Forecasting the Global Horizontal Irradiance based on Boruta Algorithm and Artificial Neural Networks using a Lower Cost

Abdulatif Aoihan Alresheedi[1], Mohammed Abdullah Al-Hagery[2]
Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia[1, 2]
BIND Research Group, College of Computer, Qassim University, Buraydah, Saudi Arabia[2]

***Abstract*—More solar-based electricity generation stations have been established markedly in recent years as new and an important source of renewable energy. That is to ensure a more efficient, reliable integration of solar power to overcome several challenges such as, the future forecasting, the costly equipment in the metrological stations. One of the effective prediction methods is Artificial Neural Networks (ANN) and the Boruta algorithm for optimal attributes selection, to train the proposed prediction model to obtain high accurate prediction performance at a lower cost. The precise goal of this research is to predict the Global Horizontal Irradiance (GHI) by building the ANN model. Also, reducing the total number of GHI prediction attributes/features consequently reducing the cost of devices and equipment required to predict this important factor. The dataset applied in this research is real data, collected from 2015-2018 by solar and meteorological stations in KSA. It provided by King Abdullah City for Atomic and Renewable Energy (KA CARE). The findings emphasize the achievement of accurate predictions of solar radiation with a minimum cost, which is considered to be highly important in KSA and all other countries that have a similar environment.**

***Keywords*—Global horizontal irradiance; artificial neural networks; feature selection; boruta algorithm; cost reduction; machine learning**

I. Introduction

Alternative energy sources increasingly form the future of the world's energy system. This is due to fossil fuel resource limitations as well as their negative side effects on climate change and environmental pollution. The objective of sustainable power supply can be achieved by renewable energy sources, such as solar power, which still unused and it is characterized by high variability in availability and production. Rapid changes in solar power output are one of the negative consequences of rapid changes in weather conditions. The intermittent nature of renewable energy sources might hinder electrical utilities from effectively utilizing them.

Higher penetrations of solar energy into the electrical grid cause a more variable power output than with higher penetrations of wind [1]. Moreover, higher penetrations of alternative sources lead to power system technical operation and design issues, such as systems protection, systems control, power factor quality, and optimal operation of power systems [2]. Also, to manage the variability and the uncertainty in solar power, that is due to high penetrations of renewables, adjustments to the power systems operation are needed including adding new ancillary services [3]. Thus, the economic feasibility of renewable energy sources is negatively affected by the expensive costs of these adjustments and requirements.

Many potential solutions can manage technical issues caused by short-term uncertainty in solar power (up to seven days ahead) [4]. For instance, increasing the level of demand-side participation, increasing the level of coordination to balance the allocation, and deploying more flexible but often also costlier energy storage systems. Still, the prediction of global solar radiation is one of the most efficient and economical ways to integrate more solar power, especially at current levels of integration. These forecasts can be utilized by balancing authorities to operate electric power systems more efficiently and reliably. In the literature, several forecasting approaches have been embraced [5]. Among these, Machine Learning (ML) algorithms are currently the most common methods to predict solar energy, because prediction is an important step in designing and assessing photovoltaic systems technically and economically.

Predictions of solar irradiance using ML algorithms were proposed in several studies, after the advent of fast computing capabilities as well as systems able to store massive data sets [6]–[9]. The ML methods include ANN, support vector regression (SVR), decision tree regression, and K-Nearest Neighbors, other methods [10], [11]. ANN is considered to be the most powerful ML method because of its capability to intrinsically deal with the nonlinear nature of solar and meteorological data. In recent studies, ANNs resulted in a lower mean absolute percentage error.

ANN was used, for example, to predict the electric load of Tai's power system [12]. Besides, in Salerno, Italy, two ANN models were developed to predict GHI and direct normal irradiance (DNI) in an hourly manner [13]. In the two later studies, the ANN model resulted in good accuracy. Also, ANN ensemble methods were applied by Alobaidi et al. to forecast the solar radiation variables utilizing satellite images. Five locations in the state of the United Arab Emirates were selected to apply the developed ANN prediction models [14]. As an application to forecasting one day-ahead solar radiation in a grid-connected-PV system, Mellit and Pavan developed ANNs that use mean daily solar radiation as well as air temperature as inputs into the prediction system [15]. Also, Lam et al.

employed the ANN to forecast the daily GHI in 40 different cities in China utilizing the observed duration of sunshine, and that study targeted areas with various thermal climates as well as sub-areas [16]. Moreover, a combination model of numerical weather prediction, using a hybridized autoregressive moving average and ANN algorithms were developed for forecasting the short-term GHI as presented [17].

An advanced embedded feature selection algorithm is known as the Boruta algorithm was implemented in this paper to choose from 13 available attributes in the dataset the most significant attributes. To the best of the author's knowledge, the feature selection approach employed by this research has not been applied to this issue before. This adds a great contribution to this paper. Since the analysis involves big datasets, authors embrace the powerful machine learning algorithm of ANN as a means of the computing system. The type of ANN used in this paper is a Multilayer-Feed-forward Back Propagation (MFBP) Network. In short, this paper introduces a novel, intelligent, a hybrid framework consisting of the ANN algorithm to conduct the training and testing processes and Boruta algorithm to select the most important features to be inputted into the ANN model. As a result, this paper introduces an ML-based model to predict GHI for each location of interest-based on the optimal number of attributes and it is an extension of limited work in [18], which focused partially on Buridah city in KSA.

The rest of this paper is organized as follows: Section 2 highlights the framework of the developed methodology including the data used and the used ML algorithms. Section 3, presents an overview of the feature selection technique used. Section 4 discusses the validation measures and metrics used to evaluate the accuracy of the proposed model. Besides, the analysis and discussion of the results were placed in Section 5. Finally, Section 6 explains the conclusions and future work.

II. Literature Review

There are many research works concentrated in the field of energy, electricity generation, solar energy prediction methods based on ML algorithms, and other methods were based on mathematical models.

Several ML algorithms employed in the energy field for prediction purpose, such as the Support Vector Machine (SVM). This is because of the ability of SVM to model nonlinearity exists in time-series metrological data. Utilizing SVR applications to predict GHI was used and the results reveal feasibility and accurate prediction performance [14]. Also, the study achieved by [19] developed an SVM model based on a firefly algorithm to predict GHI [20]. Performance comparisons between the developed model and ANN and genetic programming models were created, and the results demonstrated that the enhanced SVR model has a better prediction accuracy. Also, another wavelet-based SVR model was developed in to predict GHI in different cities in Australia.

The research work by [21] concluded that the prediction accuracy in SVR models is directly proportional to the size of training data when SVM applied to predict electric load. Similarly, the SVR model accuracy considerably relies on selecting the optimal set of parameters. A proper determination of the optimal set of SVR models' parameters is not an easy task. To solve this problem, several advanced optimization techniques have been used such as particle Swarm Optimization algorithm, Immune algorithms [22]. Besides, Genetic algorithm, for example, has been used to optimally select the SVR model parameters to forecast electricity market prices [23]. In Saudi Arabia, past studies on models of solar radiation applied various computational methods, most of which belonged to methods of empirical or artificial intelligence. The researches by [24], [25] carried out forecasting of the average of GHI per annum with good accuracy. A nonlinear Angstrom-type model was used in their study and then was compared to Bulut and Büyükalaca's trigonometric function model in [26]. Also, a related study achieved in Oman [27] for measuring various features such as the temperature, humidity, and solar radiation. The study provided statistical results compared to the NASA SSE Model.

A geostatistical methodology was used by [28] to predict GHI in the Kingdom of Saudi Arabia. This study had the purpose of producing a geographically persistent mapping system of solar irradiance, and also for every single month of the year, to draw the solar irradiance contour maps. In a mission originating from 1994 to April 2000, solar radiation measurements were taken over twelve locations in 12 cities [29]. Utilizing features of latitude, longitude, altitude, the number of months, and sunshine duration, Mohandes and Rehman used a predictive approach to forecast GHI anywhere across Saudi Arabia [30], [31]. The experiment used the 35 stations solar radiation data to test the accuracy of the prediction model where the outcomes of the forecasted values were near to the observed values to some extent. Benghanem et al. employed ANN-based prediction models to forecast daily averages of global horizontal irradiance for five years' period through the use of National Renewable Energy Laboratory repositories. Using recent data sets given by KA CARE. But in our study we add the features selection technique, to improve the results.

Almaraashi used automated fuzzy logic systems aiming at forecasting the next day's solar radiation [32]. To the best of the author's knowledge and even though many ML-based models have been introduced in the Kingdom of Saudi Arabia, no automated methods of feature selection have been examined to forecast short-term solar radiation in Saudi Arabia. In addition to the mentioned works, there are some recent researches are concentrating on the GHI forecasting but with different datasets and different strategies.

