter application is what matters in this research. By focusing on forecast time horizons, solar forecasting reduces the impact of variability and uncertainty associated with solar energy. About practical application, Fig. 1 depicts various prediction scales and related functions in solar-based power systems. For real-time battery storage management, very short-term prediction is deemed necessary [10]. Short-term prediction is vital for decision-making tasks such as unit commitment problems [11]. Medium-term forecasting is valuable for maintenance schedules and power units running [12]. Long-term prediction is essential for strategic power grid operations planning [13]. Diverse predictive methodologies for GHI forecasting have been established and are discussed in Section 1.1.

1.1. Literature review

Data-based techniques have become typical in solar radiation prediction in recent years with the advancements of data-mining methods. Specifically, different data-mining methods are used to forecast the time series of solar radiation. Machine-learning models have recently received a lot of attention due to their high accuracy. Various machine learning (ML) models, such as the artificial neural network (ANN), regression decision tree (DT), genetic programming (GA), SVR, data mining, and fuzzy logic, have been developed for GHI forecasting [14]. Besides, Alfadda et al. have proposed an hour-ahead solar irradiance prediction model concerning desert areas [15]. This model used aerosol optical depth and angstrom exponent ground-based data to capture the dust impact in such areas. The proposed model was tested and validated using four ML algorithms: multilayer perceptron (MLP), SVR, KNN, and DT. The research study concludes that the dust measures significantly enhanced the model's accuracy, where the MLP model has the best predictive capacity. Deep learning, a subfield of machine learning, has been booming lately because of rapid advancements of information technology in hardware and software. For example, the hybrid methods of convolutional neural network (CNN) and long short-term memory (LSTM) algorithms (C-LSTM) were developed by Ghimire et al. [16]. The models were trained and validated using 30-minute and hourly datasets from 2006 to 2018 in Alice Springs, Austria. Compared to CNN, LSTM, MLP, deep neural nets (DNN), and DT, the C-LSTM hybrid methods performed the highest, with predictive errors at 70% and under $\pm$ 10 W/m^2. Utilizing data series from 2017 to 2019, Huynh et al. employed the deep learning of LSTM for forecasting GHI from one minute to 30 min ahead in Bac Ninh province, Vietnam [17]. When LSTM was compared to the autoregressive integrated moving average (ARIMA), SVM, MLP, and CNN, it was found that LSTM had the maximum efficiency, with R-value greater than 0.9. Aslam et al. came to a similar conclusion [18]. Besides the algorithms mentioned above, researchers

Abbreviations

GHI	Global horizontal irradiance
BA	Boruta algorithm
SVR	Support vector regression
GOA	Grasshopper optimization algorithm
ML	Machine learning
KNN	K-nearest-neighbors
ANN	Artificial neural networks
RF	Random forest
MLP	Multilayer perceptron
DT	Decision tree regression
BPNN	Backpropagation neural network
DHI	Diffuse solar radiation
PSO	Particle swarm optimization
GA	Genetic algorithm
ANFIS	Adaptive neural fuzzy inference system
GMDH	Group method of data handling
CNN	Convolutional neural network
LSTM	Long short-term memory
DNN	Deep neural nets
ARIMA	Auto-regressive integrated moving average
ELM	Extreme learning machine
C	Regularization parameter
γ	Width of the radial basis kernel function
FSP	Feature selection process
GGA	Grouping genetic algorithm
CRO	Coral reefs optimization algorithm
SVR-GOA-BA_K	The proposed model in this paper
K	The number of input features in building the proposed model
K.A.CARE	King Abdullah City for Atomic and Renewable Energy
RBF	Radial basis function
AA	Azimuth angle
SZA	Solar zenith angle
PrevHourDNI	Direct normal irradiance at the previous hour
PrevHourDHI	Diffuse horizontal irradiance at the previous hour
PrevHourGHI	Global horizontal irradiance at the previous hour
RMSE	Root mean square error
nRMSE	Normalized root mean square error
R^2	Goodness of fit
MAPE	Mean absolute percentage error
MAE	Mean absolute error
nMAE	Normalized mean absolute error

suggested a slew of others. Autoregressive (AR) [19], ARIMA [20], multivariate adaptive regression splines (MARS) [21], and the logistic model [22] are some of them. In general, these methods outperformed empirical alternatives in terms of precision. Compared to MLP, kernels-based algorithms, tree-based algorithms, they had less computational complexity. However, they are often coupled with other models to improve their efficiency.

Only a few properties of the solar radiation time series, nevertheless, can be identified by single data-mining methods. Thus, hybrid models, which use multi-data mining algorithms, have been embraced by researchers to boost short-term solar forecasting efficiency. At a station in Kuala Terengganu, Malaysia, Halabi et al. developed conventional and hybrid ANFIS models by combining ANFIS with particle swarm optimization (PSO), GA, and differential evolution algorithm (DE) to forecast monthly GHI [23]. The study used various climatic variables such as maximum and minimum air temperature, rainfall, clearness index, and sunshine duration. The findings concluded that the hybrid ANFIS-PSO model outperforms the other models in forecasting GHI. Also, Dong et al. primarily have constructed a new predictive approach based on the deep-learning CNN algorithm and afterward applied a chaotic hybrid (GA-PSO) method to optimize CNN's network's parameters [24]. The study reveals how critical the chaotic hybrid algorithm is in mitigating the approach's imperfect efficiency. On the other hand, integrating ML algorithms with suitable decomposition methods impacts the GHI value forecasting precision. Generally, wavelet transfer (WT), empirical model decomposition (EMD), and ensemble empirical model decomposition (EEMD) are the three most commonly used decomposition methods in the GHI prediction scope. For 1-hour ahead GHI prediction, Monjoly et al. compared the three standard algorithms [25]. A traditional ANN and a hybrid model were used to test the three decomposition models' applicability. The research revealed that, after the decomposition, the predictive performance was considerably enhanced, particularly for WT. The analysis used datasets collected between 2012 to 2013 in Guadeloupe Island, France. The ANN model had an RMSE of 25.86%, while the combined method with EMD, EEMD, and WT had a decreasing RMSE of 16.91%, 14.06%, and 7.86%, respectively.

However, in solving complex nonlinear engineering problems, SVR, a kernel-based algorithm, has been proven to have higher predictive accuracy and faster speed in dealing with such issues [26]. It is considered reliable and robust, where a compact description of the learned model can be provided, which is another advantage needed by some researchers. Several researchers have discovered that SVR is more effective at forecasting GHI than other methods [27]. Using data from various locations, references [28–31] constructed SVR-based models to forecast GHI and inferred that the SVM performed satisfactorily for GHI prediction. Hassan et al. evaluated the group method of data handling (GBDT), random forest (RF), ANN, DT, and SVR algorithms finding that the SVR had the best accuracy [32]. Moreover, Quej et al. have developed predictive models utilizing various soft-computing methods to forecast solar radiation in a warm sub-humid climate [33]. The study has better predictive performance from SVR than ANFIS and ANN models. The SVR model achieves an R-value of 0.8209 while ANFIS and ANN score 0.8024 and 0.8012, respectively. Interestingly, for three different locations in Nigeria, Olatomiwa et al. have proposed a hybrid approach of the SVM, whose hyperparameters were optimized by the firefly optimization algorithm to predict the monthly average GHI [34]. In this study, the hybrid model's predictive effectiveness was compared to ANN and GA models. The comparison findings conclude that the designed model is more successful in predicting GHI values, achieving MAPE of 11.52%. Due to its capacity to grasp the uncertainty linked to time series data, SVR was the most accurate for GHI prediction in a comparative study of different MI algorithms in GHI forecasting [32]. However, the tuning of the SVR's hyperparameters is the model's major flaw [26]. The first parameter is a nonzero constant known as the regularization parameter (C). The C is the tuning parameter between two objectives: models' complexity and the model's needed predictive efficacy during the forecasting models' training phase. Also, the parameter of the kernel function (γ) is the second parameter. Thus, the suitable selection of these two hyperparameters is

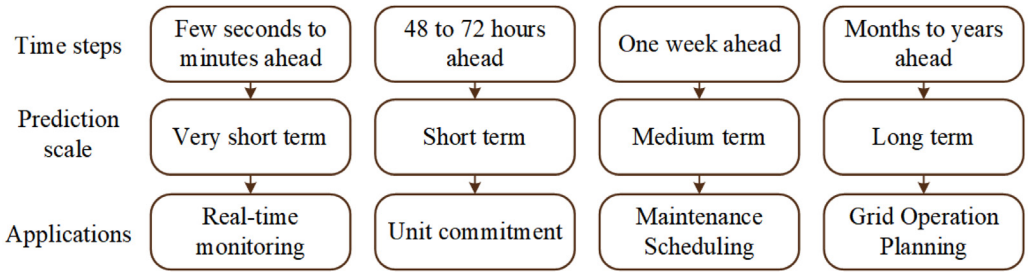


Fig. 1. Prediction scales and used applications.

critically essential for SVR to perform well. As a result, in previous studies, the conventional SVR model was combined with various optimization algorithms to enhance its efficiency. Even though the current hybrid SVR model's accuracy is good, the forecasting efficiency still needs to be improved, given the significance of GHI measurement accuracy. Hence, the employment of recent and powerful optimization techniques, especially metaheuristic optimization algorithms, is a crucial key in developing highly accurate SVR-based predictive models [35].

In ML-based applications, the feature selection process (FSP) is a crucial task as irrelevant features, used in training phases of several predictive systems, can negatively influence systems' cost, computational runtime, and overfitting problems [36]. Also, irrelevant features make predictive models' efficiency in generalization much lower [37]. Several different algorithms can be applied to address the FSP. Generally, such algorithms can be split into two distinct classes: wrapper and filter techniques [38]. Salcedo et al. have concluded that wrapper techniques are primarily used in renewable energy applications, compared to filter counterparts [39]. The methods of Relief-F, Monte Carlo uninformative variable elimination, random frog, and LSA were compared by Almaraashi [40]. In this study, the MLP-based models were trained and tested with multiple variables from eight regions in Saudi Arabia for daily GHI forecasting. He discovered that the Relief-F's combinations of features achieved the best precision. However, LSA needed lower computation time and was advised for the optimal techniques. For example, Aybar et al. used a grouping genetic algorithm (GGA) to choose the optimal features set that maximize an extreme learning machine's (ELM) efficiency in predicting GHI in Toledo, Spain [41]. Compared to the standalone ELM, findings show that, in terms of RMSE, the FSP can boost the prediction efficiency by 10%. In Spain, Salcedo et al. also built a wrapper FSP method based on a coral reefs optimization algorithm to attain a smaller number of relevant predictive variables, combined with ELM models (CRO-ELM), solving the feature selection process to predict GHI [42]. When compared to the GGA-ELM models, the CRO-ELM was found to be more accurate. Thus, identifying the optimal combination of input variables for the forecasting model is an independent, significant process.

1.2. Contribution statement

According to the presented literature, this study contributes to the present predictive analytics literature concerning global horizontal irradiance (GHI) forecasting by developing a neoteric hybrid forecasting strategy. In this strategy, the Grasshopper optimization algorithm (GOA) and the Boruta-based feature selection algorithm (BA) are incorporated along with the SVR learning algorithm to constitute the proposed predictive approach. The main contributions and innovations of this study can be summarized as follows:

1. The SVR is applied to explore the underlying patterns in complex data series and predict GHI's future values. Real-world data train the model. In general, the merits of the SVR algorithm are thoroughly discussed in Section 3.1.
2. Indeed, the SVR's predictive performance is strongly dependent on the suitable and precise selection of its hyperparameters, namely, C and γ. In this context, this paper hybridizes the SVR method with a novel optimization technique (GOA) to improve its efficiency, accuracy, and calculation speed. Therefore, the SVR's optimal hyperparameters can be automatically acquired, and this saves the manual parameters setting burden and facilitates the entire prediction process.
3. A new optimization algorithm (GOA) is introduced to optimize the SVR's hyperparameters in the forecasting approach. GOA is used in this study because of its simplicity, gradient-free structure, high local optima avoidance, and interpretation of problems as black boxes. Thus, we investigate the usage of this algorithm to solve real-world problems, as it is suggested by [43]. Further discussion about the merits of GOA compared to other optimization techniques can be found in Section 3.2.
4. A new feature selection technique based on BA is used to determine the optimal candidate inputs. The optimal candidates will be transferred to the forecast engine because of this filtering, which improves the forecasting approach's accuracy and speed, ensuring its effectiveness for real-time implementations.
5. Although this study aims to predict hourly GHI for three sites in Saudi Arabia, the findings can be used to help choose the optimal model of ML algorithms for estimating solar radiation in different geographical locations.

Finally, hybridization of the BA and GOA with the SVR for predicting short-term GHI has contributed to superior predictive performance in our novel hybrid model, which, in turn, contributes to the learning paradigm for solar energy modeling. Performance evaluations of the developed predictive model (SVR-GOA-BA_K) are conducted to validate the predictive accuracy by forecasting the hour-ahead GHI in three Saudi Arabian cities (Dhahran, Riyadh, and Jeddah). The proposed approach's efficacy and superiority are verified by benchmark tests and justified by conducting predictive comparisons with distinct predictive algorithms and optimizers.

1.3. Organization of the paper

After summarizing recent allied literature, defining gaps, and stating objectives for the current work in Section 1, the paper's remainder is organized in the following way: A detailed description of study sites, datasets, and data pre-processing and normalization is outlined in Section 2. Research methodology, including the operations of SVR, GOA, BA, the proposed model SVR-GOA-BA_K, and benchmark ML algorithms for comparison purposes, are

Table 1
The geographical coordinates and summary statistics of the sites.

Station information	Sites in Saudi Arabia		
	Dhahran	Riyadh	Jeddah
Latitude (°N)	26.30	24.71	21.49
Longitude (°E)	50.14	46.68	39.24
Elevation (m)	75	668	75
Solar station	KFUPM	K.A.CARE	KAU
Average GHI (kWh/m^2/day)	5.6	6.29	5.82
Maximum GHI (kWh/m^2/day)	8.45	8.74	8.18
Minimum GHI (kWh/m^2/day)	0.55	0.7	1.31
Average temperature (°C)	27.83	25.96	30.85
Maximum temperature (°C)	40.5	39.5	39
Minimum temperature (°C)	8.1	5	20.1

enunciated in Section 3. The results of this research study are reported and together discussed in Section 4. Finally, remarks and future work recommendations are concluded in Section 5.

