

Received April 5, 2022, accepted April 22, 2022, date of publication April 29, 2022, date of current version May 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3171230

Optimized Feature Selection Based on a Least-Redundant and Highest-Relevant Framework for a Solar Irradiance Forecasting Model

NAJIYA OMAR^{ID}, (Graduate Student Member, IEEE),
HAMED ALY^{ID}, (Senior Member, IEEE), AND **TIMOTHY LITTLE**

Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS B3H 4R2, Canada

Corresponding author: Najiya Omar (najiya.omar@dal.ca)

ABSTRACT Exogenous and endogenous variables are typically evaluated several times during the selection trial of a predictive model for Global Horizontal Irradiance (GHI). This is accomplished using various statistical measures (e.g., univariate statistical analysis, correlation analysis, etc.) that are applied to gauge redundancy and relevancy in specific variables. The main benefits of these approaches include lower computational cost, fast screening times, accurate measuring of linear and monotonic degrees of variable pairs, and the removal of features with low relevance. However, they cannot identify instances where single or groups of predictor variables are non-monotonically associated with the response variable, nor can they discern whether variables are predictive in combination with other variables or in isolation. The present study attempts to overcome these challenges by first describing monotonic and non-monotonic (Spearman's rho and Hoeffding's D, respectively) correlation statistics in combined usage for locating groups with major non-monotonic endogenous variable changes. The proposed work's novelty is subset evaluation that determines relevance using Weather Recursive Feature Elimination (WRFE). This is a novel hybrid feature reduction method that optimizes feature selection using a Least-Redundant/Highest-Relevant framework. The proposed WRFE utilizes feature importance for measuring variance reduction in Random Forest Regression (RFR) and as data perturbation in Long Short-Term Memory (LSTM). The simulation results of GHI hourly predictions demonstrate that the proposed optimal features of the training subset make the greatest contributions to the prediction target, proving that the high variability of irradiance conditions lowers training subset reliability. The results showed that the proposed WRFE is superior compared to the other models with 1.0927 % for the RMSE and the R^2 coefficient is exceeding 98%.

INDEX TERMS Feature importance, redundancy and relevancy measures, exogenous and endogenous variables, recursive feature elimination, long short-term memory, random forest regression, GHI forecasting, data perturbation, variance reduction.

NOMENCLATURE

RH_{mean} Average relative humidity.
 μ Mean.
 ρ_P Pearson correlation.
 ρ_S Spearman correlation.
 σ Standard deviation.
ANFIS Adaptive network fuzzy inference system.

ARD Automatic Relevance Determination.
BDT Boosted decision tree.
CGHI Clearsky GHI.
DHI Diffuse Horizontal Irradiance.
DNI Direct Normal Irradiance.
DOY Day of year.
DP Dew Point.
GFM Generalized Fuzzy model.
GHI Global Horizontal Irradiance.
H₀ Extraterrestrial global solar radiation.

The associate editor coordinating the review of this manuscript and approving it for publication was Dipankar Deb^{ID}.

<i>HMM</i>	Hidden Markov Model.
K_+^{ref}	Daily clearness index for the reference station.
<i>LR</i>	Linear regression.
<i>LSTM</i>	Long short-term memory.
<i>MBE</i>	Mean Bias Error.
<i>MLR</i>	Multiple linear regression.
<i>MTSF</i>	Multivariate time-series forecasting.
<i>NGA</i>	Niching Genetic Algorithms.
<i>NMIFS</i>	Normalized mutual information.
<i>P</i>	Pressure.
<i>p_C</i>	Population correlation coefficient.
<i>PW</i>	Precipitable water.
<i>R_a</i>	Extraterrestrial radiation.
<i>RBFNN</i>	Radial Basis Function neural network.
<i>RFR</i>	Random Forest Regression.
<i>RH</i>	Relative Humidity.
<i>RMSE</i>	Root Mean Square Error.
<i>S₀</i>	Sunshine duration.
<i>SA</i>	Surface Albedo.
<i>SZA</i>	Solar Zenith Angle.
<i>T</i>	Temperature.
<i>T_{max}</i>	Max air temperature.
<i>T_{min}</i>	Min air temperature.
<i>T_{mean}</i>	Average air temperature.
<i>T_{mean}</i>	Mean air temperature.
<i>UTSF</i>	Univariate time-series forecasting.
<i>VIF</i>	Variance Inflation Factors.
<i>WD</i>	Wind Direction.
<i>WRFE</i>	Weather Recursive Feature Elimination.
<i>WS</i>	Wind Speed.
<i>f(d)</i>	Probability function of Gaussian noise.
<i>M</i>	Calendar month number.

I. INTRODUCTION

As a means to better understand relationships between model components, forecasting models typically use numerous input variables. Analysis, however, can be compromised by the high dimensionality of these variables. In particular, redundant inputs can cause a variety of issues, such as increasing the computational time, heightening the chance of under/overfitting, destabilizing estimates on parameters, and preventing accurate detection of relationships pertaining to the explanatory and response variables. Unlike relevancy, redundancy does not include the response variable, whereas relevancy involves the relationship between the target and predictors. Therefore, during the process of feature selection, it is imperative to choose features relevant to the prediction, while simultaneously ensuring that there is no redundancy in them. In an earlier study [1], the behavior of Long Short-Term Memory (LSTM) models was investigated under certain geographical and meteorological conditions and according to previous data on solar irradiance. These types of variables (i.e., exogenous and endogenous, respectively) were utilized as input features for day-ahead solar irradiance forecasting

models. In the work, a comparison was made between LSTM model results and Radial Basis Function Neural Network (RBFNN) outcomes with regard to univariate time-series forecasting (UTSF) and multivariate time-series forecasting (MTSF). The results were then validated using data obtained from a region with different climatic conditions. Interestingly, the LSTM performance deteriorated when additional features were added. Next, a series of questions are presented that we will attempt to answer in the current study:

- What types of associations occur between the input features of exogenous and endogenous variables that could be considered to enhance the overall prediction results?

Most of the correlation analysis in the literature is performed to determine whether a linear relationship exists among the input features [2]–[5]. In this study, a feature selection technique based on correlation analysis for redundancy and relevancy measures will be proposed as the basis for making decisions about redundant and/or irrelevant attributes. Redundant attributes are usually measured using Pearson's correlation coefficient to find linear associations between the exogenous variables. However, we could argue that linear association is not enough to make a fully informed decision about redundant variables. Therefore, we will inspect and investigate the following:

- When two exogenous variables are correlated, should the one explaining the variation of the endogenous variable be dropped?
- When one of the attributes violates the assumptions of the Pearson correlation analysis, is this technique still valid?
- Should nonlinear associations for redundancy measures be taken into consideration?

Irrelevant attributes are measured using the Spearman rank correlation coefficient to measure monotonic associations between each of the exogenous variables and the endogenous variable.

- If variables are not monotonically related to each other, which associations does the technique overlook, if any?

For smart grid applications, weather data can be forecasted using machine-learning algorithms in univariate and multivariate time series analysis. Hybrid implementation should be used during feature selection and model design for optimal accuracy. In model design phase, the authors in [6] proposed a wind forecasting model using Support Vector Regression (SVR) variants based on wavelet transform. Evaluation of different performance indices identified the optimal one for wind forecasting. Wind power ramp events were investigated as well, and indicated an increase in ramp events when hub height rises. For short-term wind speed forecasting, the authors used ϵ -SVR, Least-square support vector regression (LS-SVR), ϵ Twin support vector regression (ϵ -TSVR), and Twin Support vector regression (TSVR), comparing them with the Persistence model for windfarm sites. Regarding absolute error, ϵ -TSVR beat TSVR, LS-SVR and ϵ -SVR, showing that machine intelligent hybrid methodology improves forecasting performance,

including ramp events. The authors in [7] investigated a hybrid method based off discrete wavelet transform (DWT) and learning algorithms, e.g., Twin Support vector regression (TSVR), random forest regression (RFR), and Convolutional neural networks (CNN) for geographical features. Wavelet transform-based signal processing extracted wind-speed features, with SVR-based prediction models giving the best results, though CNN gave better results in larger training datasets. Compared to SVM, ANN and ELM, hybrid TSVR, RFR and CNN models showed improved ramp event prediction. In considering hybrid wind-battery farms, the authors in [8] proposed a penalty-cost solution based on machine intelligent wind forecasting. They compared a wavelet-Twin support vector regression (TSVR)-based wind-power forecasting model to Random Forest, ϵ Twin support vector regression, and Gradient-boosted machines, aiming to mitigate penalty cost. Results showed that wind-power forecasting using a TSVR-based method reduces global operational costs. In general terms, predictive modeling can be described as a multivariate problem in which every variable can have an impact on other input and output variables in a variety of simple or complex ways. To date, the interactions and nonlinearities that may potentially exist between variables are not yet fully researched in the literature [3], [9]. Even so, they represent critical elements for developing robust predictive models. In this work, we will focus on the measures and attributes of redundancy and relevancy and will investigate how these can be mitigated and enhanced, respectively, to develop more accurate models. The main contributions of the present work are:

- 1) Applying Hoeffding's D and Spearman's rho to the individual weather indicators that respond non-monotonically to GHI forecasting.
- 2) Analysing the stability performance of the correlation analysis for a one-year dataset and a ten-year time-frame.
- 3) Proposing the novel Weather Recursive Feature Elimination (WRFE) method for optimizing feature selection schema according to a Least-Redundant/Highest-Relevant framework.
- 4) Employing large training and testing datasets, along with conducting comprehensive statistical analysis and testing of specific features.
- 5) Assessing the effectiveness of the proposed novel WRFE approach for hourly GHI prediction by applying it in regions with different solar irradiance and weather profiles as well as comparing it to other established models.

A. FEATURE SCREENING

In every case of predictive modeling, the model's accuracy is entirely based on data quality. Therefore, it is crucial to appropriately choose and prepare exogenous (i.e., explanatory) variables as well as to determine any variations in the endogenous (i.e., response) variables. This can be accomplished by considering the most common issues that

may occur, in particular redundancy and irrelevance. Both redundancy and irrelevance can be overcome using variable screening, followed by selecting predictive variables that best suit the specified model. Pearson is a correlation statistic approach that can be applied to measure degrees of relationships existing between weather variables as given in Equation (1) [10]. This approach, however, may be invalid if the variables do not satisfy Pearson correlation assumptions. In Pearson correlation analysis, the two assumptions which need testing are: 1) the normality assumption, and 2) linearity.

$$\rho_P(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{Y})^2}} \quad (1)$$

Pearson correlations can be highly susceptible to normality and linearity assumptions and can also easily perceive outliers. Accordingly, nonparametric correlation strategies may be preferable to Pearson correlation in some cases. Examples of nonparametric strategies include Hoeffding's D and Spearman's rho. Spearman's rho is applied to gauge the direction and strength of a monotonic relationship between two variables as given in Equation (2). This is unlike the Pearson's correlation, which gauges the direction and strength of a linear relationship between two variables. In the Spearman's rho approach, the correlation between two variables is equivalent to the Pearson's correlation between rank scores from the two variables. Furthermore, whereas Pearson's correlation measures linear relationships, Spearman's correlation determines monotonic relationships (either linear or non-linear) and ranges between -1 and 1 [10].

$$\rho_S(X, Y) = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (2)$$

Hoeffding's D is applied as a non-parametric rank-based measure for determining non-linear associations, as presented in Equation (3). The measure ranges from -0.5 to 1 when no tied ranks exist; otherwise, the measure may feature lower values. In this technique, stronger associations between variables are indicated by larger values [11].

$$D(X, Y) = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)} \quad (3)$$

II. COMPREHENSIVE DATA ANALYTICS

A. DATA COLLECTION

This research uses data from solar irradiation and weather readings from the U.S. National Solar Radiation Database (NSRDB) and the U.S. National Renewable Energy Laboratory (NREL) [12]. The data were downloaded using the NSRDB data viewer for Halifax, Nova Scotia, Canada, at coordinates 44.88 N and 63.51 W. The data were collected for hourly periods using the time-frame 2000-2018. They include solar radiation irradiance and meteorological data, such as Diffuse Horizontal Irradiance (DHI), Direct Normal Irradiance (DNI), Global Horizontal Irradiance (GHI),

TABLE 1. Descriptive statistics for dataset.

Variable	count	mean	std	min	25%	50%	75%	max
DHI	82118	127.807411	97.59748	1.000	59.000	101.000	169.000	466.000
DNI	82118	333.352456	336.5912	0.000	8.000	212.000	657.000	1018.000
GHI	82118	306.29678	254.5158	1.000	93.000	236.000	471.000	1010.000
CGHI	82118	446.278161	271.4996	2.000	218.000	423.000	675.750	1010.000
DP	82118	5.887337	8.41948	-19.000	0.000	6.000	13.000	22.000
SZA	82118	60.733491	17.82613	21.390	47.330	63.700	75.040	92.040
SA	82118	0.25704	0.284197	0.098	0.119	0.125	0.135	0.866
WS	82118	2.501217	1.320333	0.000	1.500	2.300	3.300	9.900
PW	82118	1.872482	1.197914	0.089	0.878	1.642	2.665	6.764
WD	82118	209.060881	92.49892	0.000	153.800	220.500	281.700	360.000
RH	82118	82.859269	14.04239	34.350	72.530	84.600	95.830	100.000
T	82118	9.042456	8.057104	-19.000	2.000	9.100	16.000	28.000
P	82118	1005.62316	9.771198	950.000	1000.000	1010.000	1010.000	1040.000

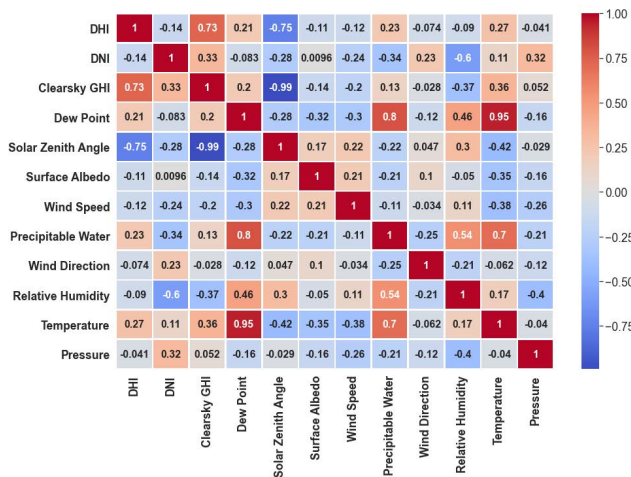


FIGURE 1. Heatmap of Pearson correlation.

Clearsky GHI (CGHI), Dew Point (DP), Solar Zenith Angle (SZA), Surface Albedo (SA), Wind Speed (WS), Precipitable Water (PW), Wind Direction (WD), Relative Humidity (RH), Temperature (T), and Pressure (P). Table 1 presents the descriptive statistics information for the dataset that includes around 82,118 observations.

B. REDUNDANCY MEASURES

Redundant attributes are usual measured by Pearson’s correlation coefficient. Figure 1 shows a heat map technique employed to visualize the correlation coefficients. As can be seen, the coefficient values (+1 to -1) measure linear associations between the exogenous variables. Inputs with linear and additive effects that have a constant rate of change on the output could be insufficient to render a full decision on the redundant variables. The effect of each input variable might have a nonlinear relationship with other

input variables, which makes the effects both nonlinear and non-additive. Thus, nonlinear associations for redundancy measure will be tested. In Tables 2 and 3, we see the Pearson and Spearman correlation coefficients for evaluating bivariate analysis and for measuring linear or monotonic relationships in the variable pairs. The coefficient values in the tables are between (+1 and -1), and the redundancy measure depends on pairwise observations of twelve common exogenous variables, namely DHI, DNI, CGHI, DP, SZA, SA, WS, PW, WD, RH, T, and P. Note that we have excluded the endogenous variable, GHI, from this analysis. All values on the diagonal are valued as 1, as the variables are perfectly correlated with themselves. We have also considered any off-diagonal elements in the matrix’s upper triangle that mirror those in the matrix’s lower triangle. The above-mentioned elements include both correlation coefficients and their respective p-values. A hypothesis test will be performed to determine any significances in the correlation coefficient and to gauge if the sample data’s linear/monotonic relationship may be sufficiently strong to apply in modeling a relationship within the population. Further, we can use the two-tailed significance test to express both the null hypothesis (H0) and alternative hypothesis (H1) of the correlation. When looking at the population correlation coefficient (p_c), we need to see if there is 95% confidence (at a 0.05 level of significance), in which case:

$H0 : p_c = 0$ (“If the population correlation coefficient equals 0, no association is detected”).