For instance, in northeast Iraq [33], a research study accomplished on a Satellite Datasets to obtain a more accurate and precise method for forecasting hourly GHI. The proposed method established based on the ANN and another training algorithm called "Levenberg Marquardt" algorithm. The obtained results showed a very high accuracy. Besides, in South Africa [34], a research study carried out for discussing probabilistic of forecasting the GHI before 24 hours, using two machine learning methods and the data collected during the period from 2009 to 2010. The study gave excellent results but not exceeded by 95.5%.

Moreover, in Croatia [35], a research study concentrated on several models that are used for estimating solar radiation.

These models assessed based on seven meteorological stations dataset, where the models studied, compared and evaluated to find out the best accuracy. As well, in southern Finland [36], a research study carried out a GHI forecasting using a data set of weather satellite imagery, using a mathematical modelling method. The results obtained show very good accuracy. In this regards, it noticed that in many countries, the number of GHI measurement locations is sufficient as in KSA and insufficient in other countries as in Korea [37], [38], where the satellite images can be helpful sources for getting the GHI over a wide area space, in these cases, usually predicted by secondary parameters such as readily obtainable climatic variables.

On the other hand, regarding efficient attributes selection and efficient data preprocessing, the feature selection techniques are utilized for minimizing and preparing data with high dimensionality for ML-based problems. Such techniques are usually categorized into either supervised algorithms, requiring information about labels, or unsupervised algorithms, that operate without a need for information about labels. The challenge is that the solar radiation intensity is influenced by a large number of parameters. Thus, by removing redundant attributes, dimensionality reduction algorithms, as well as feature selection, might positively affect the prediction accuracy of developed forecasting models. Furthermore, the need for an optimized feature space grows when broad degrees of uncertainty is involved in the considered application. Several studies have identified and discussed the need to have a feature selection approach to be embraced before forecasting GHI.

For example, Salcedo-Sanz et al. [39], examined the usage of a species-optimizing coral reef algorithm to gain a decreased collection of important features to forecast GHI. Also, Yadav et al. implemented a set of features to some specific input predictors and observed the parameters of latitude and longitude are having the slightest impact on the forecasting of solar radiation [40]. Hedar et al. applied a programming-based algorithm of adaptive memory to minimize the space of input features of a fuzzy classifier for global solar radiation [41]. They found that among nine attributes, DHI, DNI, and relative humidity have the best dependence degree values.

This paper concentrates on the prediction of the GHI in two different regions in Saudi Arabia, by building Neural Networks models whose input variables are optimally and systemically selected by the features selection algorithm named Boruta, which was used for the first time to improve the GHI forecasting results.

III. Methodology

The overall methodology steps are listed as follows:

- Data-preprocessing tasks of all datasets used in this research are carried out ahead of the training, validation, and testing process of the proposed data-based model.
- The model trained and validated utilizing the all-feature set is established.
- In a similar way to (2), the model trained and validated utilizing the most important eight-feature set determined by the Boruta Algorithm is also built.
- Moreover, the model trained and validated utilizing the most important five-feature set by Boruta Algorithm is also developed.
- The predicted values of GHI by the model with different features are carried out through the testing process.
- To evaluate the performance accuracy of the developed forecasting model, the observed and the predicted values of the GHI are compared by a set of four evaluation metrics.

The Boruta is used to pick the most important variables among a wide range of meteorological variables that could impact solar radiation in the future. A prediction processed are then independently utilized, based on the Boruta's five and eight most important variables. Even though the targeted feature-selection-based model can be designed using any number of variables, we have fixed the number of features chosen to provide a fair comparison at the various locations of this study between the forecasting developed. Strictly speaking, in this case, a smaller set of features (that is, 5 and 8) must be made in advance among the 13 features.

A. Data Collection

To build the proposed forecasting model proposed in this paper, massively big observed solar and meteorological datasets are collected from KA CARE that provided more than 25 solar and metrological variables. The total number of observations (records) of the collected data is 35735 in the Qassim dataset, while in Jeddah city is 35856 observations before the pre-processing step. After the cleaning process, the data were reduced at the Qassim region to be 17892 and in Jeddah to be 19467 observations. The GHI observations under consideration are in 1-hour time resolutions for the period from March 1. 2013, out to June 30, 2017, and they are collected from the interest locations of Jeddah and Qassim. Furthermore, the corresponding 1-hour intervals weather variables are gathered. The whole dataset (x) is splatted into three subsets namely: the training dataset, $x_{training}$, the cross-validation dataset, $x_{cross\text{-}validation}$, and the testing dataset, $x_{testing}$, such that $X = x_{training} \cup x_{cross\text{-}validation} \cup x_{testing}$. In this research, the ratio of the training, cross-validation, and testing datasets are 6:2:2, respectively.

At this time, the variables involved in this analysis include 13 independent variables: month of the year (M), day of the month (D), an hour of the day (H), air temperature (T) in (℃), relative humidity (RH) in (%), surface pressure (P) in (hPa), wind speed at 3 meters (WS) in (m/s), Wind Direction (WD) at 3 meters in (°N), Peak Wind Direction (PWD) at 3 meters in (°N), diffuse horizontal irradiance (DHI) in (Wh/m^2), direct normal irradiance (DNI) in (Wh/m^2), azimuth angle (AA) in (Â°), and solar zenith angle (SZA). The GHI in (Wh/m^2), as a target variable, is measured with a Kipp & Zonen Pyranometer.

Future work will involve an increased number of the observed features of the data as inputs into a novel proposed model. The input variables to the prediction model can be summarized in Table I. The prediction model can be expressed, in the scope of the used predictors, as shown in equation (1).

$$GHI_{predicted} = f(M, D, H, T, RH, P, WS, WD, PWD, DHI, DNI, AA, SZA) \quad (1)$$

TABLE I. NPUT VARIABLES TO THE PROPOSED PREDICTION MODEL

Input variable	Input variable abbreviation	Input variable explanation	Input variable unit
$x_1^{(i)}$	M	the month of the year	Month
$x_2^{(i)}$	D	day of the month	day
$x_3^{(i)}$	H	hour of the day	hour
$x_4^{(i)}$	T	air temperature	℃
$x_5^{(i)}$	RH	relative humidity	%
$x_6^{(i)}$	P	surface pressure	hPa
$x_7^{(i)}$	WS	wind speed at 3 meters	m/s
$x_8^{(i)}$	WD	Wind Direction	°N
$x_9^{(i)}$	PWD	the peak wind direction at 3 meters	°N
$x_{10}^{(i)}$	DHI	diffuse horizontal irradiance	Wh/m^2
$x_{11}^{(i)}$	DNI	direct normal irradiance	Wh/m^2
$x_{12}^{(i)}$	AA	azimuth angle	Â°
$x_{13}^{(i)}$	SZA	solar zenith angle	Â°

B. Data Preprocessing

After a thorough inspection of the used datasets from KA CARE, we notice that there are some errors, noise, redundant records. For that, some of the data cleaning steps are applied. These steps are very important to have high-quality datasets because unclean data can decrease the classification or regression model accuracies [42]. Fig. 1 shows the flowchart of the data preprocessing steps, and these steps can be summarized in the below sections.

1) Organizing dataset: It is a process of transforming the data received to a common format to improve visualizing and dealing with the data.

2) Removing redundant data: The performance of the forecasting algorithms depends mainly on the amount and accuracy of the used data. Using redundant data to train and test the algorithms will make the model computationally expensive as well as increase the time of executing the algorithms. The metrological datasets received from KA CARE contain some duplicated data for the same features. For example, for Air Temperature, there are data for actual temperature as well as data for uncertainty in Air Temperature. The uncertainty in Air Temperature adds no value and is redundant data. Therefore, it should be removed to avoid any complications in the data. Similarly, for wind direction, wind speed, DHI, DNI, GHI, peak wind speed, relative humidity, and barometric pressure.

3) Monitoring data errors: After organizing the data and from initial scanning, the data file shows that there are missing data entries in some features. This is shown in the data file as empty cells and MATLAB as (NaN), meaning Not a Number in a numerical file. In our final dataset, the entire day at any of the unrecorded features is removed when creating the forecasting model.

4) Feature construction/selection: Where new attributes (features) are constructed and added from the given set of attributes to help the mining process. The format of the date of KA CARE is as follows: 01/01/2014, 12:00:00 AM (MM/DD/YYYY HH:mm: SS), where MM: month, DD: day, YYYY: year, HH: hour, mm: minute and SS: second. Day, Month, and Hour are used as features for training and testing in the forecasting algorithms.

These attributes are needed to be re-constructed to improve the mining process because the date format is not suitable for advanced mining. Therefore, splitting these variables into different columns is necessary.