2. Description of the database

In this section, information about the sites considered in this research across Saudi Arabia is first given. Afterward, a complete description of the datasets utilized in developing and validating the built models is presented. Finally, dataset normalization, pre-processing, and preparation phases are also discussed below.

2.1. Study sites

Fossil fuel resources are responsible for supplying electric energy in most countries, so these resources diminish steadily annually [44]. Hence, substituting such resources with alternative energy resources, especially solar energy, has been considered. Regardless of the enormous hydrocarbon reserves in Saudi Arabia, this nation is potentially one of the best solar energy regions. For photovoltaic generation plants, the prediction of solar radiation is vital in boosting solar energy for electrical production schemes. This paper aims at forecasting the future hourly GHI (at the one-hour-ahead time horizon) for three different sites in Saudi Arabia: Dhahran (located along the eastern coast), Riyadh (situated in the middle), and Jeddah (situated along the western coast) cities. Fig. 2 displays the sites considered in this study. Table 1 summarizes exact location information and some statistical descriptions of the assessed areas.

The datasets utilized in this research were recorded at the King Fahd University of Petroleum & Minerals Solar Monitoring Station (KFUPM) for Dhahran City, K.A.CARE Solar Monitoring Station (K.A.CARE) for Riyadh city, and King Abdulaziz University Solar Monitoring Station (KAU) for Jeddah city. The data obtained by a subset of the Renewable Resource Monitoring and Mapping (RRMM) program developed and operated by The King Abdullah City for Atomic and Renewable Energy (K.A.CARE), as Saudi Arabia's leading renewable energy state agency. The RRMM program of Saudi Arabia is a modern ground-based monitoring network. It can be accessed through the online interactive Renewable Resource Atlas of Saudi Arabia through the website (K.A.CARE; <http://rratlas.energy.gov.sa>) [46]. The data are collected per 1 h by rotating shadow-band radiometric stations in the selected sites with high precision.

2.2. Datasets and feature extraction

This paper aims at forecasting the future GHI (at the one-hour-ahead time horizon) based on hourly data collected over four years (from June 1, 2013, to May 31, 2017) for all the selected sites. Because of the different variables affecting GHI values, we use the so-called feature engineering to classify the most important variables. Such variables are categorized into three groups. The first group is the time-related variables, including month, day, and hour. The second group includes climate-related variables such as temperatures, humidity levels, etc. The one-hour lag observations of GHI, DNI, and DHI have been constructed as the third group to enrich the datasets. These newly extracted features are donated in this study as PrevHourGHI, PrevHourDNI, and PrevHourDHI, respectively. This process of feature extraction is an integral part of this paper's contribution.

Table 2
Summary of input variables.

Input variable explanation	Abbreviation	Unit
The month number in a year	M	Month
The day number in a month	D	Day
The hour number in a day	H	Hour
the ambient temperature of the air	T	°C
Relative humidity	RH	%
Pressure of surface	P	hPa
Wind direction	WD	°N
Wind speed	WS	m/s
Peak wind direction	PWD	°N
Azimuth angle	AA	Â°
Solar zenith angle	SZA	Â°
Direct normal irradiance at the previous hour	PrevHourDNI	Wh/m^2
Diffuse horizontal irradiance at the previous hour	PrevHourDHI	Wh/m^2
Global horizontal irradiance at the previous hour	PrevHourGHI	Wh/m^2

Table 2 enlists all the contributing variables considered in this analysis. Many factors affect the GHI based on the geographical location of interest. The selected input variables are strongly correlated to the quantity of the GHI in the sites of interest in Saudi Arabia, based on the extensive assessment of solar energy resources over the Arabian Peninsula [47]. Besides, since there is little cloud cover, precipitation, fog, dew point, and snow in the desert and arid geographic location of Saudi Arabia, these environmental influences have little effect on the surface GHI [40]. Therefore, they are not considered in this study.

2.3. Data pre-processing and normalization

It is critically recommended that solar datasets be cleaned and filtered before introducing them to machine learning models. It is supposed to clean the night hours and retain them only between sunrise and sunset by filtering them out of the database. Also, since the data near sunset and sunrise are commonly unreliable, a solar elevation-based pre-processing operation is performed: solar radiation data shall be omitted for the solar elevation less than 10 [19]. In this analysis, four years of hourly data were used for each selected site. After the solar component data of GHI has been cleaned and filtered, the overall number of hourly data used in each dataset is around 15435. Mainly, approximately 56.25% of the solar data were not utilized: 1.25% of data were outliers and 55% when the sun elevation is below 10° or within the nighttime.

After each time-series dataset is randomly divided into training (80%) and test (20%) datasets, the training datasets use another pre-process. It is referred to as the k-fold cross-validation technique [48]. It is a method of evaluation utilized to increase the flexibility of a model and, hence, the proposed model's accuracy. Thus, the statistical analysis would generalize well into a single dataset. This analysis uses K-fold cross-validation of the training set to tune the SVR models' parameters. The K-fold cross-validation method splits the original samples of training datasets into K equal sub-samples at random. The models are then tested and validated using a single sub-sample as the validation data, while the remaining K-1 sub-samples are utilized as the training

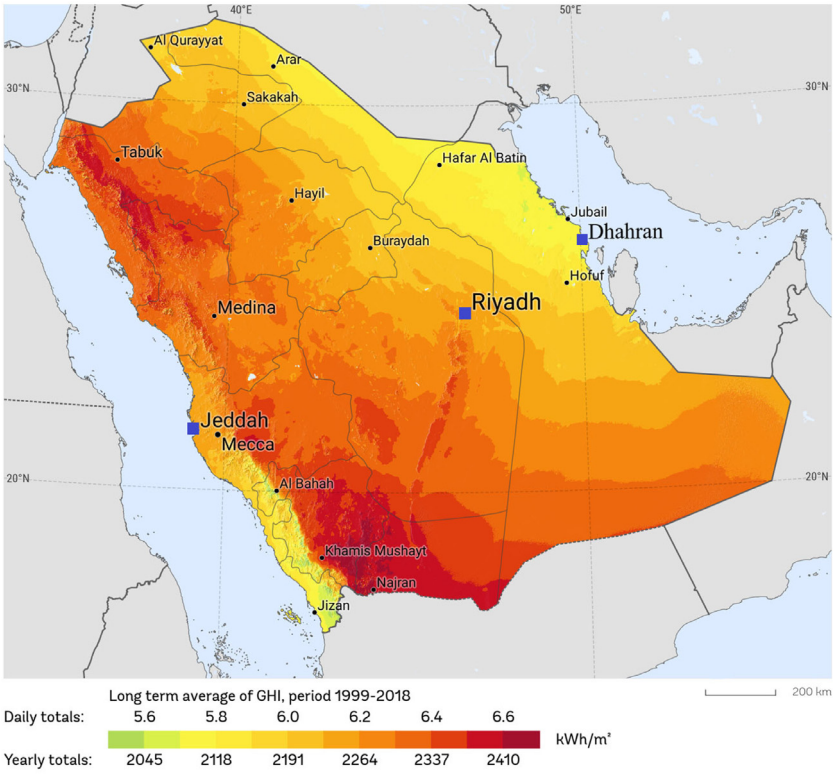


Fig. 2. The considered sites in Saudi Arabia [45].

data. These steps are repeated K times, where all the K subsamples serve precisely once as the validation dataset. After that, the K outcomes from the folds can be averaged to have a sole estimate. The average value of the reliability metrics presented in this article is the k-fold. In this analysis, the value of k taken is equal to 10, as explained in Fig. 3. Consequently, the findings are independent of the training phase's dataset because the conclusions' robustness is reduced by utilizing only one dataset (with its statistical characteristics). To summarize, the k-fold cross-validation technique is used to enhance models' generalization to be used more accurately.

Normalization of input variables data, sometimes recognized by scaling, is vital when adapting ML-based predictive models [49]. This functional implementation aims to prevent the potential superiority of input variables with prominent numerical figures over the variables with miniature figures. Also, because of the reliance of kernel quantities mostly on input vectors' inner multiplication, there are calculational complications induced by large-value input variables. Thus, overcoming numerical complexities during computation processes is another essential aspect for normalizing input vectors. In this analysis, employing Eq. (1), each input variable is scaled linearly to a range [0, 1].

$$x_i^n = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x_i is the input-variable vectors with the measured observation points; the minimum and maximum figures which connect to measured data series become x_{min} and x_{max}; x_i^n is the scaled version of x_i.

3. Research methodology

In this section, the methods that we applied in this paper are amply explained. These are a vital forecasting algorithm SVM, a metaheuristic optimization algorithm GOA, a feature selection algorithm BA, and a proposed hybrid model SVR-GOA-BA$_K$ that integrates these three algorithms. Additionally, benchmark ML algorithms used to assess and validate the proposed model's predictive efficacy are explained.

3.1. Support vector regression (SVR)

SVMs are supervised ML algorithms that can deal with classification and regression problems [50]. Based on input data types, the structure of SVMs is built and optimized. In regression forms of the SVMs, known as ε-SVRs, the initial primal objective is to learn a hypothesis whose all regression prediction errors lie within a predefined threshold, ε. However, the second fundamental objective of the learned function lies in the fact that this function has the best possible generalization capacity. The last goal is intentionally sought so that a flat model can be learned and developed, eventually. Eqs. (2) and (3) impose these two conditional objectives, which together form a standard convex quadratic optimization problem with linear constraints set:

$$\underset{w,b,\xi_i,\xi_i^*}{minimiz} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{2}$$

$$subject\ to \begin{cases} y_i - \langle w, x_i + b\rangle \le \xi_i + \varepsilon, \forall n \\ \langle w, x_i + b\rangle - y_i \le \xi_i^* + \varepsilon, \forall n \\ \xi_i, \xi_i^* \ge 0, \forall n \end{cases} \tag{3}$$

In which, for training points $(x_i, y_i), \ldots, (x_n, y_n)$, n is the number of data samples, the vectors of x_i represent input values, and y_i are corresponding output value for x_i. The upper and lower training regression errors are represented by ξ_i and ξ_i^*, respectively. Such training errors are insensitive to a specific limit characterized by ε, after which penalties will start adding up to the cost function. w is the normal vector. $C > 0$ is the regularization parameter that

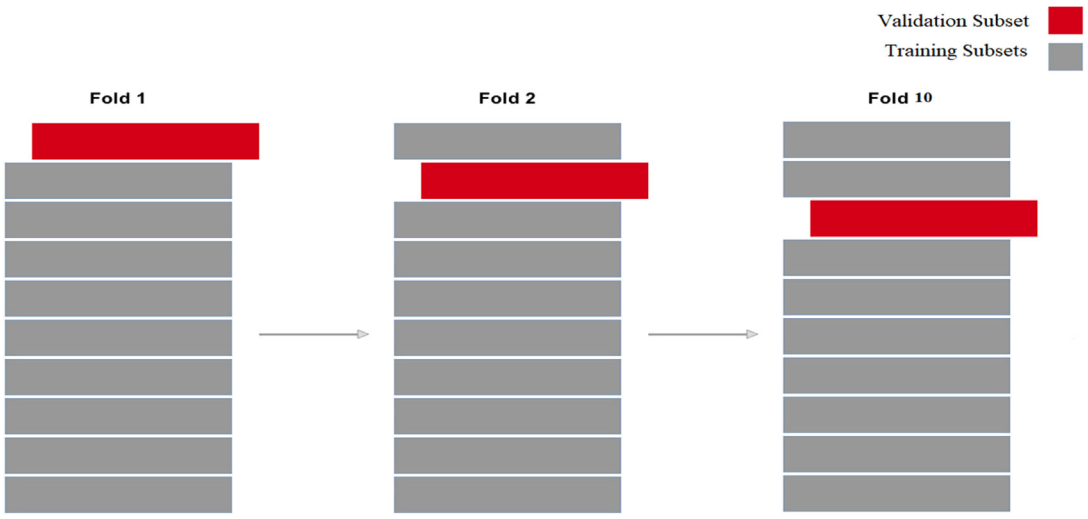


Fig. 3. K-fold cross-validation method, K = 10.

controls the tradeoff between the two different goals imposed in Eqs. (2) and (3).

To find a solution to the optimization problem of the SVR formulated by Eqs. (2) and (3), standard dual optimization through Lagrange multipliers is used. Several transformations are implemented after that the Lagrangian is computed until Eq. (4) is acquired:

$$f\left(x, \alpha_i, \alpha_i^*\right) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) \kappa\left(x, x_i\right) - b \tag{4}$$

Eq. (4) can be obtained by employing the concepts of kernels trick, Lagrange multipliers, and optimality constraints. There are almost four well-known functions in the literature review utilized as kernels: linear, radial basis function (RBF), polynomial, and sigmoid. However, for this study, the RBF is under consideration. The RBF is deliberately chosen because of its computational efficiency, where it generally outperforms the other polynomial and sigmoid functions [51]. It is also highly nonlinear, including having fewer variable parameters and infinite-dimensional mapping space [52]. The used kernel is expressed in Eq. (5) below:

$$K\left(x_i, x_j\right) = e^{-\gamma\left(\left\|x_i - x_j\right\|^2\right)} \tag{5}$$

In which $\gamma \in \mathrm{R}$, $\gamma > 0$ represents the width of the radial basis kernel function.

The architecture of SVR based upon Eq. (4) is shown in Fig. 4, where the conditions of Karush–Kuhn–Tucker's are considered to solve a quadratic optimization problem. $\left(\alpha_i - \alpha_i^*\right)$ values are nonzero support vectors, and they are employed to acquire the decision function. Optimizing the two set-by-user hyperparameters, C and γ, is considerably significant to learn a highly accurate prediction model. Instead, using optimization techniques to determine these parameters' optimal values is gaining attention in recent works.

The SVR algorithm is selected in this study due to its merits that can be summarized as follows [26]:

- It is considered remarkably accurate, reliable, simple-to-implement, and robust to the outliers.
- It can model highly nonlinear complicated patterns and trends seen in weather and solar data series, where you can select from various kernels.
- Compared to other regression models, it is less vulnerable to common overfitting problems, particularly in spaces with high dimensionalities.