$H1 : p_c \neq 0$ (“If the population correlation coefficient does not equal 0, a nonzero correlation may exist”).

As shown in Tables 2 and 3, the Pearson and Spearman correlation coefficients for CGHI and DHI are 0.73 and 0.79, respectively. Further, because $p < .0001$, $p < 0.05$ has been satisfied, indicating that the result is statistically significant, and the null hypothesis is therefore rejected. Hence, there is enough evidence at the 0.05 significance level to assume there is a strong positive linear relationship between CGHI and DHI variables across the whole population. Furthermore, there is a strongly negative linear relationship existing between SZA and DHI, with the Pearson and Spearman correlation coefficients being -0.74 and -0.81, respectively, and $p < .0001$. Therefore, between SZA and CGHI, using the Pearson and Spearman correlation coefficients, there is a robust association of -0.99 and -0.995, respectively.

Figure 2 illustrates pairwise analysis in scatterplot form with monthly variations, showing a highly skewed DHI. As shown, between DHI and the CGHI and SZA variables, the relationships are not as robustly linear as presented in the Pearson’s correlation coefficient. Additionally, neither WS, WD, P, nor SA exhibit strong relationships with other variables, which means they would not be considered redundant variables in the model. Statistical investigation of the data presented in Table 2 gives p values to test associations between DNI and SA, CGHI and WD, and P and SZA. The

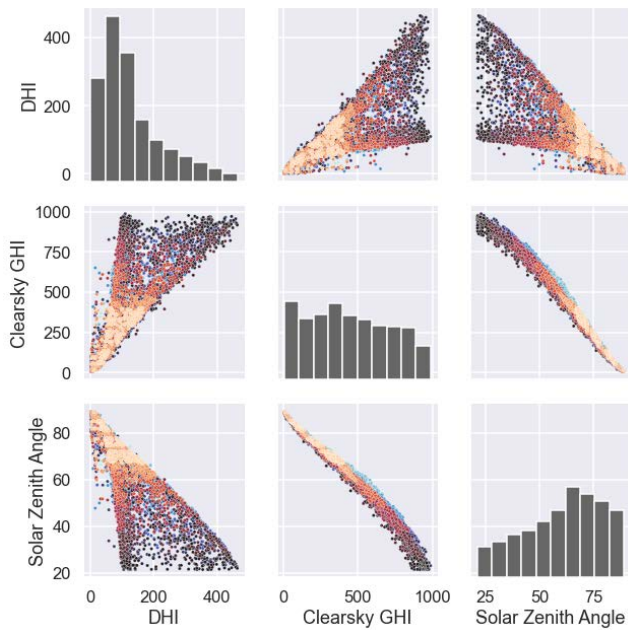


FIGURE 2. Pairwise analysis in scatterplot.

results, respectively, are 0.5269, 0.07, and 0.0577. If $p > 0.05$, the results at the 5% level are not significant, thus showing no correlation between these variables and also failing to reject the null hypothesis. In this case, the scatter plots in Figure 3 within variations of 12 clusters would be suitable for verifying these results.

C. NORMALITY ASSUMPTION TESTS

Table 4 provides descriptive statistics analyses of DHI datapoints. In order to determine whether the DHI data are normally distributed, we can apply numerical techniques by looking at kurtosis and skewness values to gauge normality according to criteria proposed in [13]. In cases where sample sizes exceed 300, histograms and absolute values of kurtosis/skewness can be considered without including z-values. If there is an absolute kurtosis exceeding 7 or an absolute skew value exceeding 2, it can be used as a reference value to determine significant levels of non-normality. Based on the above criteria, we determine that the sample data used in our test are slightly kurtotic and skewed, with kurtosis at 0.756 and skewness at 1.173. These results indicate that the sample is normally distributed according to kurtosis and skewness criteria. As a second consideration, the SAS manual [14] mentions that if the sample size exceeds 2000, the most appropriate tests are Cramer-von Mises, Kolmogorov-Smirnov, and Anderson-Darling. The Shapiro test is more suitable for sample sizes of less than 2000. In the three referenced tests for sample sizes larger than 2000, the null hypothesis applies if the data are normally distributed; otherwise, the null hypothesis will be rejected with p values < 0.05 . In our test sample, the p values show as being below .05 for

all three referenced tests, as presented in Table 5. This means that the null hypothesis is rejected and the DHI data distribution is non-normal. As a third consideration, graphical methods can be utilized in visualizing variable distribution and comparing this distribution with theoretical variable distribution by employing plots such as the Quantile-Quantile (Q-Q) plot. Figure 4 illustrates an example of DHI data that are distributed non-normally. In this instance, the Pearson's correlation may not be the most appropriate measure to find variable associations. Instead, a nonparametric approach, such as Spearman's correlation, is likely a more suitable choice. Table 6 presents both the Spearman and Pearson correlation coefficients in descending rank for DHI and a range of variables.

D. INCLUSION AND EXCLUSION CRITERIA OF EXOGENOUS VARIABLES

As mentioned earlier in this research, variable screening is an effective way to decrease excess exogenous variables, as this form of screening is able to identify variables that are redundant. In the current context, the redundancy measure for considering very high correlated variables using the Spearman's technique is coefficients larger than 0.8 in value. The working hypothesis is that the model's performance may be impeded by exogenous variables with monotonic associations. In our prior example, the two exogenous variable subsets of CGHI and SZA, along with T and DP, are all highly correlated, making them redundant. For variable inclusion, we need to investigate which exogenous variable should be dropped in cases where they are correlated. This investigation will be presented in more detail in the proposed WRFE technique. Table 7 provides a list of highly correlated variables which could potentially be redundant. As can be seen, there is a positive monotonic relationship between CGHI and DHI, where $(\rho_S) = 0.73$. Additionally, we can see that there is a negative monotonic relationship between SZA and DHI, where $(\rho_S) = -0.74$. There is also a strongly negative association between SZA and CGHI, where $(\rho_S) = -0.99$. In the same table, a negative moderate association exists between DNI and RH, where $(\rho_S) = -0.6$. Further, in comparison to other data, DP shows the strongest positive monotonic and linear relationships to T (up to $(\rho_S) = 0.95$), and a positive, monotonic, and curvilinear relationship (up to $(\rho_S) = 0.79$) to PW. There is also a moderate association of $(\rho_S) = 0.69$ between PW and T.

E. EFFECT OF SAMPLE SIZE IN CORRELATION ANALYSIS OF WEATHER DATA

To test stability visually, we made a comparison of the correlation analysis of a one-year dataset. For the year 2000, there are 4309 datapoint observations, while for the ten-year time-frame of our study period (2000-2010), there are 47,407. Thus, we observed a relatively stable magnitude of correlations both in large and small data samples. Moreover, the majority of the correlation coefficients within the dataset appeared entirely stable in relation to the dataset size,