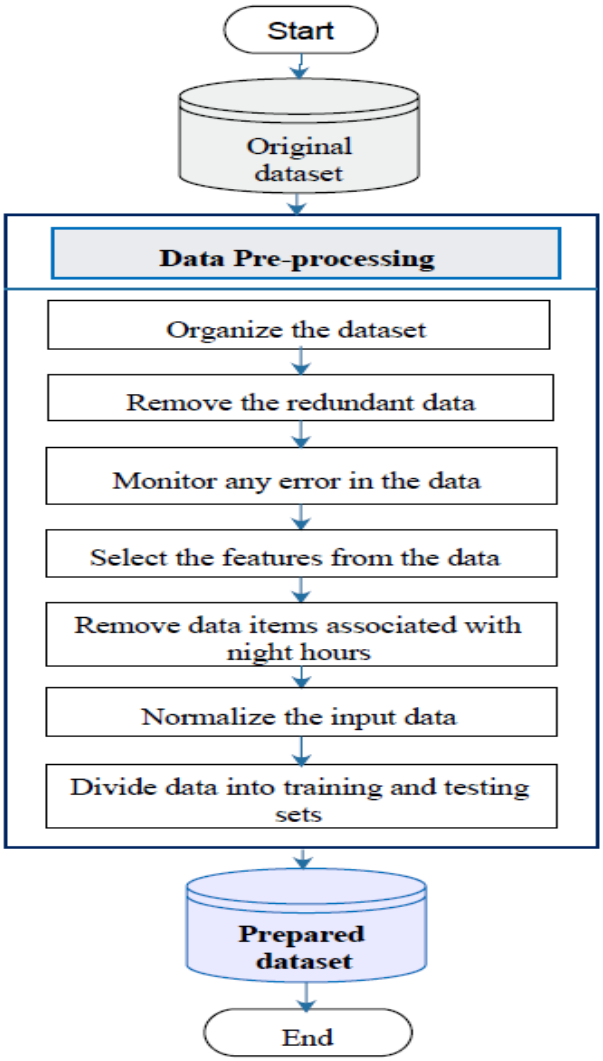

Fig. 1. Diagram of Data Preprocessing Steps.

The MATLAB code used to separate the data cells into single cells for the month, day, and hour. However, one of the problems encountered during preparing the data is inconstancy in the Date format. At each year and in each month of the year, the format of date from day 1 to day 12 is (Day, Month, Hour) while the reaming days' format is (Month, Day, Hour). This required creating a MATLAB code to change the format from day 1 to day 12 to make it smooth with the rest of the months' days (Month, Day, Hour). Accomplishing this task should be automatic because doing this manually is a very difficult task and time-consuming since we are dealing with a very huge dataset.

5) Removing GHI night hours: The main goal of this paper is to forecast the value of GHI. During the night, there is no solar radiation and the Pyranometer (a device used to measure GHI value) recorded zero values at night hours. These hours add no values to the forecasting model and removing them is very necessary. Therefore, all the night hours of the GHI on each day and all the corresponding features associated with it are removed from the dataset.

6) Dividing data into training and testing: After completing the aforementioned five steps, the data are divided into training and testing set in the ratio of 80% for the training process and 20% for the testing process.

The Pareto principle is a common rule of thumb to divide the dataset into two sub-sets; training and testing data. This is also called the 80/20 rule. The training and testing dataset are selected randomly. The objective of selecting the data randomly is to make our model be trained based on a variety of weather observations. These training and testing data are then fixed, and all the forecasting algorithms are encountering the same training and testing data input.

7) Input data normalization: Input data scaling, also known by normalization, is a very critical practical implementation when applying ANNs. The importance of this practical consideration is mainly to avoid the possible domination of attributes with greater numeric values upon those attributes with smaller ones. Another significant feature is to overcome numerical difficulties during computation processes. Because of the dependency of the kernel values on the inner products of the attribute vectors, attributes with large values cause numerical problems. In this research, each attribute is linearly normalized to the range: 0 to 1, using equation (2).

$$x_i^n = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (2)$$

Where x_i is the actual value of the feature vector; x_{min} and x_{max} are the minimum and the maximum values corresponding to the actual dataset; x_i^n is the normalized value associated with x_i.

C. Multilayer Feed-forward Back Propagating Neural Networks

Non-linear relationships between independent and dependent variables can be captured by Artificial Intelligence (AI) methods. One of the powerful non-linear forecasting algorithms used here is an ANN. ANNs mimic how human nervous systems interpret information. Being ANNs are capable of modelling non-linear processes without a need to assume the relationship form between the input and output variables is considered one of the main advantages of this technique. The type of the ANN used here is shown in Fig. 2, it is a Multi-Layer Perception (MLP) [43]. To conduct the training process, the Back-Propagation Algorithm (BP) is selected in this research because it is one of the most common ANN algorithms [44]. As Fig. 2 depicts, the usual architecture of ANNs formed with three main layers. First, the input layer, $[x_1, x_2, \dots, x_N]^T$, which is composed of an N-dimensional input vector. After that, the hidden layer, $[h_1, h_2, \dots, h_M]^T$, which includes a nonlinear activation function known as the activation function. Finally, the output layer, $[y_1, y_2, \dots, y_L]^T$, which contains a linear function. Inside hidden layers, the outputs of nodes, also known as neurons, which represent the basic component of any ANN, can be calculated as below:

$$z_j = \sum_{i=0}^{N} v_{ij} x_i, j = 1,2, \dots, M, i = 1,2, \dots, N \quad (3)$$

$$h_j = f(z_j), j = 1,2, \dots, M \quad (4)$$

In which,

- z_j is the value of the activation function of the jth node associated with the hidden layer.
- v_{ij} is the weight that connects the input i, in the input layer, with the node j in the hidden layer.
- f is known as the transfer function of the neurons, often a sigmoid function is selected $f(x) = \frac{1}{1+\exp(-x)}$.
- h_j is the output value of the node j in the hidden layer.

In the output layer, the values of the output nodes can be calculated by using the equations (5) and (6).

$$z_l = \sum_{i=0}^{M} w_{jl} h_j, l = 1,2, \dots, L, j = 1,2, \dots, M \quad (5)$$

$$y_l = f(z_l), l = 1,2, \dots, L \quad (6)$$

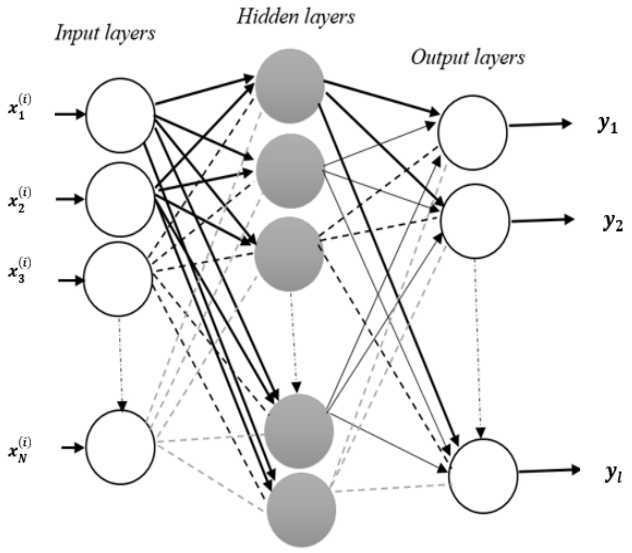

Fig. 2. The MFBP Network Model.

In which,

- z_l is the value of the activation function associated with the lth node in the output layer.
- w_{jl} the weight that connects the node j in the hidden layer with the node l in the output layer.
- f is a sigmoid activation function.
- y_l is the value of the activation function of the node l in the output layer.

Through experimenting with different choices, the required number of hidden layers, and the number of nodes in each layer were selected based on the optimal value that provides the best training prediction performance is reached. In this article, training networks for the model with three different sets of features were utilized by the MLP with the BP algorithm, while the Levenberg-Marquardt approach was the training function. One input layer, one hidden layer, and one output layer are used. The hidden layer in the ANN was constructed with 14, 8, and 5 neurons (nodes) for the model using the different sets of features; All-Feature, Eight-Feature, and Five-Feature. The input and the output of the training and testing dataset are similar for the model with the three sets.

D. Feature Selection Algorithm

In ML applications, feature selection, also known as variable selection, is frequently a crucial phase towards building a highly accurate ML-based model [45]. There are good reasons that support the use of feature selection algorithms. Nowadays, new datasets for practical model designing are usually described with a very high number of variables. Most of these variables are often irrelevant to the classification problem, and their significance is not established beforehand.

In dealing with data with too large feature sets, several disadvantages will appear. Practically, it found that dealing with massively large feature sets causes algorithms to slow down. Another reason is even more critical is the decrease in prediction accuracy is shown in many ML algorithms when dealing with higher than optimal sets of variables. Therefore, it is desirable for practical reasons to select the possibly smallest set of features that returns the best possible prediction results. This problem, known as the minimal-optimal problem, has been considerably researched where plenty of algorithms were developed to come up with manageable-size sets of features. Nevertheless, this very practical objective echoes another very important issue, which is the recognition of all features that are relevant to the classification problem under certain conditions. This is the so-called all-relevant problem. It can be very useful in itself to find all relevant features, rather than just non-redundant ones. This is especially necessary if one is interested in understanding processes related to the topic of interest, rather than simply building a predictive black-box model.