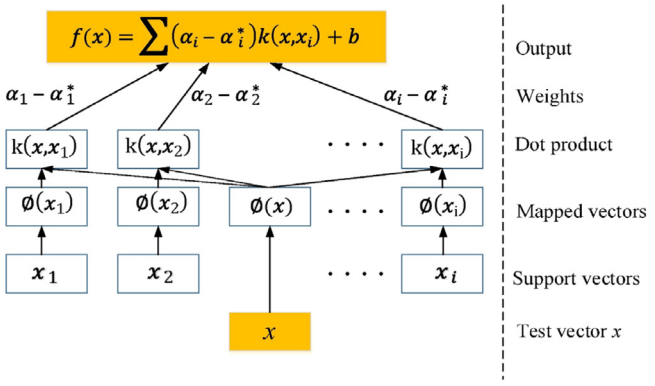


Fig. 4. The architecture of SVR in the scope of optimization solver [50].

- It provides a compact description of the learned model, allowing experiments to be replicated by interested researchers with the same results.
- The learned regression model can be easily updated.
- It performs well when the training data is small, and the number of features is extensive. This feature enhances the generalization ability of the proposed method usages for monthly and annual forecasting of solar radiation where smaller sets are available for training the models.
- The characteristics mentioned above make it one of the most used methods in solar energy forecasting.

3.2. Grasshopper optimization algorithm (GOA)

This paper adopts a modern swarm intelligence technique known as the grasshopper optimization algorithm (GOA). It is based on nature. Saremi et al. proposed this optimization tool and applied it to handle challenging structural optimization problems [43]. The suggested algorithm mathematically models and imitates the grasshopper swarms' behavior in nature to solve optimization problems. In GOA, the search process is logically divided into inclinations: exploration and exploitation. This search process is implemented by search agents: adults and nymph grasshoppers. Naturally, adult grasshoppers abruptly move long distances. Thus, they are utilized to globally search the whole search space to find better regions of food supplies. In other words, they perform the exploration process. In contrast, nymph

grasshoppers are utilized to move and target a particular neighborhood or area locally. This is known as exploitation in optimization terminology.

A smooth balance between exploration and exploitation is ensured by GOA, leading to a less complicated algorithm mathematically. Saremi et al. found a way to model the swarming behavior of grasshoppers mathematically. The mathematical model, which is used to stimulate the grasshoppers' swarming behavior, is symbolized below:

$$X_i = S_i + G_i + A_i \tag{6}$$

where X_i represents the ith grasshopper's position. S_i implies the concept of social interaction. In the meantime, while G_i symbolizes the force of gravity imposed on the ith grasshopper, the wind advection is shown by A_i. Notice that the equation can be rewritten as $X_i = r_1 S_i + r_2 G_i + r_3 A_i$ to provide random behavior, in which r_1, r_2 and r_3 are randomly selected numbers in [0-1]. Mathematically, the GOA algorithm can be implemented by the following steps:

Step 1: the component S_i of Eq. (6) is determined as follows:

$$S_i = \sum_{j=1, j\neq i}^{N} s\left(d_{ij}\right) \cdot \hat{d}_{ij} \tag{7}$$

In which, the ith and jth grasshoppers are separated by a distance denoted as d_{ij}, which is determined by $d_{ij} = \left|x_j - x_i\right|$. The unit vector between the ith and jth grasshoppers is represented by $d_{ij} = (x_i - x_j)/d_{ij}$. Finally, and to consider social forces, a function s is defined as presented in Eq. (8):

$$s(r) = f \cdot e^{-\frac{r}{l}} - e^{-r} \tag{8}$$

In which, f stands for the attraction intensity, l represents the attractive length scale and $r = \left|d_{ij}\right|$. The function s can split the space between two grasshoppers into three zones: repulsion, comfort, and attraction.

Step 2: the component G_i of Eq. (6) is determined as follows:

$$G_i = -g \cdot e_g \tag{9}$$

In which the gravitational constant is denoted by g, whereas e_g is the unit vector heading to the globe center.

Step 3: the symbolic component A_i of Eq. (6) is determined as follows:

$$A_i = u \cdot e_w \tag{10}$$

where u and e_w represent a constant drift and a unity vector in the wind direction, respectively. Notice that since Nymph grasshoppers lack wings, their motions are strongly associated with the wind direction.

By substituting the components of S_i, G_i, and A_i into Eq. (6), this equation is expanded as below:

$$X_i = \sum_{j=1, j\neq i}^{N} s\left(\left|x_j - x_i\right|\right) \cdot \frac{x_j - x_i}{d_{ij}} - g \cdot e_g + u \cdot e_w \tag{11}$$

Here N represents the grasshoppers' number.

Since the grasshoppers quickly reach their comfort zones and the swarms do not converge to certain points, the mathematical model shown in Eq. (7) cannot handle the optimization problems directly. To overcome this obstacle, an amended alternative of this equation is formulated as below:

$$X_i^d = c\left(\sum_{j=1, j\neq i}^{N} c \cdot \frac{ub_d - lb_d}{2} \cdot s\left(\left|x_j^d - x_i^d\right|\right) \cdot \frac{x_j - x_i}{d_{ij}}\right) + T_d \tag{12}$$

Here lb_d and ub_d stand for the lower and upper boundaries in the Dth dimension, respectively. T_d constitutes the location of the optimum solution it has found yet.

The decreasing coefficient c is used in Eq. (12) to shrink the comfort zone, repulsion zone, and attraction zone. It is also worth noting that Eq. (12) contains the adaptive parameter c twice for the following purposes:

- As the number of iterations increases, the target's grasshoppers' movements are reduced by the first c from the left. This parameter, in other words, balances the entire swarm's exploration and exploitation of the target.
- The attraction, comfort, and repulsion regions are reduced among the grasshoppers by the second c parameter. This reduction is proportional to the number of iterations.

Grasshoppers force a gradual and smooth balance between discovery and exploitation due to the differing comfort zone parameter c. This feature allows GOA to avoid being stuck in local optima and instead seek an accurate estimation of the global optimum. The dynamic coefficient c in every iteration can be worked out as below:

$$c = c_{\max} - l \cdot \frac{c_{\max} - c_{\min}}{L} \tag{13}$$

where c_{min} and c_{max} stand for the minimum and the maximum values of the coefficient c, respectively. While l refers to the current iteration, L represents the highest iterations. In the analysis setup, 0.00001 and 1 are the values of c_{min} and c_{max} parameters. We used high repulsion rates in this study because repulsion is a critical technique in the GOA algorithm for avoiding local solutions. The results show that high repulsion rates prevent grasshoppers from stagnating in local optima.

The GOA is mainly selected in this study to optimize the SVR's hyperparameters due to the following reasons:

- Grasshoppers efficiently locate the promising areas of an assigned search space.
- Grasshoppers experience sudden, radical changes in the early stages of optimization, which aids them in searching globally.
- In the final stages of optimization, grasshoppers head to travel locally, allowing the exploitation search of the space.
- Grasshoppers force a gradual and smooth balance between discovery and exploitation due to the differing comfort zone parameter c. This feature allows GOA to avoid being stuck in local optima and instead seek an accurate estimation of the global optimum.
- The GOA optimizer improves grasshoppers' fitness values, proving that it can significantly boost a randomly generated grasshoppers' population.
- As the number of iterations increases, the target's fitness improves, meaning that the global optimum's estimation improves proportionally to the number of iterations.
- GOA can solve real-world problems involving unknown search spaces.
- GOA outperforms other current algorithms when tackling a range of existing or new optimization problems.

To sum up, the steps in which the GOA is executed can be found in Fig. 5.

3.3. Boruta feature selection algorithm (BA)

The BA is designed as an ensemble-based feature selection algorithm [53]. It emulates the RF's working theory with additional mechanisms to achieve superior results. Its name is derived from a god of the forest in Slavic mythology. The BA is mainly built to distinguish the so-called "all relevant variables" in classification or regression problems. This method's key idea is to use statistical

1: Initialize the swarm population (grasshoppers) $X_i, where\ i = (1,2,\ldots..,N)$
2: Initialize the parameters: c_{min} , c_{max} , L
3: Calculate the fitness value of each search agent
4: Assign T to the best search agent (the individual with the highest fitness value)
5: **while** $l < L$ **do**
6: Use Eq.13 to update c
7: **for** each search agent
8: Normalize the distances between grasshoppers within [1,4]
9: Update the position of the current search gent by Eq.12
10: Bring the current search engine back when it exceeds the boundaries
11: **end for**
12: If there is a better solution, update T
13: $l = l + 1$
14: **end while**
15: Return T

Fig. 5. The pseudo-code of the GOA optimization algorithm.

testing and multiple runs of RF to compare the original predictor variables' importance with those with an increased level of randomness. This extra added randomness offers a better picture of which variables are important and relevant. The BA execution is composed of the subsequent steps:

(1) Make copies of all input variables, known as shadow attributes, through expanding the information system.
(2) Disturb the shadow attribute values to decrease the relationship with the output variable(s).
(3) Obtain the importance values of all features, including the shadow attributes, by training an RF regressor upon the new expanded dataset. Such important values are known as Z scores.
(4) Choose the maximum Z score among shadow attributes (MZSA). After that, assign any feature that scored better than MZSA to a hit.
(5) Features that have still need to be evaluated for importance are then ordered to perform the two-sided equality test with MZSF.
(6) Features with a Z score vastly larger than MZSF are labeled "important", whereas features with a Z score smaller than MZSF are labeled "unimportant".
(7) Delete each unimportant variable and the whole shadow attributes.
(8) The preceding steps are reiterated until all variables have been classified or the algorithm has exceeded a pre-specified number of the random forest runs.

Interestingly, the BA defines all relevant variables in the information system and returns the importance ranking of features from the most important to the least important. It also assigns important variables with numerical scores ranking their importance. Thus, this can help researchers construct different input combinations based on its feature importance ranking to find the optimal feature set. One should notice that the BA authentic execution, based on the standard *randomForest* R package, was computationally heavy [51]. Even though the Boruta method's applications in high-dimensional datasets were challenging, it was utilized in more than 100 studies. The ranger package is used in the 5.0 version of the Boruta package to train the random forest and estimate variables' importance. In conclusion, more information about using the Boruta package amply can be found in [53].

3.4. The establishment of the proposed model (SVR-GOA-BA$_K$)

In this section, a hybrid model, denoted SVR-GOA-BA$_K$, is proposed to boost the performance accuracy in one-hour GHI predictions. Three practical algorithms are integrated into SVR-GOA-BA$_K$: BA for feature selection, GOA for parameter optimization, and SVM regressor. The proposed model BA-GOA-SVR$_K$ is parameterized with K, which indicates the number of input features participating in building the proposed model according to the K highest scores of importance ranked by BA. In the SVR-GOA-BA$_K$, BA is first used to choose important features and delete unimportant ones from the input variables considered in this study. GOA is then employed in the training phase to optimize and set two optimal hyperparameters of the SVM (C and γ). Ultimately, we harness the proposed SVR-GOA-BA$_K$ model to execute the forecasting phase. The basic flowchart of SVR-GOA-BA$_K$ is depicted in Fig. 6, which comprises the following four steps:

Step 1: Data selection and preparation:

- From the original time series dataset, determine time and weather variables. Also, extract the previous-hour observations/lags of the three solar radiation components (PrevHourDNI, PrevHourDHI, PrevHourGHI) and then add them as new features (see Table 2, Section 2.2). There are 14 input variables to be considered in each dataset.
- The data pre-processing and normalization steps on the input GHI series dataset are executed, as explained in Section 2.3.

Step 2: Feature selection by BA:

- Apply the BA to provide unbiased and consistent selection and ranking of important and non-important input variables from the GHI series datasets.
- According to important features ranking, construct different input combinations, denoted as (K=1, K=2, …, K=All). The K=1 set includes only the first most important feature; the K=2 set includes only the two most important features, and up to the end set K=All, where it has all the important features defined by BA. The objective is to find the minimal optimal set of inputs.

Step 3: Optimization by GOA:

- Employ GOA to optimize the two hyperparameters of SVR: C and γ.
- The final regression models will be developed based on the different input combinations (SVR-GOA-BA$_1$, SVR-GOA-BA$_2$, …, SVR-GOA-BA$_{All}$).

Step 4: Forecasting by SVR-GOA-BA$_K$:

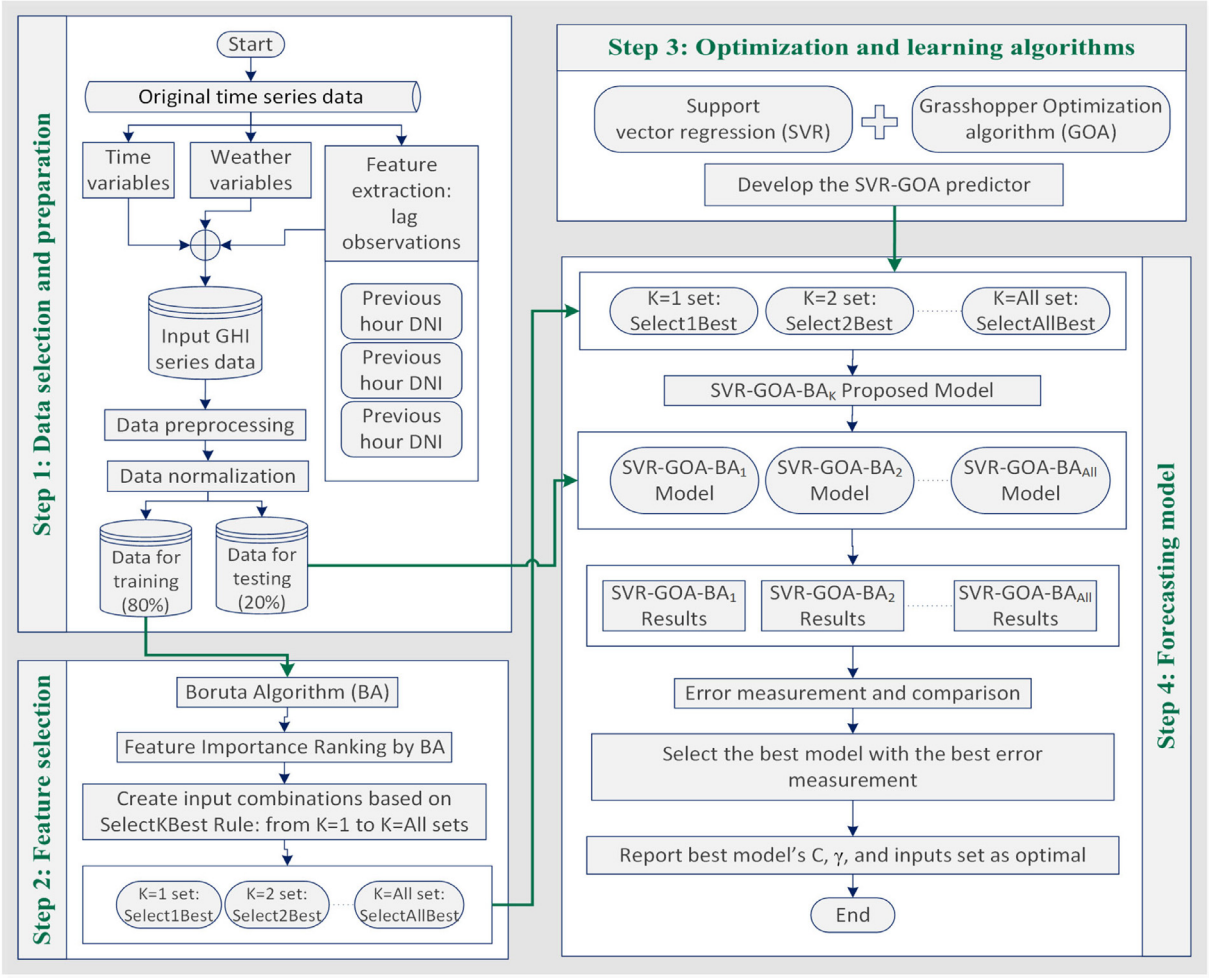


Fig. 6. Flowchart of the proposed research framework.