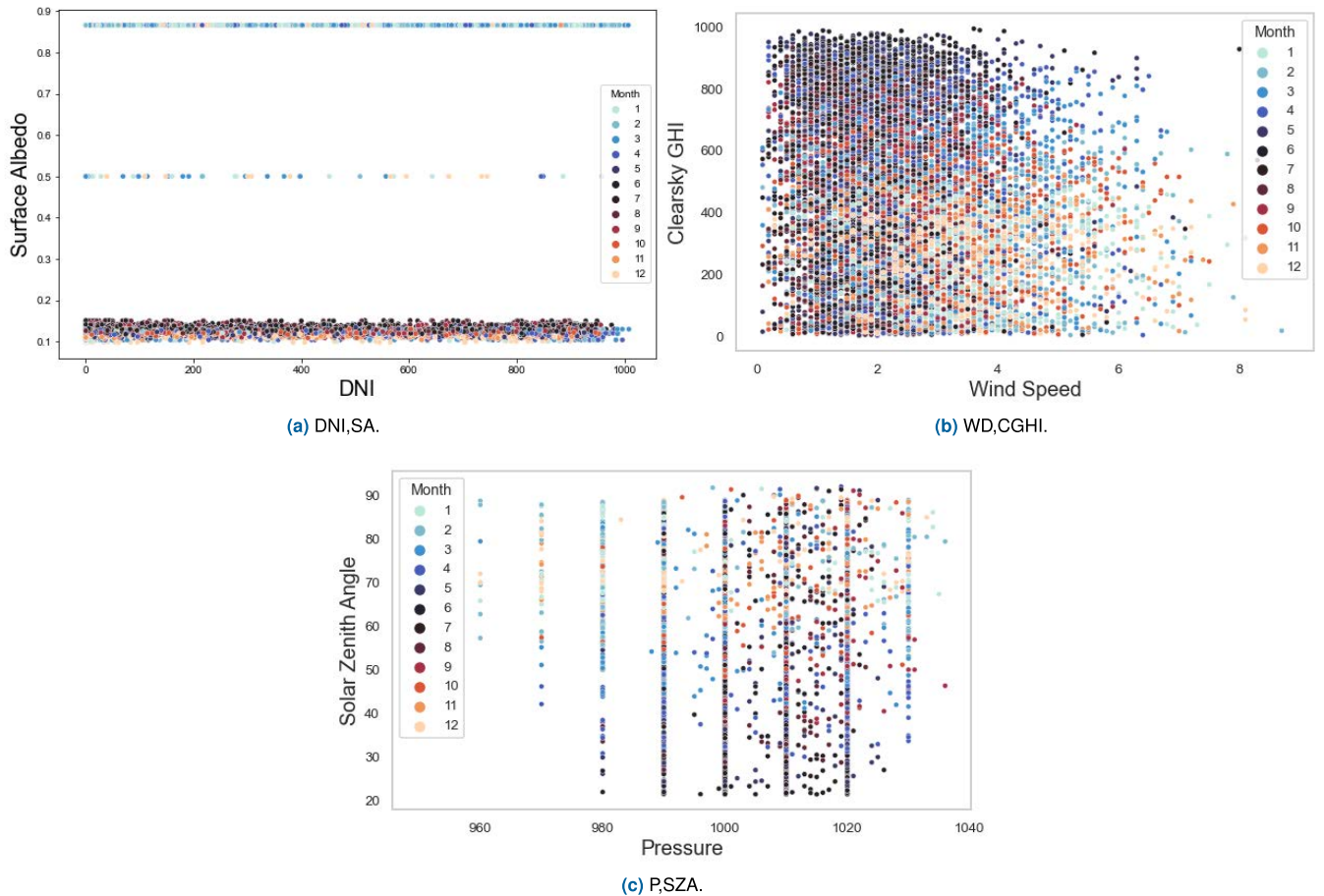


FIGURE 3. Scatter plots illustrating two variables for (a) DNI and SA, (b) WD and CGHI, (c) P and SZA.

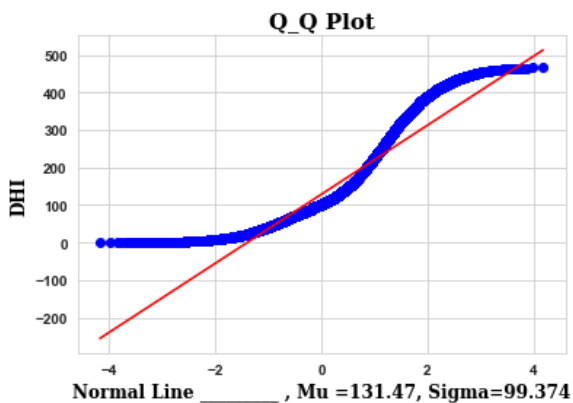


FIGURE 4. QQ plot for DHI.

as shown in Table 3 and Table 8, whereas some were stable but featured slight fluctuations around their true value. This type of deviation, however, is considered trivial and is therefore tolerable. Examples of ρ_S changes are as follows: DNI and DHI, ρ_S changed from 0.06 to 0.13; CGHI and SA, ρ_S changed from 0.125 to 0.029; CGHI and P, ρ_S changed from 0.04 to 0.02; DP and DNI, ρ_S changed from -0.04 to

-0.09 ; DP and P, ρ_S changed from -0.136 to -0.06 ; SA and T, ρ_S changed from 0.29 to -0.119 ; SA and WD, ρ_S changed from 0.016 to 0.04; and T and P, ρ_S changed from -0.028 to 0.026. As well, there were a few fluctuations in some other correlation coefficients, changing, for instance, from a significantly weak association to a significantly very weak or null association. The correlation coefficients in other instances changed from a significantly weak association direction into its opposite significantly weak association. For instance, for SA and RH, ρ_S changed from 0.07 to (-0.02) ; for SA and P, ρ_S changed from (-0.079) to (-0.05) ; and for WS and WD, ρ_S changed from (-0.007) to (0.06). The strongest deviations recorded were in the associations between SA and SZA, SA and WS, and SA and DP. These were recorded as being from -0.14 to -0.0088 , -0.17 to 0.008, and 0.3 to -0.08 , respectively. However, as our research setting makes allowances for moderate associations when using Spearman’s coefficients with ρ_S values greater than 0.4, these deviations are not considered problematic. On the other hand, as most deviations in the correlation coefficients occurred in the yearly dataset (correlation coefficients differ from year to year), they may warrant further investigation.

TABLE 2. Pearson correlation coefficients for the year 2000.

Pearson Correlation Coefficients, N = 4309												
Prob > r under H0: Rho=0												
	DHI	DNI	CGHI	DP	SZA	SA	WS	PW	WD	RH	T	P
DHI	1.0000	-0.13899	0.73209	0.21209	-0.74764	-0.10561	-0.11931	0.22914	-0.07357	-0.09034	0.26591	-0.04073
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0075
DNI	-0.13899	1.0000	0.33324	-0.08344	-0.28413	0.00964	-0.23606	-0.3391	0.22909	-0.60108	0.11489	0.31971
	<.0001		<.0001	<.0001	<.0001	0.5269	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
CGHI	0.73209	0.33324	1.0000	0.20498	-0.99034	-0.14244	-0.20453	0.13291	-0.02757	-0.37175	0.36083	0.05199
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0703	<.0001	<.0001	0.0006
DP	0.21209	-0.08344	0.20498	1.0000	-0.27856	-0.32152	-0.29901	0.79862	-0.11717	0.46139	0.95048	-0.15724
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
SZA	-0.74764	-0.28413	-0.99034	-0.27856	1.0000	0.1684	0.22045	-0.21903	0.04672	0.30492	-0.41955	-0.02892
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.0022	<.0001	<.0001	0.0577
SA	-0.10561	0.00964	-0.14244	-0.32152	0.1684	1.0000	0.2056	-0.21435	0.10165	-0.04964	-0.34834	-0.15502
	<.0001	0.5269	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.0011	<.0001	<.0001
WS	-0.11931	-0.23606	-0.20453	-0.29901	0.22045	0.2056	1.0000	-0.1099	-0.03394	0.11374	-0.37645	-0.26426
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.0259	<.0001	<.0001	<.0001
PW	0.22914	-0.3391	0.13291	0.79862	-0.21903	-0.21435	-0.1099	1.0000	-0.2461	0.54317	0.69789	-0.21472
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
WD	-0.07357	0.22909	-0.02757	-0.11717	0.04672	0.10165	-0.03394	-0.2461	1.0000	-0.20632	-0.06234	-0.12331
	<.0001	<.0001	0.0703	<.0001	0.0022	<.0001	0.0259	<.0001		<.0001	<.0001	<.0001
RH	-0.09034	-0.60108	-0.37175	0.46139	0.30492	-0.04964	0.11374	0.54317	-0.20632	1.0000	0.17011	-0.39697
	<.0001	<.0001	<.0001	<.0001	<.0001	0.0011	<.0001	<.0001	<.0001		<.0001	<.0001
T	0.26591	0.11489	0.36083	0.95048	-0.41955	-0.34834	-0.37645	0.69789	-0.06234	0.17011	1.0000	-0.03984
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.0089
P	-0.04073	0.31971	0.05199	-0.15724	-0.02892	-0.15502	-0.26426	-0.21472	-0.12331	-0.39697	-0.03984	1.0000
	0.0075	<.0001	0.0006	<.0001	0.0577	<.0001	<.0001	<.0001	<.0001	<.0001	0.0089	

TABLE 3. Spearman correlation coefficients for the year 2000.