A good discussion on why it is important to find all relevant features is given by [46]. All relevant feature selection problems are more difficult to handle than normal minimal-optimal alternatives. To determine whether the variable is important or unimportant, therefore, we need a powerful criterion to do so. The filtering methods can be used to select relevant features [47].

However, filtering methods are not the optimal choice for feature selection implementation due to the lack of a direct correlation between a particular feature and the decision that this feature is unimportant in combination with other features. Hence, one is limited to wrapper methods of feature selection that are more challenging computationally than filtering alternatives. As a black box, the classifier is used in wrapper methods to return a feature ranking, so any classifier that can return feature ranking can be used. In a short, a classifier used in feature selection problems should be both computationally effective and easy, optimally without any user-set parameters.

To find the all-relevant features in the solar radiation prediction problem, this paper utilizes the so-called R package Boruta [48]. This package is publically available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=Boruta. In this algorithm, a wrapper approach that is built around a random forest classifier is used [49].

In Slavic mythology, Boruta is the God of the forest. Selecting all relevant features, rather than just the non-redundant features. This algorithm is an extended version of the concept that Stoppiglia et al. implemented in [50] to assess relevance by comparing the relevance of the real features with that of the random probes. Although it is used in this paper as a wrapper algorithm, the concept is originally proposed in the context of filtering. In short, a brief overview of the algorithm is giving in the following section.

1) Boruta algorithm: Boruta Algorithm is a wrapper method that is designated based on the random forest regression algorithm executed in [51]. This paper utilized the regression version of the Boruta algorithm and its origin the random forest algorithm. However, to explain the basics of both algorithms, this article sticks to the classification versions of these algorithms.

The classification algorithm of the random forest is considered relatively fast, can normally run without a need for parameter setting, and it delivers a numerical approximation of the feature's importance. It is considered under the category of ensemble methods, which executes classification by acting on multiple unbiased poor classifiers recognized as decision trees. Such trees are freely and independently established upon different samples of bagging extracted from the training dataset. A feature importance metric is gained as the classification accuracy loss induced by the feature values' random permutation between objects. This measure of the importance of a feature is separately calculated for all trees available in the forest. These trees utilize a given feature for the classification task. Afterwards, the accuracy loss's mean and standard deviation are worked out. Dividing a feature's accuracy loss by its standard deviation results in the so-called Z score that can alternatively be used as the measure of the importance of a feature.

Unfortunately, since the distribution of the random forest algorithm is not N (0,1), the Z score cannot be interpreted as a

direct relation to the statistical significance of the feature importance given by the random forest algorithm [52]. In the Boruta algorithm, nonetheless, since the Z score takes into consideration the average accuracy loss fluctuations among trees in the forest, we use it as the measure of the importance of a feature. Because the Z score cannot be directly used to calculate the importance, some external reference is needed to assess whether the significance of any given feature is important, i.e., whether it is perceptible from the significance of random variations. To that point, the information system has been expanded with random design features. We create a corresponding 'shadow' feature for each original feature, whose values are acquired by mixing values across objects from the original feature. We then use all the features of this extended information system to perform regression and measure the value of all features. Because of random fluctuations, the value of a shadow feature can be nonzero. Thus, the shadow features set of importance is utilized as a guide to decide, which features are considered significant. The significance indicator differs due to the random forest classifier stochasticity. Moreover, that is very sensitive to non-important features being present in the information system (as well as shadow features). It also depends on the specific realization of shadow features. Thus, to obtain statistically valid results, we need to perform the process of re-shuffling.

In short, the Boruta method is based on the same principle that serves as the foundations of the classifier of the random forest algorithm, that is by introducing randomness to the information system and gathering results from the randomized sample ensemble one can the misleading effect of random fluctuations and correlations. Thus, this added randomness will give us a clearer picture of which attributes matter significantly. The steps in which the Boruta Algorithm is executed consist of the following:

- Add copies of all features (variables/predictors) to expand the information system. Even if the number of features in the original dataset is smaller than 5, always extend the information system by at least 5 shadow features.
- To eliminate their correlation with the target variable, shuffle the added features.
- To collect the measured Z scores, run a classifier of the random forest upon the extended information system.
- Figure out the maximum Z score amongst shadow features (MZSF), and after that give a hit to any better-scored feature than MZSF.
- A two-sided equality test with the MZSF is conducted for each feature of undetermined significance.
- Consider the features of significantly lower value than MZSF as ' unimportant' and delete them permanently from the information system.
- Consider the features of significantly higher value than MZSF as ' important'.
- Delete all shadow features.
- Repeat the process until the importance for all the features has been allocated, or the algorithm has exceeded the random forest runs previously set.

E. The Hybrid Strategy Proposed for Forecasting

In this section, the designed hybrid big data-driven strategy for short-term global solar radiation forecasting based on the ANN and Boruta Algorithm, which is applied to select the optimal set of features to be inputted to the ANN algorithm, is introduced. The whole structure of the proposed hybrid system as a GHI forecasting model, as shown in Fig. 3.

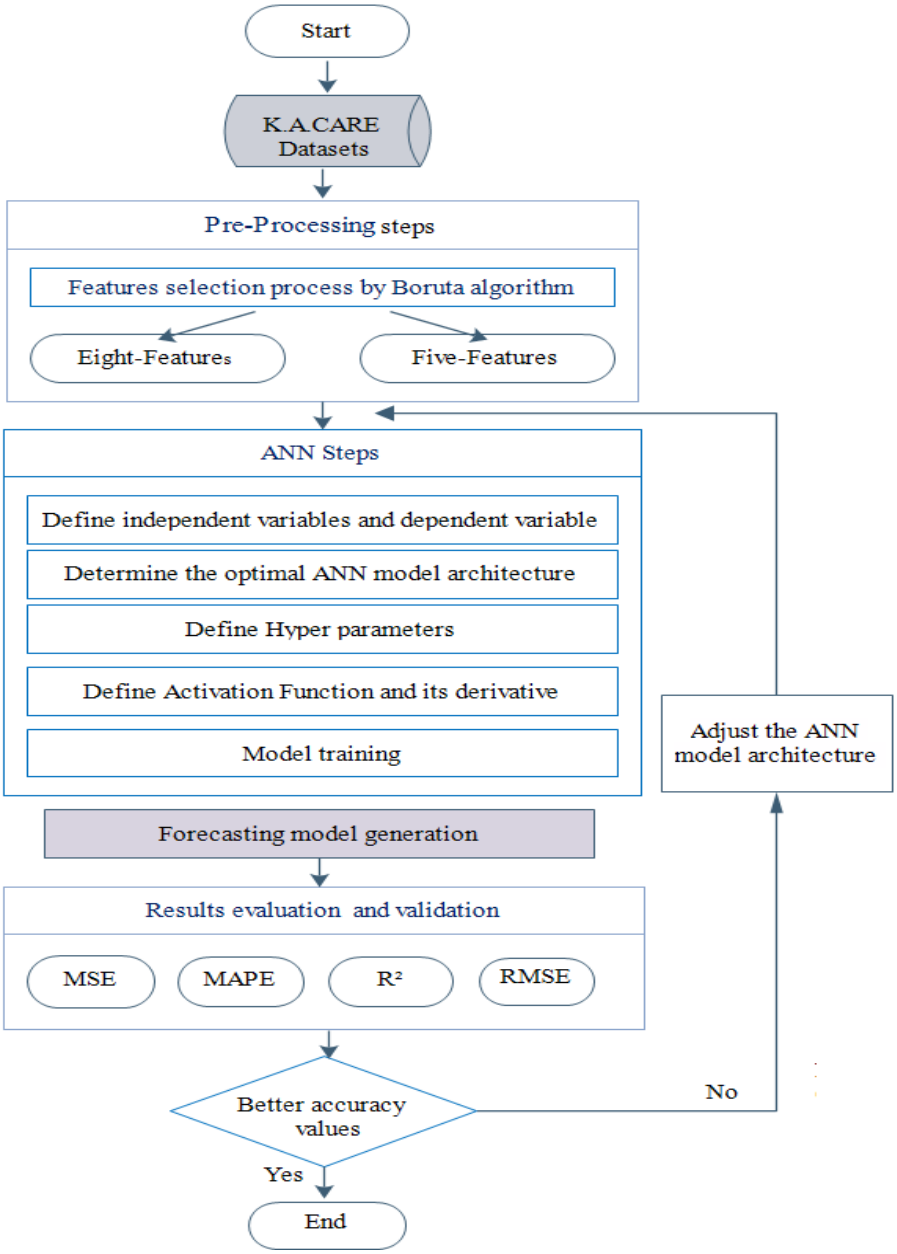

Fig. 3. The Proposed GHI Forecasting Model.