- Use the testing samples to evaluate the corresponding developed predictive models and return the prediction results.
- Compare all developed models' results based on evaluation metrics (see Section 4.1).
- Select the best model with the best error measures. Thus, the best model's input combination is considered the optimal set of features, and its hyperparameters C and γ are the optimal values for SVR.
- Repeat all previous steps for each site's dataset independently.

3.5. Benchmark machine learning algorithms

In this section, each ML algorithm used for comparative performance assessment purposes is briefly described: ANN, DT, KNN, and RF. It can be noted that each considered algorithm belongs to a particular family in supervised categories of ML regression algorithms. The ANN belongs to the non-parametric family, while KNN and DT are categorized under clustering and decision-tree-based families, respectively. From the ensemble family, the powerful RF is selected. This is intended to grand this research study of diversity, solidarity, and unbiasedness.

3.5.1. Artificial neural network (ANN)

The ANN is the most widespread machine learning algorithm for prediction intents; thus, the article gives a few details solely. A full description of the ANN can be found in [54]. ANN is a nonlinear regressor that applies a basic structure of interconnected parts. A three-layer MLP with feedforward backpropagation is the sort of ANN employed here [55]. Input data, the first layer, is processed by the hidden layer, the second layer, and an output indication is sent to the output layer, the third layer. In a feedforward MLP configuration, input variables and every neuron extradite signals to the subsequent neurons in a unidirectional fashion. A nonlinear sigmoid function was taken for the hidden layer and a linear one for the output layer [56]. The Levenberg–Marquardt (LM) learning method was chosen as the optimization tool of the MLP: multiple architectures are evaluated using different numbers of neurons in the hidden layer, and the most effective is preferred [57]. Generally practicing, the number of neurons in hidden layers in the range of 3 to n+2, in which n refers to the number of input variables in the input layer. ANN is widely used for solar energy forecasting due to its powerful nonlinear estimation ability [58].

3.5.2. Decision trees regression (DT)

The DT has become widely accepted and often used for forecasting applications in regression problems. It uses the simple notion that a tree should evolve from roots to leaves. Therefore, a DT begins with a root node that leads to other subsequent non-leaf nodes. A test is implemented at each node by evaluating a particular condition on an input variable, either binary or categorical. The branches continue to break until leaf nodes are achieved to find a potential value of the predicted output. There is, therefore, a route to pursue via decision-making from the root node to the leaf nodes.

DTs are efficient approaches applied for predictive studies of both solar and wind energy. Prediction results are comparable to specific other single data-driven methods such as artificial neural networks and vector machine support. However, DTs have the substantial benefit of being easy to comprehend when their implementation and running are reasonably sophisticated. Moreover, they have the advantage of being able to discover and spotlight complicated or hidden relationships within the data.

Multiple algorithms to construct decision trees have been established. The classification and regression tree model (CART), crafted by [59], is one of the most widely used algorithms for solar radiation forecasting applications. In forecasting applications, CART stands for classification trees whenever the predicted output is a class, while it refers to regression trees whenever the expected outcome is a number. In this research, the CART algorithm was implemented since it solves regression problems and is not limited to classification problems.

3.5.3. K-nearest neighbors

The KNN is a non-parametric pattern recognition algorithm that provides a forecast using the mean of the nearest K observations in the test dataset. KNN applies to problems of classification and regression. Distance metrics are used to assess the nearness of observations in the input space. A commonly used metric is the Euclidean distance, for example. In addition to the distance metric, K, the number of neighbors considered for predicting the output variable can be freely selected. Low K values can result in over-fitting problems, while high K values also lead to even worse performance. One mechanism for determining an optimal K value is to consider various subsets of the training data. KNN has been optimized in this research by adjusting neighbors' numbers (K) and monitoring RMSE metric values on training data series. The RMSE value was at its lowest if the neighbors' number set to K=5. Additional KNN details can be found in [60].

3.5.4. Random forest (RF)

The RF is a non-parametric, supervised ML algorithm. It exploits alternative analyses, randomness strategies, and ensemble methods to create subtle ML models without overfitting [61]. The forest is a set of regression decision trees that are trained through bagging techniques. The notable merits of RF involve the discovery of data anomalies, the detection of significant features, the discovery of data trends, and the provision of informative graphics [62]. The RF is one of the widely used algorithms for solar energy prediction because of its simplicity, efficiency, and variety. Thus, it is considered in this study. Detailed information about RF can be found in [63].

4. Results and discussion

In this section, an overview of the statistical indicators used in evaluating all developed models' effectiveness is given. Afterward, the feature selection analysis results, the proposed model performance, and the comparative performance assessment are sequentially presented.

4.1. Performance evaluation metrics

Several statistical metrics of error can be utilized to assess the established models' predictive effectiveness. This research primarily takes six metrics into account, namely: the root mean square error (RMSE), the normalized root mean square error (nRMSE), the goodness of fit (R^2), the mean absolute percentage error (MAPE), mean absolute error (MAE), and normalized mean absolute error (nMAE). Eqs. (14)–(19) mathematically represent these metrics as below:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(f_t - y_t)^2} \tag{14}$$

$$nRMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{f_t - y_t}{y_{max} - y_{min}}\right)^2} \times 100\% \tag{15}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{N}(f_i - \tilde{y})^2}{\sum_{t=1}^{N}(y_i - \tilde{y})^2} \tag{16}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{f_t - y_t}{y_t}\right| \times 100\% \tag{17}$$

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}}|\mathrm{f_t} - \mathrm{y_t}| \tag{18}$$

$$\mathrm{nMAE} = \left(\frac{\mathrm{MAE}}{\tilde{\mathrm{y}}}\right) \times 100\% \tag{19}$$

In which N reflects the number of observed points participated in the process of evaluation; f_t and y_t refer to the forecasted and observed values of the target at the time step t; y_{max} and y_{min} indicate the maximum and minimum of the observed values of the target y_t; the average of the observed values of the target variable y is $\tilde{y}$. This average is calculated on the N data. Although the RMSE determines a prediction model standard deviation, nRMSE is used to compare models with different scales expressed as percentages [49]. Also, a model's R^2 determines how well the model suits a series of observations in regression problems. To assess the performance efficacy of a predictive model as a percentage, MAPE is often used. MAE quantifies forecast errors with a focus on the mean error rather than individual extreme events. A lower value for RMSE implies a better model efficacy. Smaller MAE values, like the RMSE, denote more excellent agreement between measured and forecasted values. The model is the best model when the costs of nRMSE, nMAE, and MAPE are close to zero, and R^2 is close to one.

4.2. Feature selection analysis

To assess the importance of significant features in predicting GHI's future estimates, the BA is implemented. Initially, the proposed strategy considers 14 features for the ultimate choice of the independent input variables. The BA was implemented through its publically available R package by using 100 iterations. Indeed, no substantial changes in the study outcomes were noted beyond 100 runs. For the three sites of interest, the BA-based feature selection technique's findings are depicted by Figs. 7–9.

In Figs. 7–9, box plots show the importance of the independent variables evaluated by BA. Variables marked with the green box plots are important because of their greater predictability than the shadow features, marked in blue colors. For all three study sites, all the independent variables were identified as important. Thus, all 14 variables will be utilized in building different input combinations to forecast the future estimates of the GHI values in the three cities. There will be 14 different input sets for each site, meaning 14 predictive models to be developed for each location based on the proposed framework, as can be seen in the next section. This is intended to discover the minimal optimal set of inputs to overcome underfitting or overfitting problems. Moreover, the variables marked in red in BA outcomes have less informative capacity than the shadow attributes. Therefore, they are excluded from the ultimate set. Also, the variables with yellow are considered tentative. As a result, none of the considered

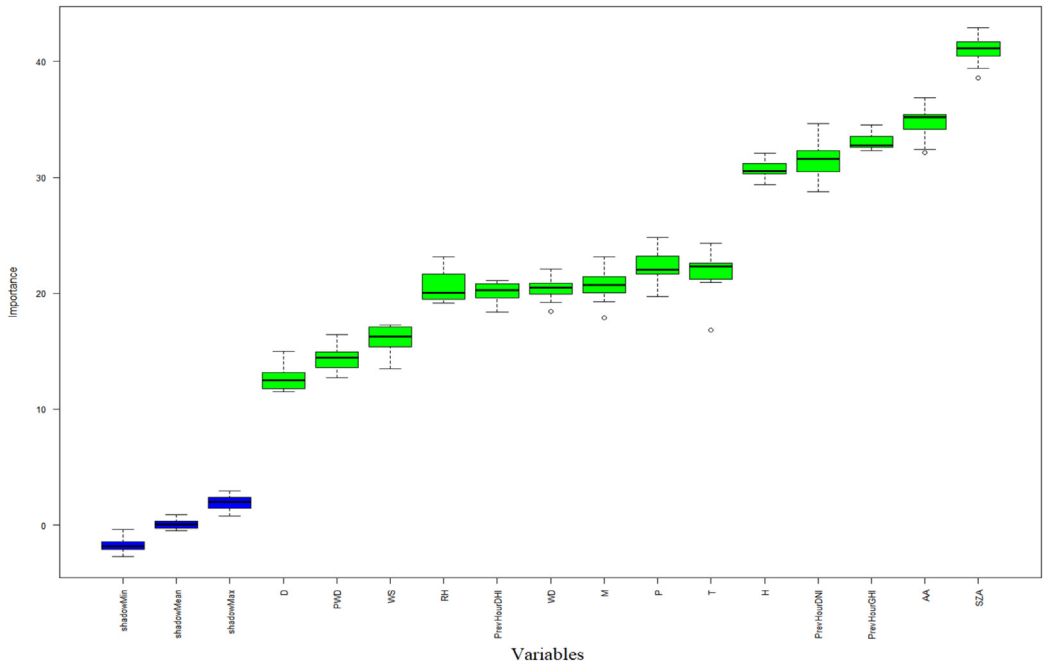


Fig. 7. The findings of the feature selection analysis on the Dhahran dataset.

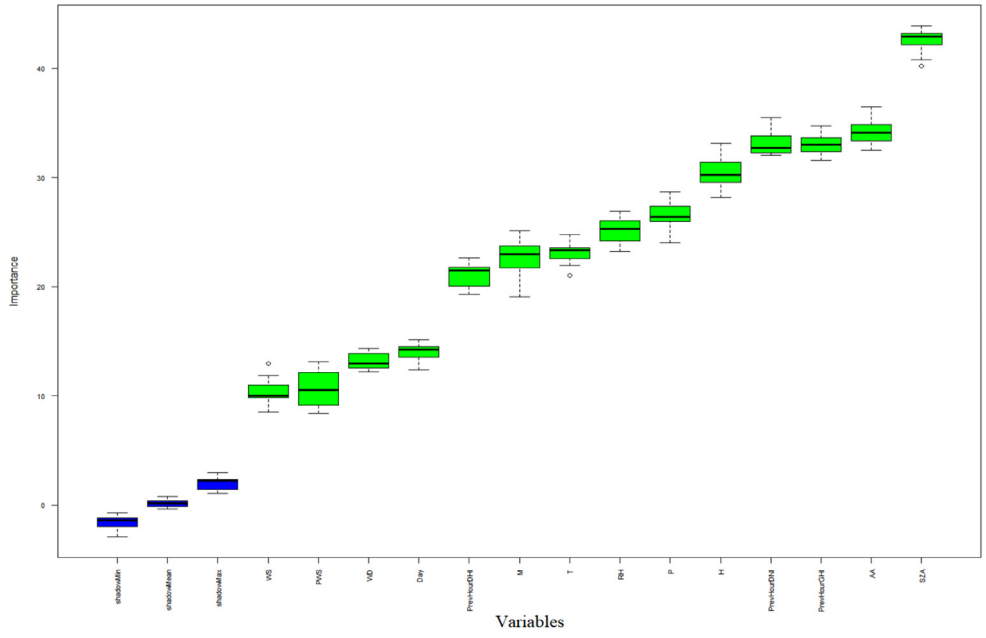


Fig. 8. The findings of the feature selection analysis on the Riyadh dataset.

independent variables emerged as tentative nor insignificant in all the three datasets associated with this study's three sites.

The SZA, AA, and PrevHourGHI variables have proven to be the most significant variables, ranked in the same descending order of importance, for all different datasets collected from the three sites. Following these variables, the PrevHourDNI variables ranked fourth in importance in both the Dhahran and Riyadh datasets, while it came ninth in the Jeddah series. Conversely, the PrevHourDHI variables became less significant, where it ranked tenth in both Dhahran and Riyadh and ninth in Jeddah. Further, the wind-related variables were often ranked the lowest in the importance of the three locations of interest. Indeed, embracing the idea of extracting new features from lag observations of the solar irradiance's three components to enrich the information system has proven helpful in increasing GHI predictability. Consequently, this research recommends that future researchers consider lags observations of GHI and DNI variables as inputs into their potential developed models.