Spearman Correlation Coefficients, N = 4309												
Prob > r under H0: Rho=0												
	DHI	DNI	CGHI	DP	SZA	SA	WS	PW	WD	RH	T	P
DHI	1.0000	0.06253	0.79752	0.18534	-0.8105	0.11218	-0.1081	0.21123	-0.09791	-0.17012	0.26578	-0.05321
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0005
DNI	0.06253	1.0000	0.36116	-0.04771	-0.32036	0.08802	-0.24591	-0.34968	0.24777	-0.60306	0.1476	0.32153
	<.0001		<.0001	0.0017	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
CGHI	0.79752	0.36116	1.0000	0.18883	-0.99505	0.12592	-0.19495	0.14136	-0.052	-0.39823	0.36101	0.04985
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0006	<.0001	<.0001	0.0011
DP	0.18534	-0.04771	0.18883	1.0000	-0.25395	0.31886	-0.2812	0.83268	-0.15325	0.39611	0.94648	-0.13649
	<.0001	0.0017	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
SZA	-0.8105	-0.32036	-0.99505	-0.25395	1.0000	-0.1417	0.20496	-0.21981	0.07845	0.34129	-0.41371	-0.02944
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0533
SA	0.11218	0.08802	0.12592	0.31886	-0.1417	1.0000	-0.17852	0.26007	0.01688	0.07367	0.29539	-0.07945
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.2679	<.0001	<.0001	<.0001
WS	-0.1081	-0.24591	-0.19495	-0.2812	0.20496	-0.17852	1.0000	-0.12072	-0.00772	0.12265	-0.35919	-0.22882
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.6122	<.0001	<.0001	<.0001
PW	0.21123	-0.34968	0.14136	0.83268	-0.21981	0.26007	-0.12072	1.0000	-0.31536	0.56267	0.72122	-0.24245
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
WD	-0.09791	0.24777	-0.052	-0.15325	0.07845	0.01688	-0.00772	-0.31536	1.0000	-0.23528	-0.10096	-0.14502
	<.0001	<.0001	0.0006	<.0001	<.0001	0.2679	0.6122	<.0001		<.0001	<.0001	<.0001
RH	-0.17012	-0.60306	-0.39823	0.39611	0.34129	0.07367	0.12265	0.56267	-0.23528	1.0000	0.11297	-0.38191
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
T	0.26578	0.1476	0.36101	0.94648	-0.41371	0.29539	-0.35919	0.72122	-0.10096	0.11297	1.0000	-0.02882
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.0586
P	-0.05321	0.32153	0.04985	-0.13649	-0.02944	-0.07945	-0.22882	-0.24245	-0.14502	-0.38191	-0.02882	1.0000
	0.0005	<.0001	0.0011	<.0001	0.0533	<.0001	<.0001	<.0001	<.0001	<.0001	0.0586	

TABLE 4. Descriptive statistics analyses of DHI.

Moments			
N	4309	Sum Weights	4309
Mean	131.470179	Sum Observations	566505
Std Deviation	99.3737954	Variance	9875.15121
Skewness	1.17306489	Kurtosis	0.75623336
Coeff Variation	75.5865675	Std Error Mean	1.51385274

TABLE 5. Statistics tests for normality assumption of DHI.

Goodness-of-Fit Tests for Normal Distribution			
Test	Statistic		pValue
Kolmogorov-Smirnov	D	0.146558	Pr > D < 0.010
Cramer-von Mises	W-Sq	27.356580	Pr > W-Sq < 0.005
Anderson-Darling	A-Sq	156.371578	Pr > A-Sq < 0.005

F. RELEVANCE MEASURES

Evaluation of irrelevancy for irrelevant attributes is commonly done using Spearman’s ranking of correlation coefficients. This can be performed by measuring monotonic associations between endogenous and exogenous variables [9], [15]. If the two measured variables present as being monotonically unrelated, key associations could be overlooked. In some cases, Hoeffding’s D statistic value can be applied in conjunction with Spearman’s analysis in order to identify non-monotonic associations which may not be identified when only using Spearman’s. As demonstrated in [16], if the Spearman rank shows as being high, this indicates a monotonic association, even if the corresponding Hoeffding’s value is low. In general, however, monotonic associations are key elements in predictive modeling. When the Hoeffding rank is high and the Spearman rank is low, the association is considered non-monotonic. This pattern of nonlinearity needs further investigation in order to gauge if and how the association might impact the model’s performance. On the other hand, if Hoeffding’s is low and Spearman’s is also low, this indicates a vulnerable association, which means the attributes are irrelevant and can be eliminated. Table 9 presents a comparison of Hoeffding’s D and Spearman’s correlation coefficients. As can be seen, CGHI, SZA, DNI, DHI, RH, and T are all deemed relevant attributes to GHI, which is the target. In this comparison, DNI, CGHI, and SZA are the highest individual relevant attributes. The results are then validated for stability via dataset testing. These datasets were collected in approximate increments of five years (2000, 2005, 2010, 2015, and 2018) as well as for the 11-year dataset for the study period (2000-2010) for Halifax, NS, as shown in Table 10. The completed results of the validation are given in the supplemental materials. Although we can employ Spearman’s rank correlation coefficients on different data

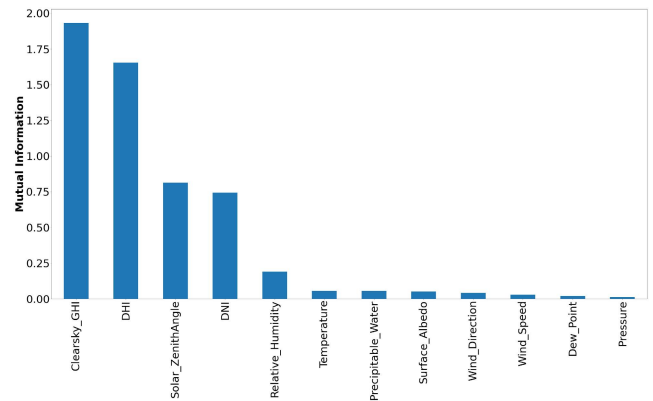


FIGURE 5. Degree of mutual information between exogenous variables and GHI.

sizes, Gilpin [17] found that with increases in sample size, the Kendall correlation coefficient is more practical. Croux and Dehon [18] agree that the Kendall correlation performs better than the Spearman correlation in this regard due to its smaller GES (gross error sensitivity), which makes it more robust, and its smaller AV (asymptotic variance), which increases its efficiency. The authors in [19], [20] mention that the Kendall correlation has a computation complexity of $O(n^2)$ in comparison to the $O(n * \log n)$ complexity of the Spearman correlation, with n being sample size. In which case, the best approach might be to use both techniques when dealing with large sample sizes. We performed an intensive screening of the features using filter methods that rely on the data’s statistical characteristics, such as parametric and non-parametric tests. Equation 4 illustrates a way to capture dependency degree between i th exogenous variables (x) and endogenous variable (y) [21]. Strong dependence shows a high degree of mutual information, which indicates greater knowledge of joint distribution $p(x, y)$ than marginal distribution $p(x)p(y)$. We applied the normalized mutual information (NMIFS) method proposed in [22] as a measure of irrelevancy. For both methods, Figure 5 validates the results, with CGHI, DHI, SZA, DNI, RH and T all being features that appear to make the greatest contributions to the prediction GHI.

$$I(X; Y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \quad (4)$$

III. WEATHER RECURSIVE FEATURE ELIMINATION (WRFE)

After implementing exploratory data analysis to the response and explanatory variables, we can conclude that six features (CGHI, SZA, DNI, DHI, RH, and T) appear to have the closest relation with the target (GHI). Even so, we need to consider our previous finding, which was that CGHI and SZA (along with DP and T) are redundant attributes. Here, we can eliminate DP, because it is an unrelated variable (i.e., it is not one of our six above-mentioned features). However, for the correlated variables, CGHI and SZA,

TABLE 6. Nonparametric measure of association with DHI for a 95 % confidence interval.

Variable	Pearson Rank	Spearman Rank	Pearson Coefficient	Pearson P-Value	Spearman Coefficient	Spearman P-Value
SZA	1	1	-0.8105	<.0001	-0.74764	<.0001
CGHI	2	2	0.79752	<.0001	0.73209	<.0001
T	3	3	0.26578	<.0001	0.26591	<.0001
PW	4	4	0.21123	<.0001	0.22914	<.0001
DP	5	5	0.18534	<.0001	0.21209	<.0001
RH	6	9	-0.17012	<.0001	-0.09034	<.0001
SA	7	8	0.11218	<.0001	-0.10561	<.0001
WS	8	7	-0.1081	<.0001	-0.11931	<.0001
WD	9	10	-0.09791	<.0001	-0.07357	<.0001
DNI	10	6	0.06253	<.0001	-0.13899	<.0001
P	11	11	-0.05321	0.0005	-0.04073	0.0075

TABLE 7. List of high correlated variables.