IV. Evaluation Measures

Several statistical measures used to evaluate the prediction performance accuracy of the developed model. This research mainly considers three indicators, namely: mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), and goodness of fit (R^2). Such measures are mathematically represented by the equations (7) to (10).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|f_i - y_i|}{y_i} \times 100\% \quad (7)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(f_i - y_i)^2 \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2} \quad (9)$$

$$R^2 = \frac{\sum_{i=1}^{N}(f_i - \tilde{y})^2}{\sum_{i=1}^{N}(y_i - \tilde{y})^2} \quad (10)$$

Where N represents the number of data points involved in the analysis; y_i is the observed value of the target, f_i is the predicted value of the target; $\tilde{y}$ is the mean of the observed value of the target y_i. While MAPE is utilized to evaluate the model performance accuracy as a percentage, RMSE measures how the observed values deviate from the corresponding predicted values [13]. In regression problems, the R^2 of a model describes how well the model fits a set of observations. R^2 ranges from zero to the preferable number 1.

V. Results and Discussion

The forecasted model is employed to predict GHI at two selected sites in Saudi Arabia, namely Qassim and Jeddah. In this research, the prediction module utilized based on

- All-Features;
- Eight-Feature; and
- Five-Feature.

In this research, the model was implemented using MATLAB. The data was first cleaned and after that normalized to increase the performance of the forecasting and feature selection algorithms.

A set of eight and five features out of the available 13 features were identified to create the so-called Eight-Feature and Five-Feature forecasting model. The choice of these features was based on the output of the feature selection algorithm, Boruta. Eight and five features were selected to demonstrate the performance of the forecasting model with a different number of features. Fig. 4 and Fig. 5 rank the features based on their importance to our outcome (GHI) at Qassim and Jeddah, respectively.

Fig. 4 and Fig. 5 show that in the Qassim region, the Eight-Feature model is built based on the following features: DNI, Zenith Angle, DHI, Hour, Month, Pressure, and Azimuth Angle. On the other hand, Jeddah's Eight-Feature forecasting model is based on the following features: DNI, Zenith angle, DHI, pressure, hour, month, Azimuth angle, and temperature. For the Five-Feature model, the Qassim region forecasting model is based on the following features: DNI, Zenith Angle, DHI, Hour, and Month, while Jeddah's Five-Feature forecasting model is created by using the following features: DNI, Zenith angle, DHI, pressure, and hour. The model developed was utilizing the MLP with a BP, while the Levenberg-Marquardt approach was the training function. One input layer, one hidden layer, and one output layer are used. The hidden layer in the ANN was constructed with 14, 8, and 5 neurons (nodes) for All-Feature, Eight-Feature, and Five-Feature model.

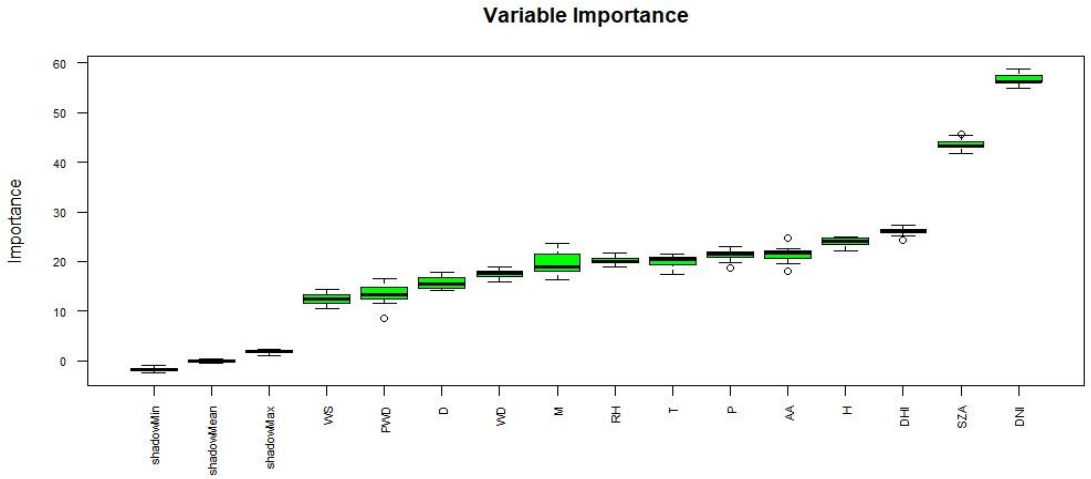

Fig. 4. Variable Importance for Weather Data for the Qassim Region using the Boruta Algorithm.

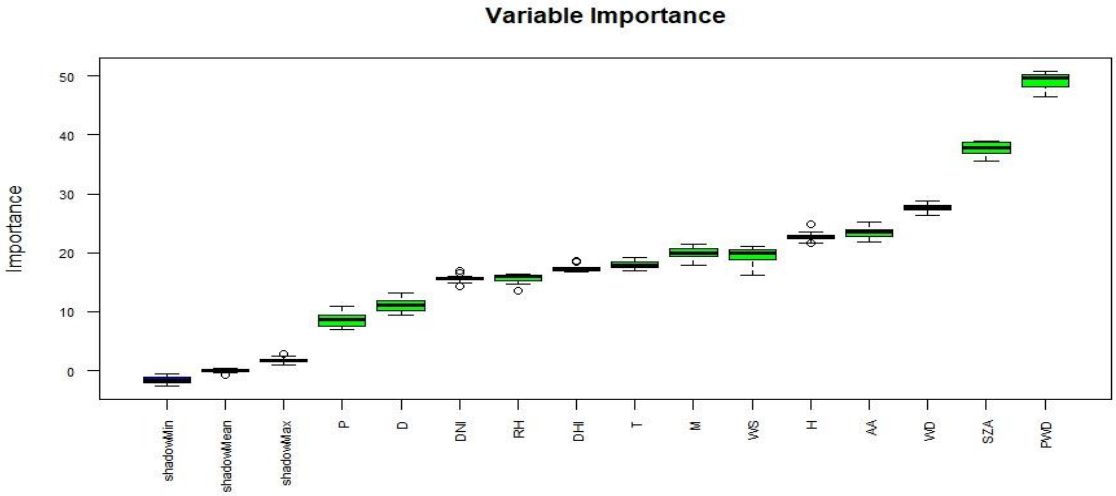

Fig. 5. Variable Importance for Weather Data for the Jeddah Region using the Boruta Algorithm.

The comparisons were conducted between the results of the All-Feature model with the results based on both the Eight-Feature and Five-Feature model in terms of their forecasting accuracy at Qassim and Jeddah sites. For this goal, the accuracy outputs of the testing model were tested based on the following metrics: MAPE, MSE, RMSE, and R^2. Table II and Table III compare the results of the model for the three different sets of features at Qassim and Jeddah.

MAPE determines the accuracy and errors ratio between measured and predicted data, while MSE and RMSE measure the relative error and expressed in ($Watt/m^2$). Table II and Table III contain the hourly forecasting of GHI based on the study of the model at Qassim and Jeddah. In Qassim and from Table II, with the All-Feature model MAPE value found to be 13.496% (16.532% with Eight-Feature and 26.563% with Five-Feature).

The MSE and RMSE values of All-Feature model are 1708.957 $Watt/m^2$ (1756.145 $Watt/m^2$ with Eight-Feature and 2360.716 $Watt/m^2$ with Five-Feature) and 41.339 $Watt/m^2$ (41.906 $Watt/m^2$ with Eight-Feature and 48.587 $Watt/m^2$ with Five-Feature), respectively. The correlation scores of the forecasting model at Qassim was found to be 0.99124, 0.99167, and 0.9794 for All-Feature, Eight-Feature, and Five-Feature, respectively.

The results indicate that the proposed All-Feature model has the best performance compared to the eight and Five-Feature model. However, the Eight-Feature model performance is high in a way that can be compared with the All-Feature model, while Five-Feature can be considered the poorest model, still, it has satisfactory results. According to Table III that presents Jeddah results, the Five-Feature model can be considered as the best model followed by Eight-Feature and All-Feature model, respectively. Unlike Qassim site, Jeddah forecasting model with merely five and eight features prove the significance of using a feature selecting approach and how adding more feature may lead to overfitting forecasting model. In Jeddah, the MAPE value was determined to be 13.7013% with All-Feature (12.2024% with Eight-Feature and 9.6936% with Five-Feature).