Finally, the seen consistency in BA's findings asserts that the proposed algorithm has the strength to provide an unbiased and stable selection of important and non-important features from an artificial dataset. As integrating more variables can cause overfitting problems, the novel BA's ability to rank features in descending order of importance can help researchers determine which features are relevant to GHI prediction. Therefore, discarding irrelevant or even less relevant variables can ease the computational complexity and time associated with optimizing

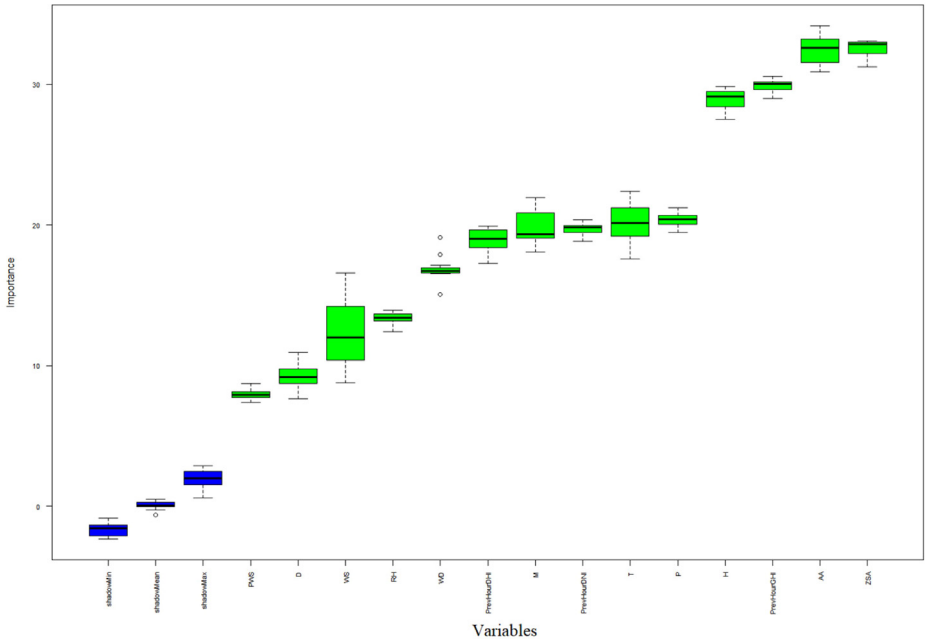


Fig. 9. The findings of the feature selection analysis on the Jeddah dataset.

Table 3
Predictive performance of the SVR-GOA-BA_Kmodel.

	Dhahran	Riyadh	Jeddah
Best model	SVR-GOA-BA_{11}	SVR-GOA-BA_9	SVR-GOA-BA_{12}
Number of inputs	11	9	12
RMSE (W/m^2)	45.0903	49.8129	41.1592
nRMSE (%)	4.35	4.67	3.96
R^2	0.98823481	0.98863249	0.98883136
MAPE (%)	9.13	7.24	6.45
MAE (W/m^2)	24.7827	23.7835	18.7913
nMAE (%)	2.37	2.18	1.78
C	7214.03906	1079.400744	1618.60915
γ	0.525261	1.610183	1.61277

the proposed scheme's hyperparameters, as can be seen in the next section.

4.3. The performance of the proposed model SVR-GOA-BA_K

For conducting predictive analytics, and after identifying the ranking of importance of the independent features for each site's dataset by BA, an SVR model kernelized with RBF is applied to the time-series datasets. SVR's two hyperparameters, C and γ, are optimized through GOA in the training phases. For each site, fourteen SVR-GOA-BA_K models (SVR-GOA-BA_1 to SVR-GOA-BA_{14}) are built based on fourteen different input variable sets (K=1 set to K=14 set). In each site's dataset, the K=1 set includes only the first most important feature, while the K=14 set includes all the 14 important inputs ranked by BA (see Figs. 7–9). This analysis employs the mean absolute percentage error measure (MAPE), also utilized as the GOA's objective function, as the selection criterion between SVR-GOA-BA_K model architectures. The target for all the models is to forecast the GHI values at the hourly level for the chosen sites.

During the execution of the experiment, each time-series dataset is randomly divided into training (80%) and test (20%) datasets. The training set is used to build the predictive models, whereas the test data sample is used to assess the forecasting accuracy. Notably, the same training and test sets are constantly introduced to all the developed models. Such models are all being developed in the MATLAB environment. For GOA, the maximum number of iterations of all built models is fixed to 50, and the search agents' number is at 30. The analysis employs lower and upper bounds of the two hyperparameters of SVR C and γ set to [0–10000] and [0–2]. Also, the MAPE function is selected as the GOA's cost function to be minimized. All the input variables are normalized between 0 and 1 for the performance metrics calculations and for mitigating the computational complexity during models' hyperparameters searching. The predictive efficacy of the built SVR-GOA-BA_K models is assessed with new datasets (such data are not utilized during the model's development phase). Table 3 summarizes the proposed models' best predictive performance, corresponding values of the SVR's hyperparameters optimized by GOA, and minimal optimal set of features for all the sites.

From Table 3, it can be observed that the values of MAPE are lower than 9.13% for all three sites. nRMSE's estimates are also lower than 4.67%. The average difference between the forecast and the observed GHI, given by nMAE, is less than 2.37%. It is estimated that the values of RMSE, MAE, and R^2 are at a low level. By comparing the findings of RMSE and RMSE values for all locations, it is apparent that very substantial errors in forecasting are unlikely to have happened. No site needs the whole set of important input variables (K=14 set) to achieve the best predictive performance. Hence, the inference can be drawn that the efficacy of SVR-GOA-BA_K in predicting the next-hour GHI predictions for Dhahran, Riyadh, and Jeddah sites are exceptional. Further, a closer look at the performance metrics helps gain more insights. The Jeddah site's MAPE, nRMSE, and nMAE values are lesser than the other sites, and the number of its model's inputs is the highest at 12. Contrarily, the optimal number of inputs into the Riyadh site's predictor is the lowest at nine variables. In terms of MAPE, nRMSE, nMAE, and R^2, it is possible to label GHI patterns in the Jeddah site as more predictable than Dhahran and Riyadh's sites. Therefore, it can be assumed that GHI dynamics in Jeddah city in Saudi Arabia are more systematic than the other two cities. Accordingly, the findings will enable policymakers to shape future policies. Fig. 10 depicts the performance metrics values to amply assess all the built SVR-GOA-BA_K models' efficacy based on different combinations of important inputs (K=1 set to K=14 set) for the three sites through heatmaps plots.

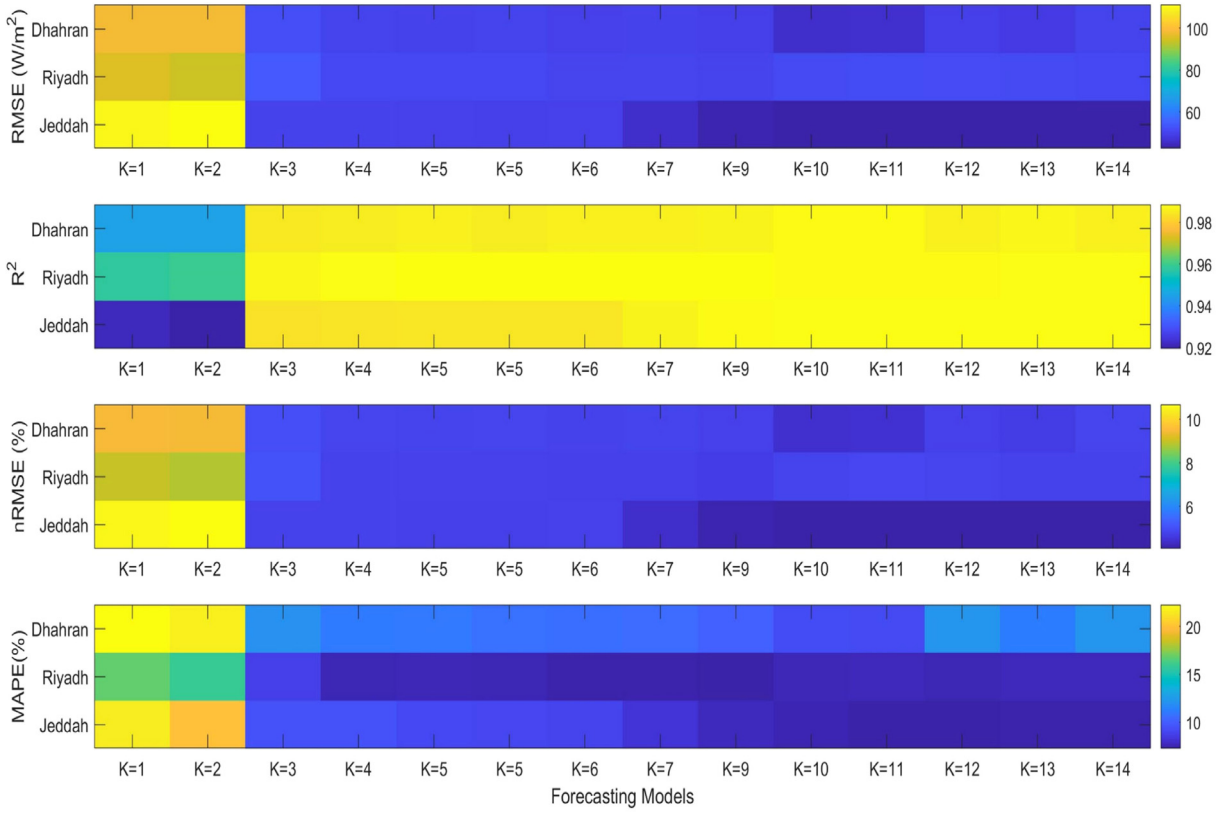


Fig. 10. Performance evaluation of the SVR-GOA-BA_K model with different input variables for all sites.

From Fig. 10, it is evident that the MAPE and nRMSE estimates acquired from all the developed SVR-GOA-BA_K models in all sites are almost always lower than 22% and 10%, respectively, even if only a single input is forwarded to the regressor. For example, the obtained values of MAPE and nRMSE indicators are 16.77% and 8.97%, respectively, in Riyadh when only one input is used to build the regressor. Furthermore, it can be seen from Fig. 10 that BA-GOA-SVR_1 and BA-GOA-SVR_2 models are associated with the worst forecasting performance at all study sites. For instance, in Dhahran, the MAPE, nRMSE, and R^2 values obtained for SVR-GOA-BA_1 are 22.26%, 9.46%, and 0.9450, respectively. On the other hand, the built models become more accurate by utilizing the three most important features and more for all the sites. For example, in Jeddah, the obtained MAPE ranges between 9.443% and 7.3323% for the models from SVR-GOA-BA_3 to SVR-GOA-BA_{14}. Thus, it can be inferred that utilizing only the three most important features of each site's dataset selected and ranked by BA would provide reliable forecasting results compared to the best-proposed model for each location. Also, similar results are found with RMSE, nRMSE, and R^2. This phenomenon is visually depicted in Fig. 11 through the scatter plots of the observed vs. forecasted GHI values generated by the proposed models using the Jeddah site data. In Fig. 11, the subgraph with red dots refers to the model associated with Jeddah's best predictive efficiency, SVR-GOA-BA_{12}.

4.4. Comparative performance assessment

For comparison of the efficacy, the proposed SVR-GOA-BA_K model is firstly compared to the standalone SVR model. All the fourteen informative features considered in this research's analysis have been selected as independent input variables for training the standalone SVR. Table 4 presents the values of the six metrics of performance for all the considered sites.

Table 4
Comparison of SVR-GOA-BA_K and SVR.

Error metric	Site	Standalone SVR	SVR-GOA-BA_K[a]
RMSE (W/m^2)	Dhahran	56.311	45.0903
	Riyadh	64.8953	49.8129
	Jeddah	47.04534	41.1592
nRMSE (%)	Dhahran	5.44	4.35
	Riyadh	6.09	4.67
	Jeddah	4.52	3.96
R^2	Dhahran	0.981685	0.98823481
	Riyadh	0.980694	0.98863249
	Jeddah	0.983072	0.98883136
MAPE (%)	Dhahran	15.14	9.13
	Riyadh	10.67	7.24
	Jeddah	9.98	6.45
MAE (W/m^2)	Dhahran	31.0125	24.7827
	Riyadh	35.9057	23.7835
	Jeddah	28.4202	18.7913
nMAE (%)	Dhahran	2.97	2.37
	Riyadh	3.3	2.18
	Jeddah	2.65	1.78

[a] $K = 11$ for Dhahran, $K = 9$ for Riyadh, and $K = 12$ for Jeddah.

From Table 4, It is amply clear that the SVR-GOA-BA_K models' obtained MAPE figures are remarkably lower than the standalone SVR model in all sites. For instance, approximately a 40%-better solution has been acquired in Dhahran after applying the proposed prediction scheme. Likewise, the prediction accuracy has improved by around 32% and 35% in Riyadh and Jeddah. In terms of RMSE and MAE, the high variations of forecasting errors seen in standalone SVR models are outstandingly reduced by the proposed models. Therefore, in terms of predictive accuracy, the inference can be derived that SVR-GOA-BA_K exceptionally outperforms the standard SVR model in predicting hourly GHI in the

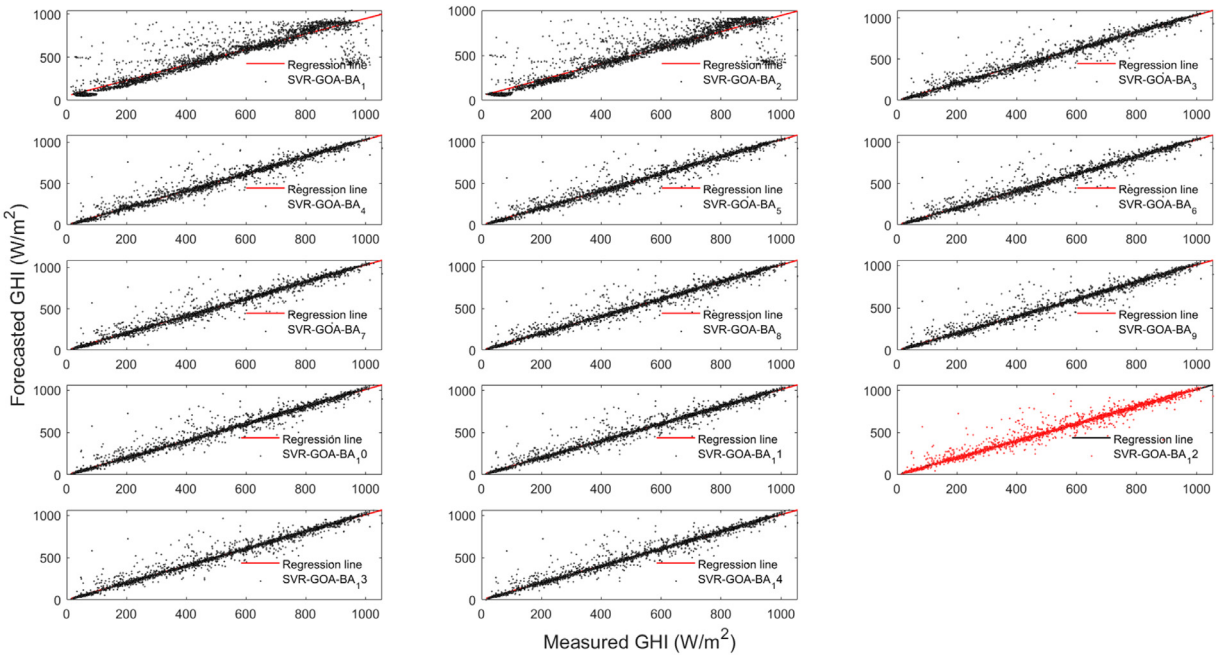


Fig. 11. Scatter plots of the observed vs. measured GHI prediction by different proposed models for Jeddah, the best in blue SVR-GOA-BA_{12}.

chosen sites in Saudi Arabia. The integration of GOA for optimizing the SVR's parameters and BA for the feature selection process in the proposed SVR-GOA-BA_K model is mainly responsible for improving the forecasts. Nonetheless, the rationalization of the SVR-GOA-BA_K model's effectiveness upon the told comparative assessment might not mirror the performance's actual nature.