First Feature	Second Feature	Spearman Coefficients
DHI	CGHI	0.73
DHI	SZA	-0.74
CGHI	SZA	-0.99
DNI	RH	-0.6
DP	T	0.95
DP	PW	0.79
T	PW	0.69
RH	PW	0.56

we need to consider the issue mentioned earlier, namely regarding which of the two exogenous variables should be dropped if they are correlated. To resolve this issue, we simply need to look at the mutual information and correlation coefficients between these two variables and GHI. In this case, CGHI is obviously more relevant to the response variable. However, in other cases, the variables could be non-predictive when in isolation, but highly predictive in combination with others. When this occurs, we need to perform a subset evaluation to determine relevance. This can be done using a hybrid feature reduction method that utilizes Weather Recursive Feature Elimination (WRFE). In this method, the feature selection process is implemented through designing two different machine learning models. The idea here is to measure each explanatory variable’s contribution to the final prediction, which can be done by considering the importance measures for the various features of each model.

A. METHODOLOGY

When adopting this approach, one first needs to design LSTM and RFR models, using the six mentioned features. Next, RMSE needs to be used to calculate the performance, followed by a calculation of feature importance by looking at the impurity measure (variance reduction) for the RFR model and data perturbation for the LSTM model. The final step is

to remove the least importance feature and then to design the two models using the remaining features. Once this is done, the performance of the new models can be compared with that of the full model performance by using the new RMSE. If the new RMSE is calculated to be larger than the full model’s RMSE, the eliminated feature is important and should be kept. We can also compare any reductions in performance. If there is a reduction in performance in comparison to a user-defined threshold (here considered 2.5), the feature should be eliminated in cases where the drop is smaller than the threshold. In cases where it is larger, the feature should be retained. The threshold of 2.5 was firstly selected arbitrary with performed model tuning, then selected it based on systematic observation for the set of the performance’s drop. Figure 6 demonstrates the flowchart of the proposed WRFE, with algorithm (1) showing the pseudo-code of the proposed procedure.

1) MODEL IMPLEMENTATION

Accordingly, we measure the feature importance for the Halifax, NS, dataset for the year 2000. The dataset includes the six above-mentioned features. As our first step of algorithm (1), we design and train an LSTM model as establishing in [1] and calculate predictions of the model. Next, as stated in [23], we perturb each feature by adding the random Gaussian distribution noise (mean $\mu = 0$, standard deviation σ), with probability function $f(d)$ as defined in Equation 5, and then calculate the perturbed prediction.

$$f(d) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d^2}{2\sigma^2}} \tag{5}$$

where d is the Euclidean distance between original feature (x_i) and perturbed feature (\hat{x}_i), as defined in 6

$$d(x_i, \hat{x}_i) = \sqrt{\sum_{i=1}^n (x_i - \hat{x}_i)^2} \tag{6}$$

When that is completed, we measure the perturbation effects in gradients. This is done by calculating RMSE

TABLE 8. Spearman correlation coefficients for the years 2000-2010.

Spearman Correlation Coefficients, N = 47407
 Prob >|r| under H0: Rho=0

	DHI	DNI	CGHI	DP	SZA	SA	WS	PW	WD	RH	T	P
DHI	1.0000	0.13588 <.0001	0.78745 <.0001	0.17698 <.0001	-0.7967 <.0001	0.06925 <.0001	-0.12588 <.0001	0.18016 <.0001	-0.07869 <.0001	-0.19838 <.0001	0.2552 <.0001	-0.01956 <.0001
DNI	0.13588 <.0001	1.0000	0.33417 <.0001	-0.09756 <.0001	-0.29018 <.0001	0.06314 <.0001	-0.17986 <.0001	-0.34686 <.0001	0.25302 <.0001	-0.59642 <.0001	0.08659 <.0001	0.26499 <.0001
CGHI	0.78745 <.0001	0.33417 <.0001	1.0000	0.17999 <.0001	-0.9931 <.0001	0.02914 <.0001	-0.16265 <.0001	0.14329 <.0001	-0.05763 <.0001	-0.36557 <.0001	0.32809 <.0001	0.02208 <.0001
DP	0.17698 <.0001	-0.09756 <.0001	0.17999 <.0001	1.0000	-0.26517 <.0001	-0.08946 <.0001	-0.2888 <.0001	0.87228 <.0001	-0.19075 <.0001	0.42865 <.0001	0.95317 <.0001	-0.06206 <.0001
SZA	-0.7967 <.0001	-0.29018 <.0001	-0.9931 <.0001	-0.26517 <.0001	1.0000	-0.00886 0.0536	0.18242 <.0001	-0.23831 <.0001	0.08875 <.0001	0.30181 <.0001	-0.40054 <.0001	-0.01005 0.0287
SA	0.06925 <.0001	0.06314 <.0001	0.02914 <.0001	-0.08946 <.0001	-0.00886 0.0536	1.0000	0.00834 0.0692	-0.07357 <.0001	0.0483 <.0001	-0.02468 <.0001	-0.11921 <.0001	-0.05911 <.0001
WS	-0.12588 <.0001	-0.17986 <.0001	-0.16265 <.0001	-0.2888 <.0001	0.18242 <.0001	0.00834 0.0692	1.0000	-0.20134 <.0001	0.06741 <.0001	0.0614 <.0001	-0.34695 <.0001	-0.26374 <.0001
PW	0.18016 <.0001	-0.34686 <.0001	0.14329 <.0001	0.87228 <.0001	-0.23831 <.0001	-0.07357 <.0001	-0.20134 <.0001	1.0000	-0.32957 <.0001	0.58005 <.0001	0.76939 <.0001	-0.13749 <.0001
WD	-0.07869 <.0001	0.25302 <.0001	-0.05763 <.0001	-0.19075 <.0001	0.08875 <.0001	0.0483 <.0001	0.06741 <.0001	-0.32957 <.0001	1.0000	-0.25423 <.0001	-0.13707 <.0001	-0.12098 <.0001
RH	-0.19838 <.0001	-0.59642 <.0001	-0.36557 <.0001	0.42865 <.0001	0.30181 <.0001	-0.02468 <.0001	0.0614 <.0001	0.58005 <.0001	-0.25423 <.0001	1.0000	0.16244 <.0001	-0.29188 <.0001
T	0.2552 <.0001	0.08659 <.0001	0.32809 <.0001	0.95317 <.0001	-0.40054 <.0001	-0.11921 <.0001	-0.34695 <.0001	0.76939 <.0001	-0.13707 <.0001	0.16244 <.0001	1.0000	0.0267 <.0001
P	-0.01956 <.0001	0.26499 <.0001	0.02208 <.0001	-0.06206 <.0001	-0.01005 0.0287	-0.05911 <.0001	-0.26374 <.0001	-0.13749 <.0001	-0.12098 <.0001	-0.29188 <.0001	0.0267 <.0001	1.0000

TABLE 9. Nonparametric relevance measure for a 95% confidence interval (Year 2000).

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Coefficient	Spearman p-value	Hoeffding Coefficient	Hoeffding p-value	Kendall Coefficient	Kendall p-value
CGHI	1	1	1	0.83719	<.0001	0.34438	<.0001	0.66575	<.0001
SZA	2	2	2	-0.81605	<.0001	0.30864	<.0001	-0.63577	<.0001
DNI	3	3	3	0.77333	<.0001	0.23712	<.0001	0.58447	<.0001
DHI	4	4	4	0.61329	<.0001	0.18468	<.0001	0.47476	<.0001
RH	5	5	5	-0.57718	<.0001	0.11794	<.0001	-0.40491	<.0001
T	6	6	6	0.34318	<.0001	0.03636	<.0001	0.23185	<.0001
WS	7	7	7	-0.27082	<.0001	0.02228	<.0001	-0.18482	<.0001
P	8	8	8	0.20816	<.0001	0.01179	<.0001	0.15667	<.0001
SA	9	9	9	0.14156	<.0001	0.00923	<.0001	0.09669	<.0001
DP	10	10	10	0.12775	<.0001	0.00659	<.0001	0.08435	<.0001
WD	11	11	11	0.09434	<.0001	0.00477	<.0001	0.06071	<.0001
PW	12	12	12	-0.07328	<.0001	0.00284	<.0001	-0.04688	<.0001

for the perturbed and original forecasts. In our case, the calculation is given via the gradient values we obtained from performing a differentiation operation on the forecasts' input sequences. A large difference in RMSE indicates the high importance of the variable in the system (see Table 12). We also design and train an RFR model and calculate

feature importance according to reductions of variance (node impurity). The capability of RFR as an ensemble learning-based technique that leverages the power of numerous decision trees for processing large data and enhancing forecasting decision capabilities and for handling the variance reduction criteria [24]. The designed forest model is an

TABLE 10. Nonparametric relevance measure for a 95% confidence interval (Years 2000-2010).