TABLE II. RESULTS OF HOURLY FORECASTING GHI AT QASSIM

	MAPE %	MSE $Watt/m^2$	RMSE $Watt/m^2$	R^2
All Feature	13.50	1708.96	41.34	0.99
Eight Feature	16.50	1756.15	41.91	0.99
Five Feature	26.60	2360.72	48.59	0.98

TABLE III. RESULTS OF HOURLY FORECASTING GHI AT JEDDAH

	MAPE %	MSE $Watt/m^2$	RMSE $Watt/m^2$	R^2
All Feature	13.70	994.37	31.54	0.99
Eight Feature	12.20	939.87	30.66	0.99
Five Feature	9.70	913.84	30.23	0.99

The MSE and RMSE values of All-Feature model are 994.369 $Watt/m^2$ (939.869 $Watt/m^2$ with Eight-Feature and 913.84 $Watt/m^2$ with Five-Feature) and 31.533 $Watt/m^2$ (30.6572 $Watt/m^2$ with Eight-Feature and 48.587 $Watt/m^2$ with Five-Feature), respectively.

The correlation scores of the forecasting model at Qassim were found to be 0.99124, 0.99168, and 0.99184 for All-Feature, Eight-Feature, and Five-Feature, respectively. For further visualization, the measured GHI values are plotted against the output of the forecasting model by three different sets of Features at Qassim as in Fig. 6 and Jeddah as in Fig. 7. Where Fig. 6 confirms that the All-Feature model has high accuracy results at Qassim, and Fig. 9 shows how the Five-Feature model performs the better compare to All-Feature and Eight-Feature model in Jeddah. The model results of All-Feature, Eight-Feature, and Five-Feature are plotted all together with measures GHI values in Fig. 8 at Qassim and Fig. 9 at Jeddah for twenty random hours.

Fig. 8 shows that the All-Feature model has superior performance in tracking the original GHI value at Qassim, and Fig. 9 confirms the ability of the Five-Feature model in following the measured GHI values at Jeddah. In regards to the cost reduction by the generated model, Table IV illustrates the prices of devices and equipment required for the prediction purpose of the GHI. This table also shows the cost reduction by the forecasting model using eight and five features for the two regions.

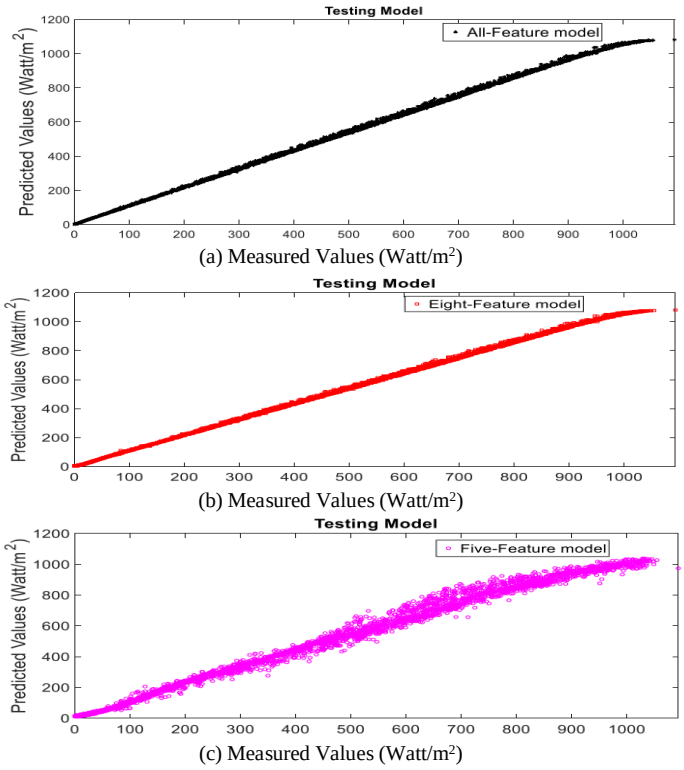

Fig. 6. Measured vs. Predicted Values of GHI for Qassim Region, (a) All-Feature Model, (b) Eight-Feature Model, (c) Five-Feature Model.

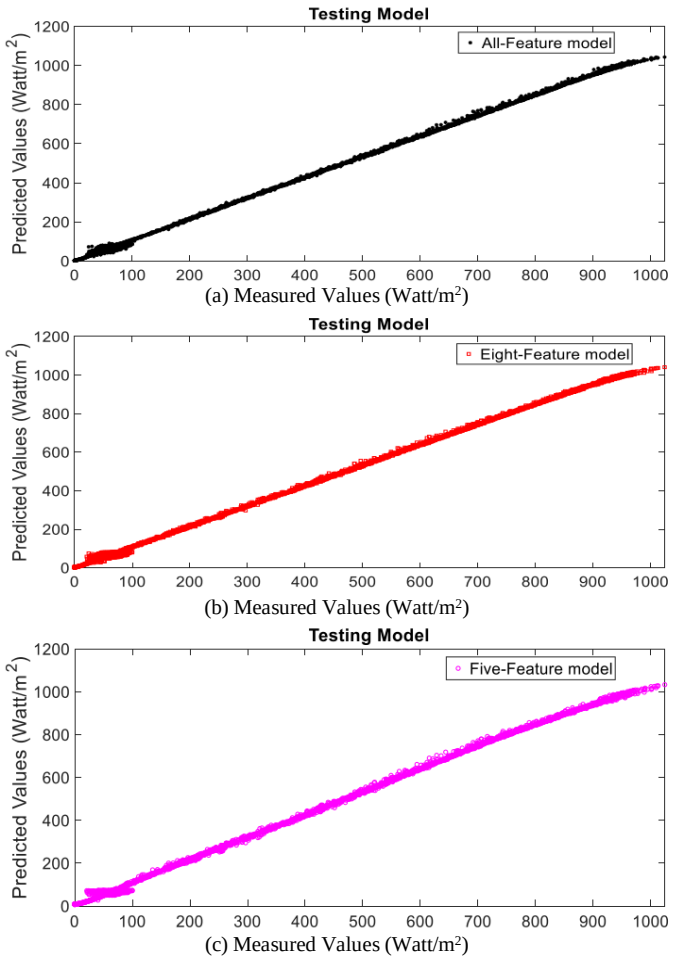

Fig. 7. Measured vs. Predicted Values of GHI for the Jeddah Region, (a) All-Feature Model, (b) Eight-Feature Model, (c) Five-Feature Model.

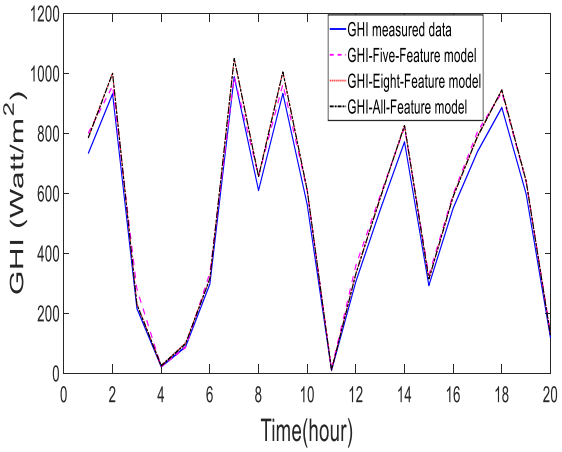

Fig. 8. The Forecasted GHI Values vs Measured GHI Values at Qassim.

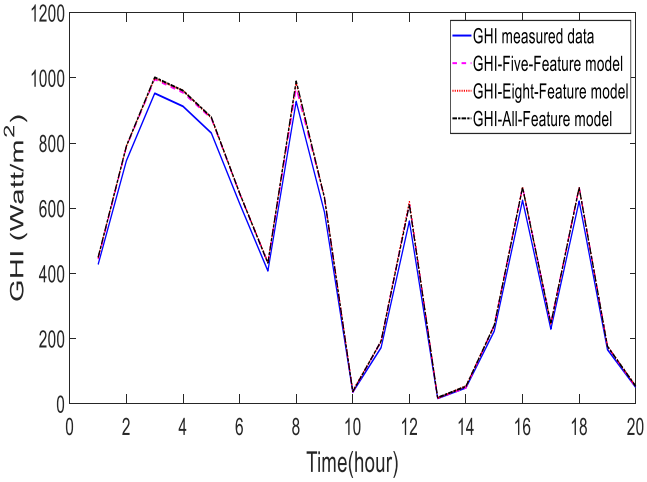

Fig. 9. The Forecasted GHI Values vs. Measured GHI Values at Jeddah.

The feature selection algorithm helped to decrease the cost of solar monitoring stations by reducing the number of features. For example, the less reduction in Jeddah for the Five-Features model, the cost decrease from 113990 RS to 60026 RS. The reduction percentage is 47%. As well in Jeddah using 8 features, the cost reduced from 113990 RS to 62065 RS, where the reduction percentage was equal to 46%. The rate of reduced cost is big although, it does not include the costs of maintenance, cables, and other accessories prices, On the other hand, in Qassim region, the cost reduced when using five features from 113990 to 76886 with a percentage =33%. As well when using 8 features in this region, the cost reduced from 113990 RS to 109333 RS, with a reduction percentage equal only 4%, as shown in Table IV.