The SVR-GOA-BA_K model's efficacy should be compared to several excellently known models to assess its superiority. In this article, four different models are used to achieve this objective: DT, KNN, ANN, and RF. However, neither GOA nor BA has been incorporated or applied to these models. In selecting the independent input variables for training DT, KNN, ANN, and RF, all fourteen informative features considered in the analysis of this research have been chosen. In DT, the maximum tree depth was fixed at 21, and the minimum number of observations was fixed at 5 for the leaf nodes. A backpropagation approach by Levenberg–Marquardt (LM) learning algorithm was used in optimizing the ANN model, with one hidden layer consisting of 14 hidden nodes. Also, Sigmoid activation functions are utilized when the learning rate magnitude is fixed at 0.1. For KNN, the number of neighbors k was set at 5. In implementing RF, the number of trees to grow (ntree parameter) was set to 500, while the maximum tree depth (mtry) was fixed as the default value. MATLAB and Python programming environments have established all these models. All the programs have been executed in a 64-bit Windows operating system using the Intel Core i5-7200U (2.50 GHz) processor with 8.00 GB of RAM. Table 5 summarizes the findings of the comparative analysis exhibited by all considered evaluation metrics.

It becomes apparent fromTable 5 that the MAPE, nRMSE, and nMAE of the proposed SVR-GOA-BA_K model in all selected sites is lower than all the other benchmark models. Notably, among the competing models and in terms of MAPE, RF's performance is better than the rest of the models. There is no primary difference in both the RF and ANA results. On the contrary, DT's predictability, followed by KNN, has appeared to be the least satisfactory among all models as its MAPE, nMAE, nRMSE, MAE, R^2, and RMSE are higher than any other model for all locations. The evaluation indicators' values demonstrate that the proposed SVR-GOA-BA_K framework is statistically superior to the four models used for comparison at every site. Fig. 12 shows, for example, the predictive efficacy of the proposed BA-GOA-SVR_K for Jeddah, in which the forecasted data effectively pursues the actual data with minor deviations. Also, Fig. 12 shows a slight variation in the magnitude of residual errors, and substantial errors are unlikely to have occurred. The statistical error metrics in Table 5 indicate that RMSE, nRMSE, MAE, nMAE, R^2, and MAPE are relatively low. Fig. 13 depicts the GHI prediction results of the developed and contrast ML models for Riyadh.

Additionally, the boxplots are created to provide a more thorough evaluation of the predictive performance of the developed forecasting models. In Fig. 14, a box and whisker plot (BWP) depicts the distribution of the predictive errors (nAME) when all locations are grouped together. A numerically distant data point from the rest of the data is referred to as an outlier. When examining a BWP, an outlier (indicated by black circles) is defined as a data point that is situated outside the whiskers of the box plot with a value that is greater than 1.5 times the interquartile range from the box's top or bottom. In line with the prior inferences, the proposed SVR-GOA-BA_K models outperform the RF, ANN, Standalone SVR, KNN, and DT models in all locations, owing to the lower spread of outliers in BWP plots. The results of the comparative assessment rationalize the proposed SVR-GOA-BA_K model's superiority over the others. Hence, the importance of integrating a devoted feature engineering technique through BA with the learning ability of SVR regressor that is optimized by GOA to model the patterns for predicting the hourly GHI is appropriately validated.

4.5. Comparison with benchmark optimization algorithms

Multiple experiments are carried out in this section to validate that the synthesis of SVR, GOA, and BA provides the best results. This simulation employs three optimizers to acquire the SVR hyperparameters. The first significant optimization technique is PSO, which demonstrates the traditional optimization performance for SVR's hyperparameters tuning. Eberhart and Kennedy proposed PSO in 1995, and further information on this popular algorithm can be found in [64]. Compared to a recent evolutionary optimizer, the Cuckoo Optimization Algorithm (COA) is widely

Table 5
Comparative performance assessment.

Error metric	Model	Dhahran	Riyadh	Jeddah
RMSE (W/m^2)	DT	79.8467	90.7218	78.6825
	KNN	59.2137	64.8506	51.6999
	ANN	46.5183	52.424	43.4204
	RF	46.0629	49.9736	42.5572
	SVR-GOA-BA_K[a]	45.0903	49.8129	41.1592
nRMSE (%)	DT	7.71	8.51	7.56
	KNN	5.72	6.08	4.97
	ANN	4.49	4.92	4.17
	RF	4.45	4.69	4.09
	SVR-GOA-BA_K[a]	4.35	4.67	3.96
R^2	DT	0.96334225	0.962361	0.95922436
	KNN	0.97970404	0.98069409	0.98227921
	ANN	0.98743969	0.98743969	0.98743969
	RF	0.98763844	0.98843364	0.988036
	SVR-GOA-BA_K[a]	0.98823481	0.98863249	0.98883136
MAPE (%)	DT	21.17	20.45	18.68
	KNN	16.54	14.08	11.45
	ANN	9.95	8.10	9.09
	RF	9.22	7.60	7.25
	SVR-GOA-BA_K[a]	9.13	7.24	6.45
MAE (W/m^2)	DT	57.5645	47.9761	40.1606
	KNN	40.7822	44.1707	31.3185
	ANN	27.0484	29.3686	22.7145
	RF	25.2624	25.9656	20.3843
	SVR-GOA-BA_K[a]	24.7827	23.7835	18.7913
nMAE (%)	DT	5.51	4.72	3.81
	KNN	3.91	4.06	2.97
	ANN	2.59	2.69	2.15
	RF	2.43	2.39	1.93
	SVR-GOA-BA_K[a]	2.37	2.18	1.78

[a]K = 11 for Dhahran, K = 9 for Riyadh, and K = 12 for Jeddah.

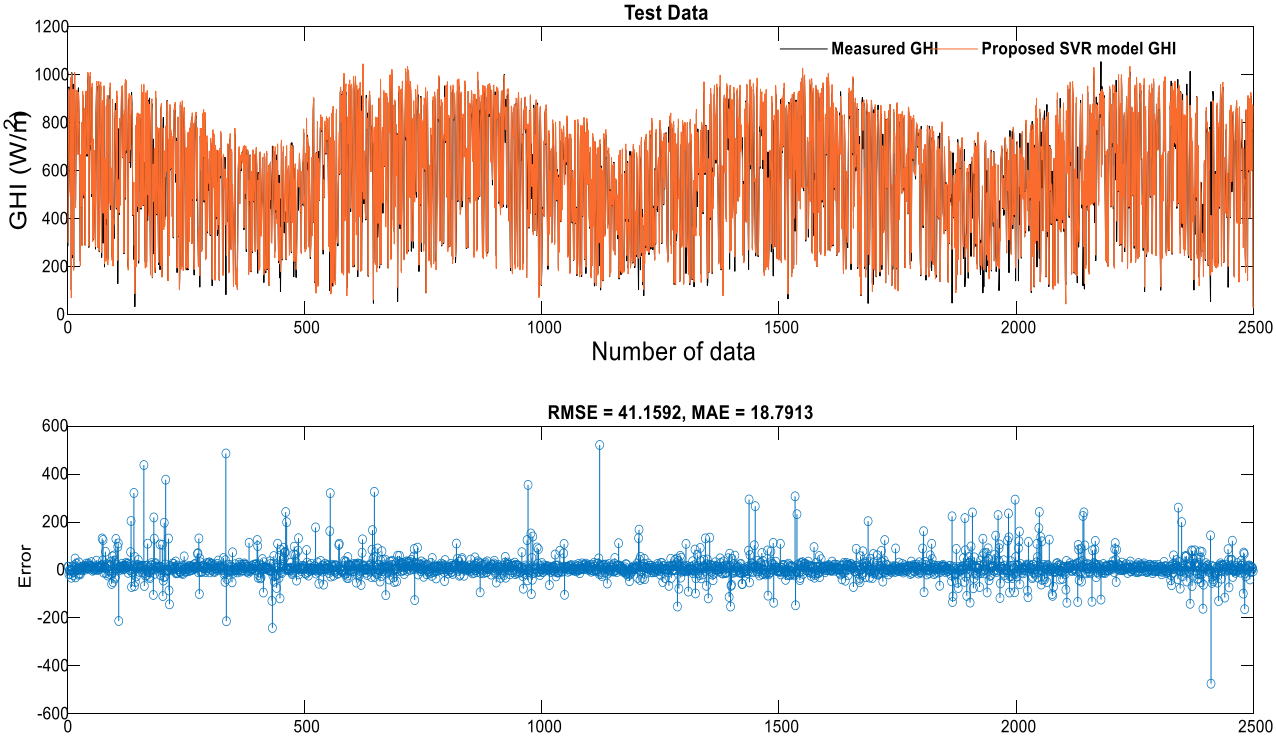


Fig. 12. Comparison of the measured and forecasted values of GHI for Jeddah.

used in parameters tuning for machine learning algorithms [65]. The third optimizer, one of the latest optimization algorithms, the dynamic optimization model Neural Network Algorithm (NNA), is employed to examine the new GOA's outstanding performance in tuning SVR's hyperparameters [66]. SVR-PSO-BA_K, SVR-COA-BA_K, and SVR-NNA-BA_K models are established to evaluate the novel optimization algorithm GOA for 1-hour ahead GHI forecasting compared to the recent optimizers of PSO, COA, and NNA. K in the later models is equal to the optimal number of inputs found with SVR-GOA-BA_K for each site. In short, the fundamental algorithms

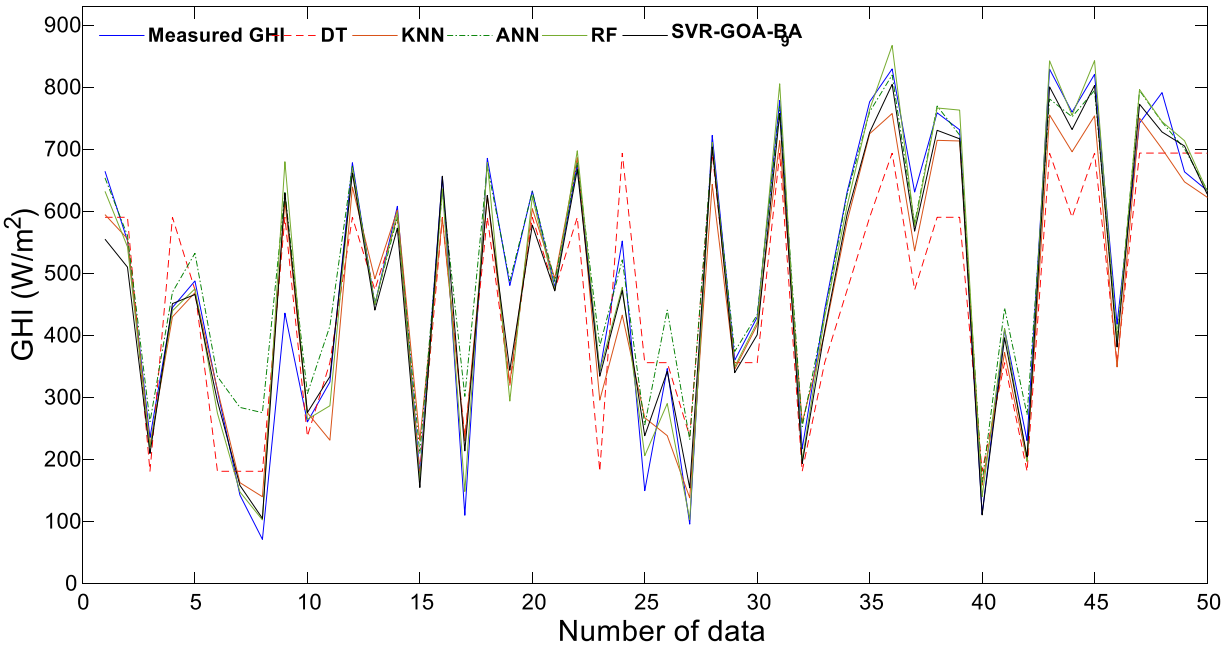


Fig. 13. Forecasting results of the proposed model and the contrast algorithms for Riyadh.

are chosen from among those new optimization algorithms as there have been numerous optimization algorithms proposed in recent years.

Fig. 15 depicts the predictive accuracy of these four hybrid models for the three considered sites. The numerical results for the SVR's hyperparameters tuning by the four considered optimizers are shown in Table 6. For the same GHI datasets, the predictive accuracy of SVR-GOA-BA_K is greater than that of SVR-PSO-BA_K, SVR-COA-BA_K, and SVR-NNA-BA_K, as it can be drawn from Table 6. Consequently, GOA's ability to search for the SVR's hyperparameters is more efficient than PSO, COA, and NAA optimizers. Simply put, the SVR-GOA-BA_K method achieves high accuracy for 1-hour GHI forecasting and has the best efficiency and consistency among all basic techniques. As a result, the hybrid SVR-GOA-BA_K model is chosen for predicting in the study, and future researchers are advised to use it in their research based on their interests.