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Coefficient	Spearman p-value	Hoeffding Coefficient	Hoeffding p-value	Kendall Coefficient	Kendall p-value
CGHI	1	1	1	0.79765	<.0001	0.3044	<.0001	0.62727	<.0001
DNI	2	3	2	0.78819	<.0001	0.24958	<.0001	0.5996	<.0001
SZA	3	2	3	-0.77335	<.0001	0.27088	<.0001	-0.59547	<.0001
DHI	4	4	4	0.64949	<.0001	0.21172	<.0001	0.50311	<.0001
RH	5	5	5	-0.5651	<.0001	0.10911	<.0001	-0.39649	<.0001
T	6	6	6	0.28267	<.0001	0.02411	<.0001	0.19036	<.0001
WS	7	7	7	-0.22272	<.0001	0.01454	<.0001	-0.15107	<.0001
P	8	8	8	0.17158	<.0001	0.00757	<.0001	0.12928	<.0001
WD	9	9	9	0.10625	<.0001	0.00455	<.0001	0.06894	<.0001
PW	10	11	10	-0.08609	<.0001	0.00279	<.0001	-0.0565	<.0001
DP	11	10	11	0.08551	<.0001	0.00337	<.0001	0.05617	<.0001
SA	12	12	12	0.06015	<.0001	0.00261	<.0001	0.04032	<.0001

ensemble of T decision trees, each comprising split and leaf nodes, as inspired by those proposed in [25], [26]. Each split node (s) consists of a normalized feature F_n and a threshold τ . We calculate the variance for every single leaf node (l_j) that is related to a particular split node as given in Equation 7:

$$\sigma_l^j = \frac{\sum_{j=1}^m (x_j - \mu_j)^2}{N_j} \tag{7}$$

Then we compute the variance of each (s) as the weighted average variance of (l), as given in Equation 8:

$$w(\sigma_s^n) = \frac{\sum_{j=1}^m w_j * \sigma_l^j}{\sum_{j=1}^m w_j} \tag{8}$$

where w_j denotes the weight applied to x_j values in (s). The optimal splitting selection rules are determined by running repeated selections to minimize the variance of a specific split node. The greater the reduction in variance, the higher that feature’s importance is in the system (see Table13). Subsequently, we project input data into lower-dimensional feature space by finding an optimal input feature subset. This is done using both statistics descriptors and a hybrid technique for detecting interactions that may occur between features. Our optimal subset includes CGHI, RH, DNI, and DHI. Forecasting models have been trained using hourly observation data from 2000 to 2002; they have also been tested using data from 2003. The training dataset contains 12937 hours, while the testing dataset contains 4310 hours. The forecasting models have been designed using Keras, as applied to TensorFlow 2.0. Table 11 shows the hyperparameter values for the proposed LSTM and RFR models, while Figure 7 demonstrates the inspection of feature importance according to data perturbation and variance reduction.

TABLE 11. Hyperparameter values for LSTM and RFR.

LSTM Hyperparameter	Values	RFR Hyperparameter	Values
Learning rate	0.001	No. of tress	100
Batch size	24	Max feature	6
Optimizer	Adam	Max depth	10
No. of Epochs	120	Min samples split	4
Input shape	3-D	Min samples leaf	1
No. of hidden layer	3	Criterion	variance reduction
No. of units in each hidden layer	100	Class weight	balanced
No. of units in output layer	1	Min weight fraction leaf	0.1
Dropout rate	0.1	Random state	0

TABLE 12. Feature inspection via LSTM model.

Features	RMSE1 (W/m ²)	RMSE2 (W/m ²)	Drop in Performance	Decision
CGHI	45.65	49.41	3.76	Keep
DNI	45.65	48.9	3.25	Keep
DHI	45.65	48.36	2.71	keep
SZA	45.65	43.86	2.42	Eliminate
RH	45.65	48.29	2.64	Keep
T	45.65	43.23	1.79	Eliminate

TABLE 13. Feature inspection via RFR model.

Features	RMSE1 (W/m ²)	RMSE2 (W/m ²)	Drop in Performance	Decision
CGHI	52.83	57.39	4.56	Keep
DNI	52.83	56.27	3.44	Keep
DHI	52.83	56.21	3.38	keep
SZA	52.83	55.01	2.18	Eliminate
RH	52.83	56.73	3.9	Keep
T	52.83	55.13	2.3	Eliminate

2) FORECASTING RESULTS AND ANALYSIS

LSTM models are employed to discover seasonality pattern of the previous 24 hours for the respective input features.

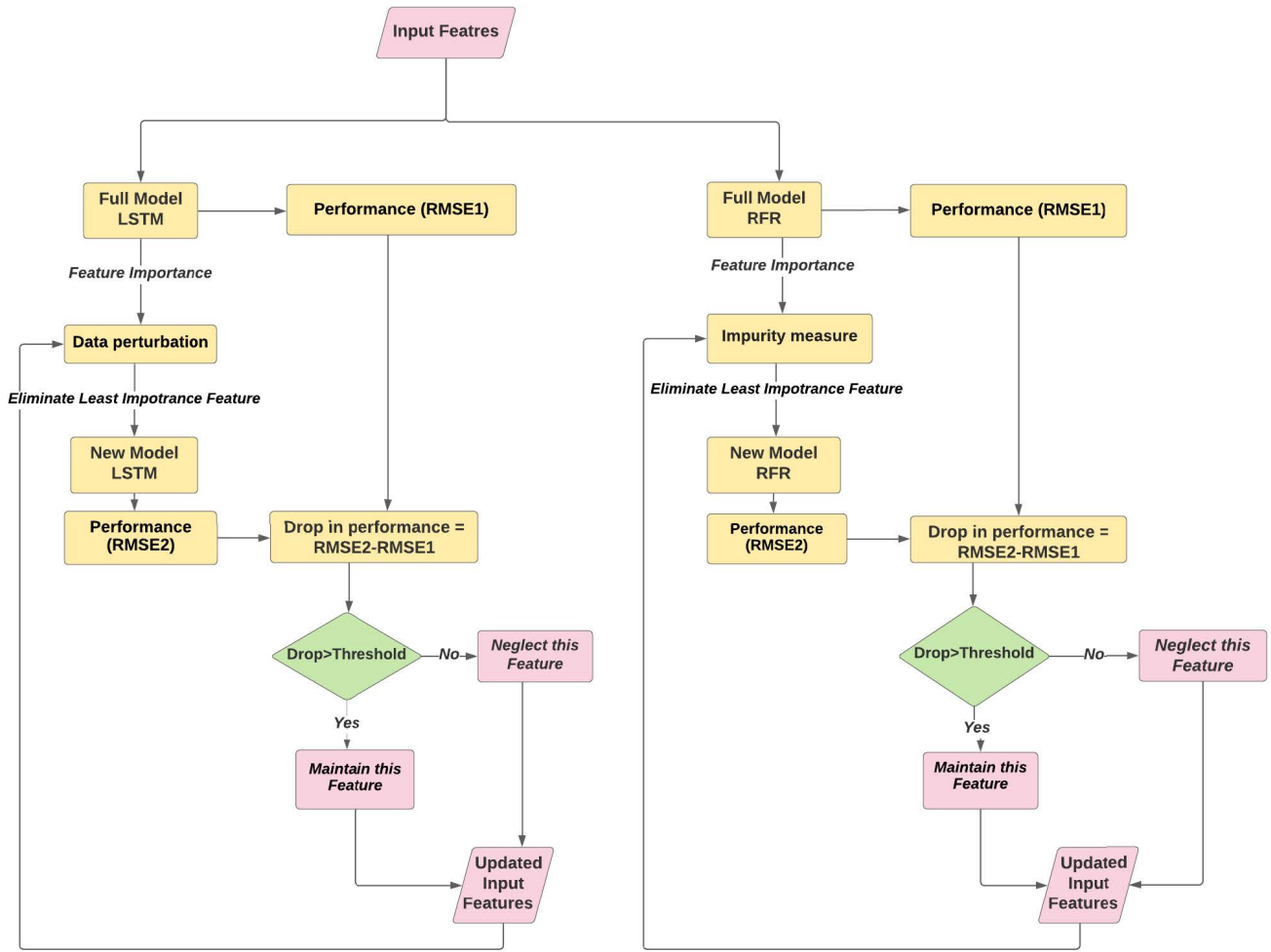


FIGURE 6. Flowchart of proposed WRFE.