At the site of Jeddah, and according to the findings of the feature selection algorithm, using the five-feature model resulted in the best prediction performance of the GHI compared to the prediction values of the model used a larger number of eight or all features. Also, the rate of cost reduction in Jeddahwas very high, as presented in Table IV, where the cost reduction values for All, eight, and five attributes for that model were calculated.

On the other hand, in the Qassim area, the best prediction values were obtained by using the model with all-feature although the model with few attributes gives good results, where the value R^2 gives approximately the same values for All-Features. This means the eight and five features also satisfying good results can apply in the future with a lower cost than the cost of all features, although the total costs reduced were small value it remains positive. Consequently, the feature selection algorithm helps to decrease the cost of solar monitoring stations. The reduced costs do not include specialists, maintenance, cables, and accessories prices. This, in turn, gives special importance to the research finding.

TABLE IV. COST REDUCTION FOR THE GENERATED MODEL IN THE TWO REGIONS IN SAUDI RIYALS (RS)

I	Features	Unit price	All Features	Qassim Region		Jeddah Region	
				Eight Features	Five Features	Eight Features	Five Features
1	Month Of The Year(M)	-	√	Eight	Five	√	-
2	Day Of The Month(D)	-	√	-	-	-	-
3	Hour Of The Day (H)	-	√	√	√	√	√
4	Air Temperature (T)	730	√	√	√	√	-
5	Relative Humidity (RH)	730	√	√	-	-	-
6	Surface Pressure (P)	3013	√	-	-	-	-
7	Wind Speed At 3 Meters (WS)	1309	√	√	-	√	-
8	Wind Direction (WD)	1309	√	-	-	√	√
9	Peak Wind Direction At 3 Meters (PWD)	1309	√	-	-	√	√
10	Diffuse Horizontal Irradiance(DHI)	35037	√	-	-	-	-
11	Direct Normal Irradiance (DNI)	13145	√	√	√	-	-
12	Azimuth Angle (AA)	28704	√	√	√	√	√
13	Solar Zenith Angle (SZA)	28704	√	√	√	√	√
Total		-	113990	109333	76886	62065	60026
Cost Reduction Rate		-	-	4%	33%	46%	47%

VI. CONCLUSIONS

The use of an advanced embedded feature selection algorithm and ANN is addressed in this paper to forecast the hourly solar radiation at two sites in the Kingdom of Saudi Arabia. The data from two stations; Qassim and Jeddah in Saudi Arabia, it was obtained to examine the prediction performance of the developed model. The five and eight most important variables among a wide range of metrological variables that could impact solar radiation in the future were optimally and systematically determined by employing a recent feature selection technique named as Boruta algorithm. For the comparison reasons of the model results with different features. The all-feature model was used to assess the benefits of using a feature selection method. The 13 input variables are the maximum number of features considered for developing a data-driven forecasting model.

At the site of Jeddah, and according to the findings of the feature selection algorithm, using the five-feature model resulted in the best prediction performance compared to the prediction values of the model when used a larger number of eight or all features. Also, the rate of cost reduction in Jeddahwas was very high. In the Qassim area, the best prediction values were obtained by using the all-feature model although the model of other features with few attributes is good where the value R^2 gives approximately the same values for All-Features. This means the 8 and 5 features also satisfying very good results can apply in the future with a lower cost than the cost of all features, but, it noted that the total costs reduced were small value. Using feature selection methods may successfully exploit the larger interdependent variables relevant to hourly global horizontal irradiance prediction without sacrificing predictive efficiency. The findings, therefore, emphasize the importance of using feature selection techniques when using the model for computational intelligence to achieve accurate predictions of solar radiation. On the other hand, the cost rate of the GHI prediction was reduced for the generated model, as presented in Table IV above, where the cost reduction values of the model using all, eight and five attributes were calculated. Consequently, From the discussed results, it found that the feature selection algorithm helps to decrease the cost of instruments and equipment required for solar monitoring stations for a high rate. Although, the costs that were reduced do not include the cost of specialists, maintenance, cables, and accessories prices. This, in turn, gives strength and special importance to the research finding. Besides, the lower-cost models can be used in future to collect new data for the coming years for forecasting.

VII. FUTURE WORK

The research results are leading all researchers who have the same interest to achieve important extensions in the future to the current finding, it can include the following:

1) Expanding the samples of the study, using other ML tools, and going deeper into the data analysis based on other selection features methods.

2) Studying more locations across Saudi Arabia to address the geographical effects.

3) Investigating more ML algorithms and comparing prediction performances.

4) Considering different types of feature selection methods.

5) Going more into the dataset analysis deeply to find other important insights that can also help in KSA community services in the future.

6) Considering different test locations with different climate conditions, to investigate their effects on the performance of the prediction model enhanced by feature selection techniques can form another research potential in the scope of the solar prediction.

ACKNOWLEDGMENT

The authors would like to express their great thanks to King Abdullah City for Atomic and Renewable Energy for providing the required datasets for this research.

REFERENCES

[1] D. Lew et al., "Sub-Hourly Impacts of High Solar Penetrations in the Western United States," 2012.

[2] M. Sandhu and T. Thakur, "Issues, Challenges, Causes, Impacts and Utilization of Renewable Energy Sources - Grid Integration," J. Eng. Res. Appl., vol. 4, no. 3, pp. 636–643, 2014, [Online]. Available: http://www.ijera.com/papers/Vol4_issue3/Version 1/DH4301636643.pdf.

[3] B. . Hernandez, "The Religiosity and spirituality scale for youth : The Developmeny and initial validation," Anesthesiology, vol. 115, no. 3, p. A13, 2011, doi: 10.1097/ALN.0b013e3182318466.

[4] A. Tuohy et al., "Solar Forecasting: Methods, Challenges, and Performance," IEEE Power Energy Mag., vol. 13, no. 6, pp. 50–59, 2015, doi: 10.1109/MPE.2015.2461351.

[5] M. Hossain, S. Mekhilef, M. Danesh, L. Olatomiwa, and S. Shamshirband, "Application of extreme learning machine for short term output power forecasting of three grid-connected PV systems," J. Clean. Prod., vol. 167, pp. 395–405, 2017, doi: 10.1016/j.jclepro.2017.08.081.

[6] L. Martín, L. F. Zarzalejo, J. Polo, A. Navarro, R. Marchante, and M. Cony, "Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning," Sol. Energy, vol. 84, no. 10, pp. 1772–1781, 2010, doi: 10.1016/j.solener.2010.07.002.

[7] A. Alzahrani, J. W. Kimball, and C. Dagli, "Predicting solar irradiance using time series neural networks," in Procedia Computer Science, 2014, vol. 36, pp. 623–628, doi: 10.1016/j.procs.2014.09.065.

[8] A. Sharma and A. Kakkar, "Forecasting daily global solar irradiance generation using machine learning," Renewable and Sustainable Energy Reviews, Elsevier Ltd, vol. 82, no. 5, pp. 2254–2269, 2018, doi: 10.1016/j.rser.2017.08.066.

[9] I. A. Ibrahim and T. Khatib, "A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm," Energy Convers. Manag., vol. 138, pp. 413–425, 2017, doi: 10.1016/j.enconman.2017.02.006.

[10] S. Ghimire, R. C. Deo, N. J. Downs, and N. Raj, "Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities," Remote Sens. Environ., vol. 212, pp. 176–198, 2018, doi: 10.1016/j.rse.2018.05.003.

[11] R. Kumar, R. K. Aggarwal, and J. D. Sharma, "Comparison of regression and artificial neural network models for estimation of global solar radiations," Renewable and Sustainable Energy Reviews, Elsevier Ltd, vol. 52, pp. 1294–1299, 2015, doi: 10.1016/j.rser.2015.08.021.

[12] W.-Y. Chang, "Short-Term Load Forecasting Using Radial Basis Function Neural Network," J. Comput. Commun., vol. 03, no. 11, pp. 40–45, 2015, doi: 10.4236/jcc.2015.311007.

[13] C. Renno, F. Petito, and A. Gatto, "Artificial neural network models for predicting the solar radiation as input of a concentrating photovoltaic system," Energy Convers. Manag., vol. 106, pp. 999–1012, 2015, doi: 10.1016/j.enconman.2015.10.033.

[14] M. H. Alobaidi, P. R. Marpu, T. B. M. J. Ouarda, and H. Ghedira, "Mapping of the solar irradiance in the UAE using advanced artificial neural network ensemble," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 7, no. 8, pp. 3668–3680, 2014, doi: 10.1109/JSTARS.2014.2331255.