5. Conclusions and future work

Forecasting global horizontal irradiance (GHI) levels is vital because of its positive effects on climate changes, environmental pollutions, renewable-energy-based innovations, economic-social growth, etc. Due to the high levels of random patterns in the time series of GHI datasets, this process is considered immensely strenuous. The SVR-GOA-BA_K model was proposed in this paper, a hybrid learning approach to hourly-level predictive analytics of GHI future values. The experimental studies were conducted on Saudi Arabia's hourly GHI figures from Dhahran, Riyadh, and Jeddah cities. The critical research results are summarized below:

- In time-series forecasting models, utilizing lag observations of solar energy components as defining features has appeared as an ideal strategy. They can be used efficiently to predict future values of GHI.
- In combination with SVR, the Boruta-based feature selection algorithm provided superior predictions on all sites' datasets of this paper interest.
- Grasshopper optimization algorithm (GOA), an advanced optimizer, is proven to work effectively in searching for the optimal values of SVR's hyperparameters for the GHI prediction with high precision. The explanation for this effectiveness is that this optimizer has a compelling capacity to discover and harness the optimal solutions to high-dimensional problems.
- The proposed strategy SVR-GOA-BA_Khas outperformed distinct models such as standalone SVR, DT, KNN, ANN, and RF. Thus, the SVR-GOA-BA_K model is an effective method for forecasting GHI values at the hourly level.
- Among the five competing models, the RF model has statistically achieved better outcomes than the remaining four models. RF's ensemble nature is primarily responsible for this.
- Regarding SVR-GOA-BA_K's prediction findings, a conclusion can be made that the Jeddah site is relatively more predictable than the other two cities.

In short, the paper's final results can be fruitfully used for the stakeholders' practical intents. As Dhahran and Riyadh's sites are found to be less predictable, GHI values are less stable than the location in Jeddah. Subsequently, lawmakers can amend policies concerning sustainable growth and accelerate the development of alternative-energy-based projects. The findings acquired should also assist solar energy developers and decision-makers.

The predictive performance of the proposed model is superior to others, as evidenced by the series of outcomes presented earlier in this study. It can be used in other applications of interest. However, the regression problem that is to be solved by the proposed model is limited to the one-hour-ahead GHI prediction. Also, it should be noted that the findings of the proposed model are only restricted to the locations of interest or sites with similar climate conditions. Regardless of the superior results, the input variables used in the training phase of the predictive strategy are limited to the available aforesaid meteorological parameters. Furthermore, despite the multiple capabilities of the powerful GOA, its use in this work is limited to tuning the SVR's hyperparameters.

Based on this discussion, we recommend that additional research could be conducted. Firstly, more potential implementations of the developed method in tackling other regression

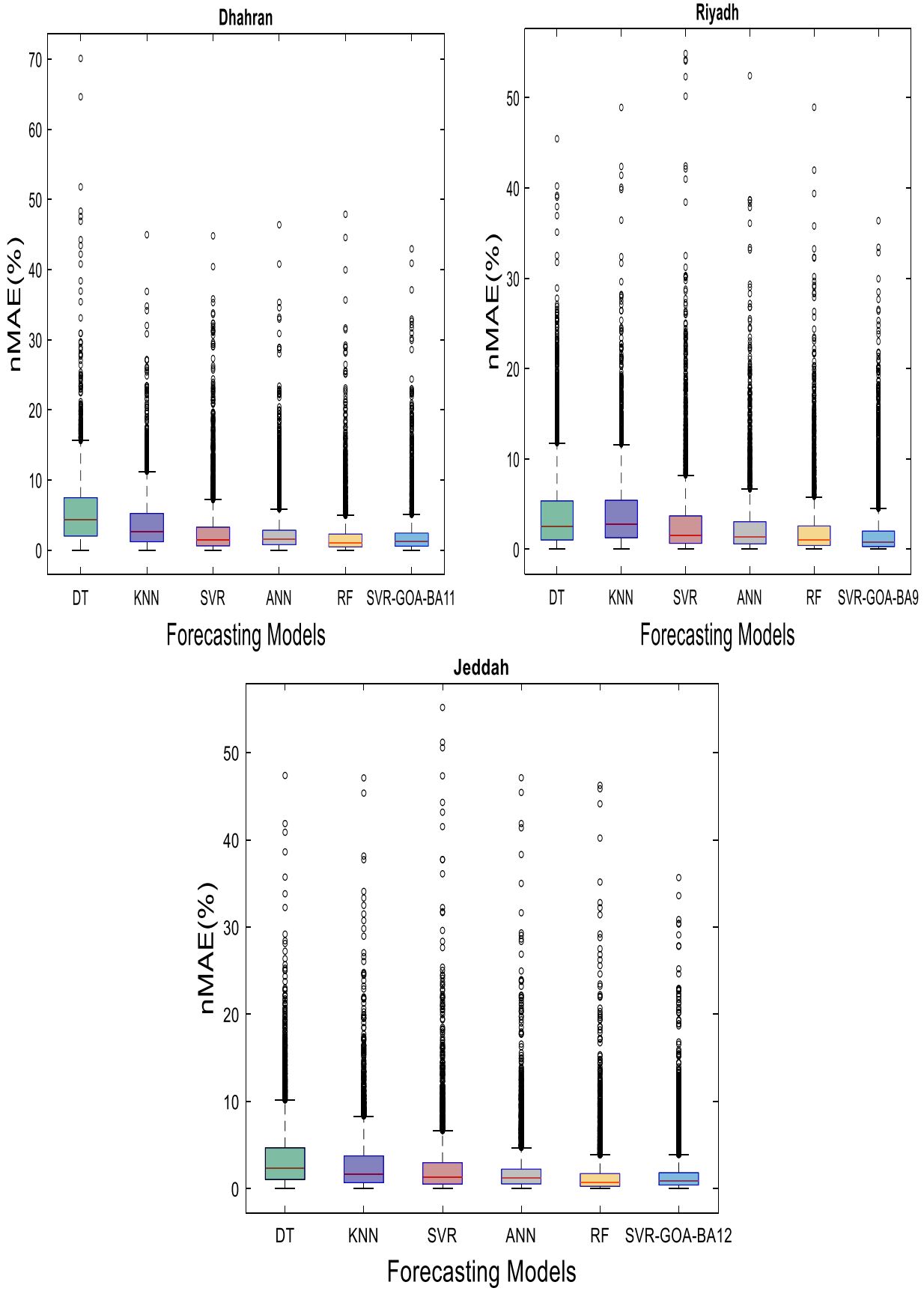


Fig. 14. Boxplots of the nMAE encountered by the developed forecasting models for all tested locations in Saudi Arabia.

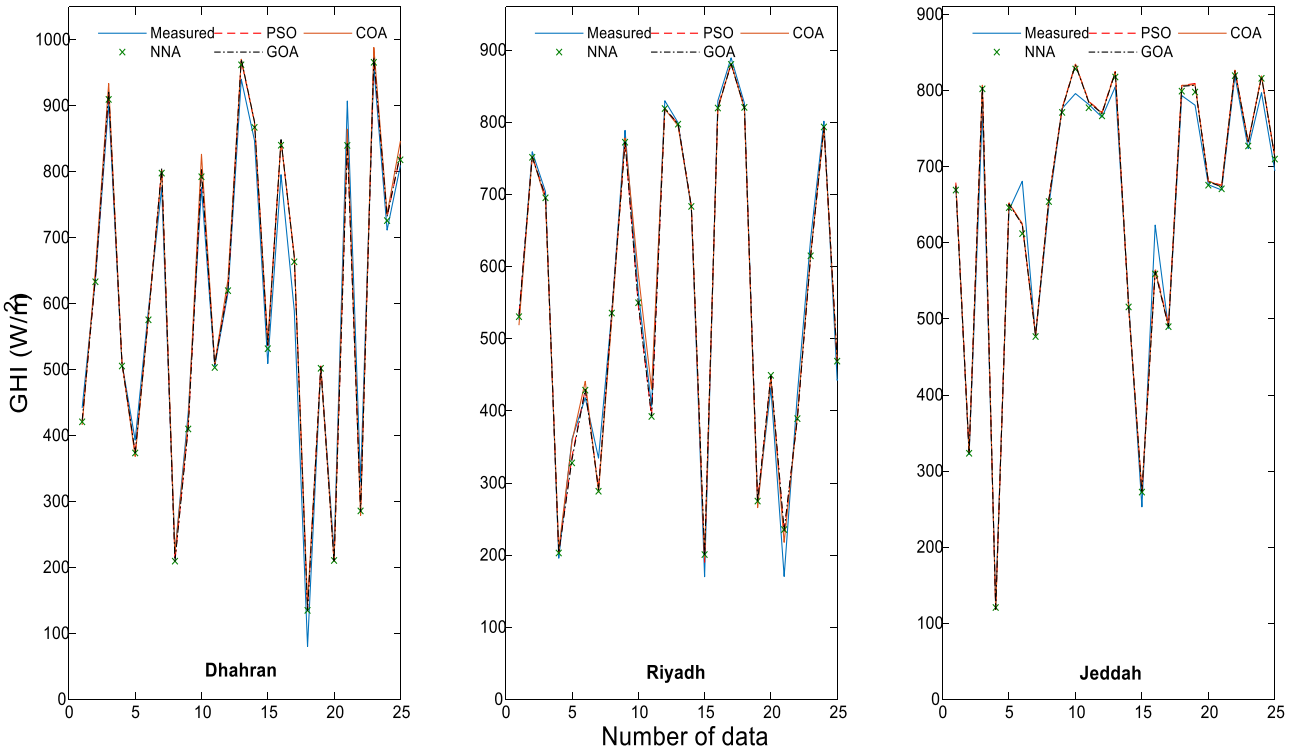


Fig. 15. The results of the proposed model with different optimization techniques.

Table 6
Performance comparison of different optimizers for GHI forecasting.

Model	Criteria	Dhahran	Riyadh	Jeddah
SVR-GOA-BA_K	RMSE (W/m^2)	45.0903	49.8129	41.1592
	nRMSE (%)	4.35	4.67	3.96
	R^2	0.98823481	0.98863249	0.98883136
	MAPE (%)	9.13	7.24	6.45
	MAE (W/m^2)	24.7827	23.7835	18.7913
	nMAE (%)	2.37	2.18	1.78
SVR-PSO-BA_K	RMSE (W/m^2)	46.0340	49.9497	42.7246
	nRMSE (%)	4.41	4.69	4.0486
	R^2	0.98763844	0.98843364	0.98783721
	MAPE (%)	9.14	7.29	7.2544
	MAE (W/m^2)	24.6016	23.9754	20.6106
	nMAE (%)	2.39	2.20	1.95
SVR-COA-BA_K	RMSE (W/m^2)	50.2129	51.0776	42.604330
	nRMSE (%)	4.8097	4.6938	4.03717
	R^2	0.98545329	0.9880360	0.987940418
	MAPE (%)	10.06	7.3045	7.2436
	MAE (W/m^2)	28.6227	24.0608	20.47934
	nMAE (%)	2.7416	2.21	1.94
SVR-NNA-BA_K	RMSE (W/m^2)	49.78560	49.94471	42.03924
	nRMSE (%)	4.768736	4.68	3.9836
	R^2	0.98565396	0.988507931	0.988257265
	MAPE (%)	9.932	7.27	7.3080
	MAE (W/m^2)	28.3724	23.8033	19.7262
	nMAE (%)	2.72	2.19	1.87

* $K = 11$ for Dhahran, $K = 9$ for Riyadh, and $K = 12$ for Jeddah.

problems about the energy data analysis field should be investigated. Secondly, the predictive performance of the proposed model in predicting the GHI values with different time horizons like intra-hour, daily, monthly, or annually should be tested. Thirdly, to ensure the generalization of the results of the proposed framework, more research on GHI forecasts in different locations is required, as the radiation on the earth's surface varies depending on the climatic conditions of the place. Fourthly, the utilization of GOA as a feature selection tool should be investigated and its performance then compared with the BA. Fifthly, studying the potential of GOA as an optimizer to tune the hyperparameters of other machine learning algorithms is needed, and then compare the results with this study's findings. Finally, future research should compare our approach to other cutting-edge algorithms, especially deep learning algorithms, as they are beyond the scope of this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Massoud Alrashidi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Musaed Alrashidi:** Methodology, Software, Validation, Formal analysis, Data curation, Visualization. **Saifur Rahman:** Validation, Formal analysis, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We sincerely thank the Qassim University of Saudi Arabia, King Abdullah City for Atomic and Renewable Energy of Saudi Arabia, and Advanced Research Institute at Virginia Tech, USA, for financial, material, and educational supports, respectively.

References

[1] D. Londoño-Pulgarin, G. Cardona-Montoya, J.C. Restrepo, F. Muñoz-Leiva, Fossil or bioenergy? Global fuel market trends, Renew. Sustain. Energy Rev. 143 (2021) 110905, http://dx.doi.org/10.1016/j.rser.2021.110905.

[2] J. Dong, M.M. Olama, T. Kuruganti, A.M. Melin, S.M. Djouadi, Y. Zhang, Y. Xue, Novel stochastic methods to predict short-term solar radiation and photovoltaic power, Renew. Energy 145 (2020) 333–346, http://dx.doi.org/10.1016/j.renene.2019.05.073.

[3] O. Ruhnau, P. Hennig, R. Madlener, Economic implications of forecasting electricity generation from variable renewable energy sources, Renew. Energy 161 (2020) 1318–1327, http://dx.doi.org/10.1016/j.renene.2020.06.110.

[4] M. Amir, S.Z. Khan, Assessment of renewable energy: status, challenges, COVID-19 impacts, opportunities, and sustainable energy solutions in Africa, Energy Built Environ. (2021) http://dx.doi.org/10.1016/j.enbenv.2021.03.002.

[5] H. Tang, S. Wang, H. Li, Flexibility categorization, sources, capabilities and technologies for energy-flexible and grid-responsive buildings: State-of-The-Art and future perspective, Energy (2020) 119598, http://dx.doi.org/10.1016/j.energy.2020.119598.

[6] C. Lupangu, R.C. Bansal, A review of technical issues on the development of solar photovoltaic systems, Renew. Sustain. Energy Rev. 73 (2017) 950–965, http://dx.doi.org/10.1016/j.rser.2017.02.003.