[15] A. Mellit, A. H. Arab, N. Khorissi, and H. Salhi, "An ANFIS-based forecasting for solar radiation data from sunshine duration and ambient temperature," 2007, doi: 10.1109/PES.2007.386131.

[16] J. C. Lam, K. K. W. Wan, and L. Yang, "Solar radiation modelling using ANNs for different climates in China," Energy Convers. Manag., vol. 49, no. 5, pp. 1080–1090, 2008, doi: 10.1016/j.enconman.2007.09.021.

[17] C. Voyant, M. Muselli, C. Paoli, and M. L. Nivet, "Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation," Energy, vol. 39, no. 1, pp. 341–355, 2012, doi: 10.1016/j.energy.2012.01.006.

[18] P. a Guti, "Hybrid Artificial Neural Networks : Models ," vol. 9, no. 2, pp. 177–184, 2011.

[19] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petković, and C. Sudheer, "A support vector machine-firefly algorithm-based model for global solar radiation prediction," Sol. Energy, vol. 115, pp. 632–644, 2015, doi: 10.1016/j.solener.2015.03.015.

[20] R. C. Deo and M. Şahin, "An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland," Environ. Monit. Assess., vol. 188, no. 2, pp. 1–24, Feb. 2016, doi: 10.1007/s10661-016-5094-9.

[21] M. Mohandes, "Support vector machines for short-term electrical load forecasting," Int. J. Energy Res., vol. 26, no. 4, pp. 335–345, 2002, doi: 10.1002/er.787.

[22] L. M. Saini, S. K. Aggarwal, and A. Kumar, "Parameter optimisation using genetic algorithm for support vector machine-based price-forecasting model in National electricity market," IET Gener. Transm. Distrib., vol. 4, no. 1, pp. 36–49, 2010, doi: 10.1049/iet-gtd.2008.0584.

[23] A. Hepbasli and Z. Alsuhaibani, "A key review on present status and future directions of solar energy studies and applications in Saudi Arabia," Renewable and Sustainable Energy Reviews, vol. 15, no. 9. pp. 5021–5050, 2011, doi: 10.1016/j.rser.2011.07.052.

[24] A. A. El-Sebaii, A. A. Al-Ghamdi, F. S. Al-Hazmi, and A. S. Faidah, "Estimation of global solar radiation on horizontal surfaces in Jeddah, Saudi Arabia," Energy Policy, vol. 37, no. 9, pp. 3645–3649, 2009, doi: 10.1016/j.enpol.2009.04.038.

[25] J. C. Lam, K. K. W. Wan, and L. Yang, "Solar radiation modelling using ANNs for different climates in China," Energy Convers. Manag., vol. 49, no. 5, pp. 1080–1090, 2008, doi: 10.1016/j.enconman.2007.09.021.

[26] H. Bulut and O. Büyükalaca, "Simple model for the generation of daily global solar-radiation data in Turkey," Appl. Energy, vol. 84, no. 5, pp. 477–491, 2007, doi: 10.1016/j.apenergy.2006.10.003.

[27] H. A. Kazem, "Solar Radiation, Temperature and Humidity Measurements in Sohar-Oman," Int. J. Comput. Appl. Sci., vol. 1, no. 1, pp. 15–20, 2016, doi: 10.24842/1611/0003.

[28] S. Rehman and S. G. Ghori, "Spatial estimation of global solar radiation using geostatistics," Renew. Energy, vol. 21, no. 3–4, pp. 583–605, 2000, doi: 10.1016/S0960-1481(00)00078-1.

[29] A. A. El-Sebaii, A. A. Al-Ghamdi, F. S. Al-Hazmi, and A. S. Faidah, "Estimation of global solar radiation on horizontal surfaces in Jeddah, Saudi Arabia," Energy Policy, vol. 37, no. 9, pp. 3645–3649, 2009, doi: 10.1016/j.enpol.2009.04.038.

[30] M. Mohandes, S. Rehman, and T. O. Halawani, "Estimation of global solar radiation using artificial neural networks," Renew. Energy, vol. 14, no. 1–4, pp. 179–184, 1998, doi: 10.1016/S0960-1481(98)00065-2.

[31] M. Benghanem, A. Mellit, and S. N. Alamri, "ANN-based modelling and estimation of daily global solar radiation data: A case study," Energy Convers. Manag., vol. 50, no. 7, pp. 1644–1655, Jul. 2009, doi: 10.1016/j.enconman.2009.03.035.

[32] M. Almaraashi, "Short-term prediction of solar energy in Saudi Arabia using automated-design fuzzy logic systems," PLoS One, vol. 12, no. 8, pp. 1-16, doi: 10.1371/journal.pone.0182429.

[33] B. Ameen, H. Balzter, C. Jarvis, and J. Wheeler, "Modelling hourly global horizontal irradiance from satellite-derived datasets and climate variables as new inputs with artificial neural networks," Energies, vol. 12, no. 1, 2019, doi: 10.3390/en12010148.

[34] P. Mpfumali, C. Sigauke, A. Bere, and S. Mulaudzi, "Day ahead hourly global horizontal irradiance forecasting—application to South African data," Energies, vol. 12, no. 18, pp. 1–28, 2019, doi: 10.3390/en12183569.

[35] T. Betti, I. Zulim, S. Brkić, and B. Tuka, "A Comparison of Models for Estimating Solar Radiation from Sunshine Duration in Croatia," Int. J. Photoenergy, vol. 2020, 2020, doi: 10.1155/2020/9605950.

[36] V. Kallio-Myers, A. Riihelä, P. Lahtinen, and A. Lindfors, "Global horizontal irradiance forecast for Finland based on geostationary weather satellite data," Sol. Energy, vol. 198, no. 1, pp. 68–80, 2020, doi: 10.1016/j.solener.2020.01.008.

[37] Y. H. Koo, M. Oh, S. M. Kim, and H. D. Park, "Estimation and mapping of solar irradiance for Korea by using COMS MI satellite images and an artificial neural network model," Energies, vol. 13, no. 2, 2020, doi: 10.3390/en13020301.

[38] J. Fan, X. Wang, F. Zhang, X. Ma, and L. Wu, "Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data," J. Clean. Prod., vol. 248, p. 1-14, 2020, doi: 10.1016/j.jclepro.2019.119264.

[39] S. Salcedo-Sanz, S. Jiménez-Fernández, A. Aybar-Ruiz, C. Casanova-Mateo, J. Sanz-Justo, and R. García-Herrera, "A CRO-species optimization scheme for robust global solar radiation statistical downscaling," Renew. Energy, vol. 111, pp. 63–76, 2017, doi: 10.1016/j.renene.2017.03.079.

[40] A. K. Yadav, H. Malik, and 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," Renewable and Sustainable Energy Reviews, Elsevier Ltd, vol. 52. pp. 1093–1106, Aug. 2015, doi: 10.1016/j.rser.2015.07.156.

[41] A. R. Hedar, A. E. Abdel-Hakim, and M. Almaraashi, "Granular-based dimension reduction for solar radiation prediction using adaptive memory programming," in GECCO 2016 Companion - Proceedings of the 2016 Genetic and Evolutionary Computation Conference, Jul. 2016, pp. 929–936, doi: 10.1145/2908961.2931648.

[42] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Data cleaning for classification using misclassification analysis," J. Adv. Comput. Intell. Intell. Informatics, vol. 14, no. 3, pp. 297–302, 2010, doi: 10.20965/jaciii.2010.p0297.

[43] M. Mohandes, "Support vector machines for short-term electrical load forecasting," Int. J. Energy Res., vol. 26, no. 4, pp. 335–345, 2002, doi: 10.1002/er.787.

[44] S. Leva, A. Dolara, F. Grimaccia, M. Mussetta, and E. Ogliari, "Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power," Math. Comput. Simul., vol. 131, pp. 88–100, 2017, doi: 10.1016/j.matcom.2015.05.010.

[45] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/s0004-3702(97)00043-x.

[46] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time," J. Mach. Learn. Res., vol. 8, pp. 589–612, 2007.

[47] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3. pp. 1157–1182, 2003, doi: 10.1162/153244303322753616.

[48] R. C. Team, "The R Project for Statistical Computing," Http://Www.R-Project.Org/, pp. 1–12, 2013, [Online]. Available: https://www.r-project.org/.

[49] A. Ng and K. Soo, "0.1 Die Weisheit der Crowd," Data Sci. ist das Eig., vol. 45, pp. 5–32, 2018, doi: 10.1007/978-3-662-56776-0_10.

[50] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1399–1414, 2003.

[51] J. Hebebrand, "Editorial: Contents of this issue," Obesity Facts, vol. 3, no. 6. pp. 343–344, 2010, doi: 10.1159/000323281.

[52] M. INUIGUCHI, "Rough Sets and Current Trends in Computing 2004," Syst. Control Inf., vol. 48, no. 11, p. 473, 2004, doi: 10.11509/isciesci.48.11_473.