[7] B. Li, J. Zhang, A review on the integration of probabilistic solar forecasting in power systems, Sol. Energy 210 (2020) 68–86, http://dx.doi.org/10.1016/j.solener.2020.07.066.

[8] G. Reikard, Comment on verification of deterministic solar forecasts: Choice of models, and estimation procedure, Sol. Energy 210 (2020) 47–48, http://dx.doi.org/10.1016/j.solener.2020.04.018.

[9] H. Tao, A.A. Ewees, A.O. Al-Sulttani, U. Beyaztas, M.M. Hameed, S.Q. Salih, A.M. Armanuos, N. Al-Ansari, C. Voyant, S. Shahid, Z.M. Yaseen, Global solar radiation prediction over north dakota using air temperature: Development of novel hybrid intelligence model, Energy Rep. 7 (2021) 136–157, http://dx.doi.org/10.1016/j.egyr.2020.11.033.

[10] A. Masoumi, S. Ghassem-zadeh, S.H. Hosseini, B.Z. Ghavidel, Application of neural network and weighted improved PSO for uncertainty modeling and optimal allocating of renewable energies along with battery energy storage, Appl. Soft Comput. 88 (2020) 105979, http://dx.doi.org/10.1016/j.asoc.2019.105979.

[11] F.J. Rodríguez-Benítez, C. Arbizu-Barrena, J. Huertas-Tato, R. Aler-Mur, I. Galván-León, D. Pozo-Vázquez, A short-term solar radiation forecasting system for the Iberian Peninsula. Part 1: Models description and performance assessment, Sol. Energy 195 (2020) 396–412, http://dx.doi.org/10.1016/j.solener.2019.11.028.

[12] S. Ghimire, R.C. Deo, N.J. Downs, N. Raj, Global solar radiation prediction by ANN integrated with European centre for medium range weather forecast fields in solar rich cities of Queensland Australia, J. Clean. Prod. 216 (2019) 288–310, http://dx.doi.org/10.1016/j.jclepro.2019.01.158.

[13] M. Emmanuel, K. Doubleday, B. Cakir, M. Marković, B.-M. Hodge, A review of power system planning and operational models for flexibility assessment in high solar energy penetration scenarios, Sol. Energy 210 (2020) 169–180, http://dx.doi.org/10.1016/j.solener.2020.07.017.

[14] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renew. Energy 105 (2017) 569–582, http://dx.doi.org/10.1016/j.renene.2016.12.095.

[15] A. Alfadda, S. Rahman, M. Pipattanasomporn, Solar irradiance forecast using aerosols measurements: A data driven approach, Sol. Energy 170 (2018) 924–939, http://dx.doi.org/10.1016/j.solener.2018.05.089.

[16] S. Ghimire, R.C. Deo, N. Raj, J. Mi, Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms, Appl. Energy 253 (2019) 113541, http://dx.doi.org/10.1016/j.apenergy.2019.113541.

[17] A.N.-L. Huynh, R.C. Deo, D.-A. An-Vo, M. Ali, N. Raj, S. Abdulla, Near real-time global solar radiation forecasting at multiple time-step horizons using the long short-term memory network, Energies 13 (2020) 3517, http://dx.doi.org/10.3390/en13143517.

[18] M. Aslam, J.-M. Lee, H.-S. Kim, S.-J. Lee, S. Hong, Deep learning models for long-term solar radiation forecasting considering microgrid installation: A comparative study, Energies 13 (2020) 147, http://dx.doi.org/10.3390/en13010147.

[19] P. Lauret, C. Voyant, T. Soubdhan, M. David, P. Poggi, A benchmarking of machine learning techniques for solar radiation forecasting in an insular context, Sol. Energy 112 (2015) 446–457, http://dx.doi.org/10.1016/j.solener.2014.12.014.

[20] M.H. Alsharif, M.K. Younes, J. Kim, Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea, Symmetry (Basel) 11 (2019) 240, http://dx.doi.org/10.3390/sym11020240.

[21] D.H.W. Li, W. Chen, S. Li, S. Lou, Estimation of hourly global solar radiation using multivariate adaptive regression spline (MARS) – A case study of Hong Kong, Energy 186 (2019) 115857, http://dx.doi.org/10.1016/j.energy.2019.115857.

[22] A.K. Yadav, H. Malik, S.S. Chandel, Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India, Renew. Sustain. Energy Rev. 52 (2015) 1093–1106, http://dx.doi.org/10.1016/j.rser.2015.07.156.

[23] L.M. Halabi, S. Mekhilef, M. Hossain, Performance evaluation of hybrid adaptive neuro-fuzzy inference system models for predicting monthly global solar radiation, Appl. Energy 213 (2018) 247–261, http://dx.doi.org/10.1016/j.apenergy.2018.01.035.

[24] N. Dong, J.-F. Chang, A.-G. Wu, Z.-K. Gao, A novel convolutional neural network framework based solar irradiance prediction method, Int. J. Electr. Power Energy Syst. 114 (2020) 105411, http://dx.doi.org/10.1016/j.ijepes.2019.105411.

[25] S. Monjoly, M. André, R. Calif, T. Soubdhan, Hourly forecasting of global solar radiation based on multiscale decomposition methods: A hybrid approach, Energy 119 (2017) 288–298, http://dx.doi.org/10.1016/j.energy.2016.11.061.

[26] A. Zendehboudi, M.A. Baseer, R. Saidur, Application of support vector machine models for forecasting solar and wind energy resources: A review, J. Clean. Prod. 199 (2018) 272–285, http://dx.doi.org/10.1016/j.jclepro.2018.07.164.

[27] R. Meenal, A.I. Selvakumar, Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters, Renew. Energy 121 (2018) 324–343, http://dx.doi.org/10.1016/j.renene.2017.12.005.

[28] S. Belaid, A. Mellit, H. Boualit, M. Zaiani, Hourly global solar forecasting models based on a supervised machine learning algorithm and time series principle, Int. J. Ambient Energy (2020) 1–12, http://dx.doi.org/10.1080/01430750.2020.1718754.

[29] S. Belaid, A. Mellit, Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate, Energy Convers. Manag. 118 (2016) 105–118, http://dx.doi.org/10.1016/j.enconman.2016.03.082.

[30] M. Guermoui, A. Rabehi, K. Gairaa, S. Benkaciali, Support vector regression methodology for estimating global solar radiation in Algeria, Eur. Phys. J. Plus 133 (2018) 1–9, http://dx.doi.org/10.1140/epjp/i2018-11845-y.

[31] Ü. Ağbulut, A.E. Gürel, Y. Biçen, Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison, Renew. Sustain. Energy Rev. 135 (2021) 110114, http://dx.doi.org/10.1016/j.rser.2020.110114.

[32] M.A. Hassan, A. Khalil, S. Kaseb, M.A. Kassem, Exploring the potential of tree-based ensemble methods in solar radiation modeling, Appl. Energy 203 (2017) 897–916, http://dx.doi.org/10.1016/j.apenergy.2017.06.104.

[33] V.H. Quej, J. Almorox, J.A. Arnaldo, L. Saito, ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment, J. Atmos. Solar-Terrestrial Phys. 155 (2017) 62–70, http://dx.doi.org/10.1016/j.jastp.2017.02.002.
[34] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petković, C. Sudheer, A support vector machine–firefly algorithm-based model for global solar radiation prediction, Sol. Energy 115 (2015) 632–644, http://dx.doi.org/10.1016/j.solener.2015.03.015.
[35] M. Alrashidi, S. Rahman, M. Pipattanasomporn, Metaheuristic optimization algorithms to estimate statistical distribution parameters for characterizing wind speeds, Renew. Energy 149 (2020) 664–681, http://dx.doi.org/10.1016/j.renene.2019.12.048.
[36] H. Bouzgou, C.A. Gueymard, Minimum redundancy – Maximum relevance with extreme learning machines for global solar radiation forecasting: Toward an optimized dimensionality reduction for solar time series, Sol. Energy 158 (2017) 595–609, http://dx.doi.org/10.1016/j.solener.2017.10.035.
[37] D. Niu, K. Wang, L. Sun, J. Wu, X. Xu, Short-term photovoltaic power generation forecasting based on random forest feature selection and CEEMD: A case study, Appl. Soft Comput. 93 (2020) 106389, http://dx.doi.org/10.1016/j.asoc.2020.106389.
[38] Y. Zhou, Y. Liu, D. Wang, X. Liu, Y. Wang, A review on global solar radiation prediction with machine learning models in a comprehensive perspective, Energy Convers. Manag. 235 (2021) 113960, http://dx.doi.org/10.1016/j.enconman.2021.113960.
[39] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, R. García-Herrera, Feature selection in machine learning prediction systems for renewable energy applications, Renew. Sustain. Energy Rev. 90 (2018) 728–741, http://dx.doi.org/10.1016/j.rser.2018.04.008.
[40] M. Almaraashi, Investigating the impact of feature selection on the prediction of solar radiation in different locations in Saudi Arabia, Appl. Soft Comput. 66 (2018) 250–263, http://dx.doi.org/10.1016/j.asoc.2018.02.029.
[41] A. Aybar-Ruiz, S. Jiménez-Fernández, L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo, P. Salvador-González, S. Salcedo-Sanz, A novel grouping genetic algorithm–extreme learning machine approach for global solar radiation prediction from numerical weather models inputs, Sol. Energy 132 (2016) 129–142, http://dx.doi.org/10.1016/j.solener.2016.03.015.
[42] S. Salcedo-Sanz, R.C. Deo, L. Cornejo-Bueno, C. Camacho-Gómez, S. Ghimire, An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia, Appl. Energy 209 (2018) 79–94, http://dx.doi.org/10.1016/j.apenergy.2017.10.076.
[43] S. Saremi, S. Mirjalili, A. Lewis, Grasshopper optimisation algorithm: theory and application, Adv. Eng. Softw. 105 (2017) 30–47, http://dx.doi.org/10.1016/j.advengsoft.2017.01.004.
[44] A. Khosravi, R.N.N. Koury, L. Machado, Thermo-economic analysis and sizing of the components of an ejector expansion refrigeration system, Int. J. Refrig. 86 (2018) 463–479, http://dx.doi.org/10.1016/j.ijrefrig.2017.11.007.
[45] The World Bank, Solar resource data: Solargis, Glob. Sol. Atlas 2.0, 2019, https://solargis.com/maps-and-gis-data/download/saudi-arabia (accessed August 28, 2020).
[46] K.A. CARE, Renewable Resource Atlas, King Abdullah City At. Renew. Energy (K.A.CARE), Saudi Arab, 2015, https://rratlas.energy.gov.sa/ (accessed August 28, 2020).
[47] H.P. Dasari, S. Desamsetti, S. Langodan, R. Attada, R.K. Kunchala, Y. Viswanadhapalli, O. Knio, I. Hoteit, High-resolution assessment of solar energy resources over the Arabian Peninsula, Appl. Energy 248 (2019) 354–371, http://dx.doi.org/10.1016/j.apenergy.2019.04.105.
[48] A. Rohani, M. Taki, M. Abdollahpour, A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I), Renew. Energy 115 (2018) 411–422, http://dx.doi.org/10.1016/j.renene.2017.08.061.
[49] R. Blaga, A. Sabadus, N. Stefu, C. Dughir, M. Paulescu, V. Badescu, A current perspective on the accuracy of incoming solar energy forecasting, Prog. Energy Combust. Sci. 70 (2019) 119–144, http://dx.doi.org/10.1016/j.pecs.2018.10.003.
[50] V. Vapnik, The Nature of Statistical Learning Theory, 2nd edition, Springer science & business media, 2013.
[51] M.B. Kursa, Robustness of random forest-based gene selection methods, BMC Bioinformatics 15 (2014) 8, http://dx.doi.org/10.1186/1471-2105-15-8.
[52] M.K. Goyal, B. Bharti, J. Quilty, J. Adamowski, A. Pandey, Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS, Expert Syst. Appl. 41 (2014) 5267–5276, http://dx.doi.org/10.1016/j.eswa.2014.02.047.
[53] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (2010) 1–13, http://dx.doi.org/10.18637/jss.v036.i11.
[54] R. Lippmann, An introduction to computing with neural nets, IEEE Assp Mag. 4 (1987) 4–22, http://dx.doi.org/10.1109/MASSP.1987.1165576.
[55] R.C. Deo, M. ¸ Sahin, Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland, Renew. Sustain. Energy Rev. 72 (2017) 828–848, http://dx.doi.org/10.1016/j.rser.2017.01.114.
[56] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control. Signals Syst. 2 (1989) 303–314, http://dx.doi.org/10.1007/BF02551274.
[57] Ö. ¸ Celik, A. Teke, H.B. Yıldırım, The optimized artificial neural network model with Levenberg–Marquardt algorithm for global solar radiation estimation in Eastern Mediterranean Region of Turkey, J. Clean. Prod. 116 (2016) 1–12, http://dx.doi.org/10.1016/j.jclepro.2015.12.082.
[58] S.A. Kalogirou, Applications of artificial neural networks in energy systems, Energy Convers. Manag. 40 (1999) 1073–1087, http://dx.doi.org/10.1016/S0196-8904(99)00012-6.
[59] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC press, 1984, http://dx.doi.org/10.1002/cyto.990080516.
[60] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, 2013, http://dx.doi.org/10.1007/978-1-4614-6849-3.
[61] J. Cai, K. Xu, Y. Zhu, F. Hu, L. Li, Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest, Appl. Energy 262 (2020) 114566, http://dx.doi.org/10.1016/j.apenergy.2020.114566.
[62] L. Benali, G. Notton, A. Fouilloy, C. Voyant, R. Dizene, Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components, Renew. Energy 132 (2019) 871–884, http://dx.doi.org/10.1016/j.renene.2018.08.044.
[63] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.
[64] M. Oprea, A general framework and guidelines for benchmarking computational intelligence algorithms applied to forecasting problems derived from an application domain-oriented survey, Appl. Soft Comput. 89 (2020) 106103, http://dx.doi.org/10.1016/j.asoc.2020.106103.
[65] R. Rajabioun, Cuckoo optimization algorithm, Appl. Soft Comput. 11 (2011) 5508–5518, http://dx.doi.org/10.1016/j.asoc.2011.05.008.
[66] A. Sadollah, H. Sayyaadi, A. Yadav, A dynamic metaheuristic optimization model inspired by biological nervous systems: Neural network algorithm, Appl. Soft Comput. 71 (2018) 747–782, http://dx.doi.org/10.1016/j.asoc.2018.07.039.