These patterns are then utilized for predicting subsequent GHI for time (T+1). The input training set formation considers adjacent 24 values for the N respective input features for time T. Specifically, it creates a 3-D tensor (12913, 24, N) and the GHI output training set at size (12913, 1). RMSE values are utilized for performance verification of the designed models that belong to weather data for different locations in Canada. From Figure 8, we can see that models with the resulting optimal set of (CGHI, DNI, DHI, RH) gave a better performance than the rest in GHI prediction, with the lowest RMSE given by the model with the six features (CGHI, DNI, DHI, RH, SZA). When separately adding exogenous variables to the LSTM models, we can see that the RMSE between observed and estimated GHI is affected. Table 14 presents dataset from regions described by different climatic conditions. As shown, adding T to LSTM model bumps the RMSE up to 2.068 %, while adding SZA reduces the RMSE to 1.824 %. This study of investigating the changes of seasonality effects on the LSTM’s learning task that has proposed in [27], [28]. We believe these changes warrant future investigation of seasonality

patterns in the weather data through capturing nonlinearity patterns embedded in the exogenous and endogenous variables.

3) MODEL VALIDATION

Gray’s relational analysis [29] is applied to the stage of model selection to calculate grey relational degree and determine the influence measure of the primary behavior of each set of the input feature to the model’s performance. This analysis includes of measuring Gray’s correlation coefficient as given in Equation 9.

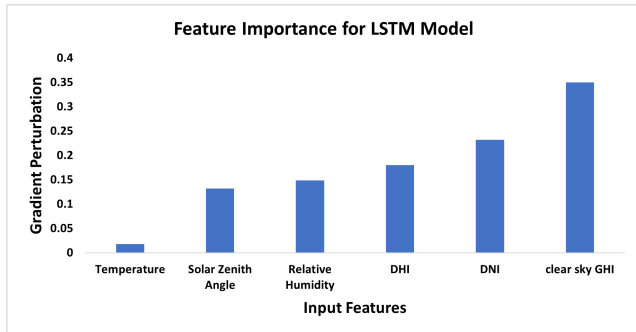
$$\gamma(x_0(k), x_i(k)) = \frac{x(\min) + \zeta x(\max)}{\Delta_{0_i}(k) + \zeta x(\max)} \quad (9)$$

$\Delta_{0_i}(k)$ is the deviation sequence given in Equation 10, and $\zeta = 0.5$ is distinguishing coefficient.

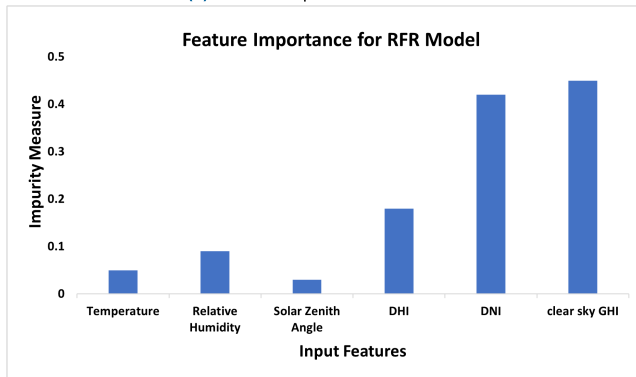
$$\Delta_{0_i}(k) = x_0(k) - x_i(k) \quad (10)$$

TABLE 14. Performance comparison of the proposed WRFE for different locations in Canada.

Locatin	Input featur	RMSE(%)	MBE(%)	R ² (%)	t-statistics	Rank	critical t-value
Halifax, NS	CGHI, DNI, DHI, RH	1.1044	0.6723	98.5112	3.0923	1	3.9043
	CGHI, DNI, DHI, RH, T	2.0682	0.9760	97.5383	3.2093	3	
	CGHI, DNI, DHI, RH, SZA	1.8246	0.8508	98.4306	3.9042	2	
	CGHI, DNI, DHI, RH, P	3.1186	1.2097	96.0105	3.9910	5	
	CGHI, DNI, DHI, RH, PW	2.8403	1.0376	96.9954	4.0154	4	
Calgary, AB	CGHI, DNI, DHI, RH	2.0153	0.8297	98.0233	3.9042	1	4.1003
	CGHI, DNI, DHI, RH, T	2.9085	0.9842	97.2129	4.0283	3	
	CGHI, DNI, DHI, RH, SZA	2.2847	1.0096	98.0054	4.1029	2	
	CGHI, DNI, DHI, RH, P	3.4913	1.1760	96.1143	2.8903	4	
	CGHI, DNI, DHI, RH, PW	4.0202	1.0982	96.0932	2.9043	5	
Thunder Bay, ON	CGHI, DNI, DHI, RH	3.1934	1.1043	97.4372	5.0214	1	5.8091
	CGHI, DNI, DHI, RH, T	4.5213	1.3060	96.0854	4.9063	4	
	CGHI, DNI, DHI, RH, SZA	3.7783	1.1034	97.7086	4.6790	2	
	CGHI, DNI, DHI, RH, P	4.1802	1.0990	95.3947	4.9042	3	
	CGHI, DNI, DHI, RH, PW	4.8842	1.3011	95.0130	4.0127	5	
Victoria, BC	CGHI, DNI, DHI, RH	1.0927	0.5092	98.3333	2.8035	1	3.2064
	CGHI, DNI, DHI, RH, T	1.6315	0.9894	97.8704	3.0852	3	
	CGHI, DNI, DHI, RH, SZA	1.4184	1.5209	98.8653	3.6013	2	
	CGHI, DNI, DHI, RH, P	3.0529	1.0371	96.6132	3.4072	5	
	CGHI, DNI, DHI, RH, PW	2.0092	1.2093	96.9422	4.0252	4	



(a) Feature Importance for LSTM.



(b) Feature Importance for RFR.

FIGURE 7. Feature importance according to data perturbation and variance reduction.

This step is followed by calculating Gray relational degree/grade as given in Equation 11

$$\gamma(x_0, x_i) = \frac{1}{n} \sum_{k=1}^n \gamma(x_0(k), x_i(k)) \quad (11)$$

Algorithm 1 Proposed WRFE Method

- 1: Input: a data set of n features $M(F_1, F_2, \dots, F_n)$
- 2: Output: Optimal feature subset M_{best}
- 3: Phase I- Modeling LSTM
- 4: Design LSTM utilizing the n features
- 5: Calculate RMSE1 (RMSE for full model performance)
- 6: **for** $i \leftarrow 1, n$ **do**
- 7: Eliminating least importance feature f_i
- 8: Perturbing the n features
 $f_1 + d_1, f_2 + d_2, \dots, f_n + d_n$
- 9: Calculating perturbed prediction error RMSE2 from the perturbed dataset $M(F_{pn})$
- 10: Calculating the drop in the performance
 $E = |RMSE2| - |RMSE1|$
- 11: **if** $E < \text{threshold}$ **then**
- 12: remove f_i
- 13: **else**
- 14: keep f_i
- 15: $M_{best} \leftarrow M_{best} + f_i$
- 16: **end if**
- 17: **end for**
- 18: Phase II: Modeling RFR
- 19: Design RFR utilizing the n features
- 20: Calculate RMSE1 (RMSE for full model performance)
- 21: **for** $i \leftarrow 1, n$ **do**
- 22: Eliminating least importance feature f_i
- 23: Measuring impurity of the n features (variance reduction)
 $\text{var}((f_1, f_2, \dots, f_n)) = \sum_{i=1}^n \frac{|f_i|}{|M(F_1, F_2, \dots, F_n)|} \text{var}(f_i)$
- 24: Calculating prediction error RMSE2 from the new dataset $M(F_{nn})$
- 25: Calculating the drop in the performance
 $E = |RMSE2| - |RMSE1|$
- 26: **if** $E < \text{threshold}$ **then**
- 27: remove f_i
- 28: **else**
- 29: keep f_i
- 30: $M_{best} \leftarrow M_{best} + f_i$
- 31: **end if**
- 32: **end for**

We tested the five subsets and concluded that the subset of CGHI, RH, DNI, and DHI is ranked most efficient. The

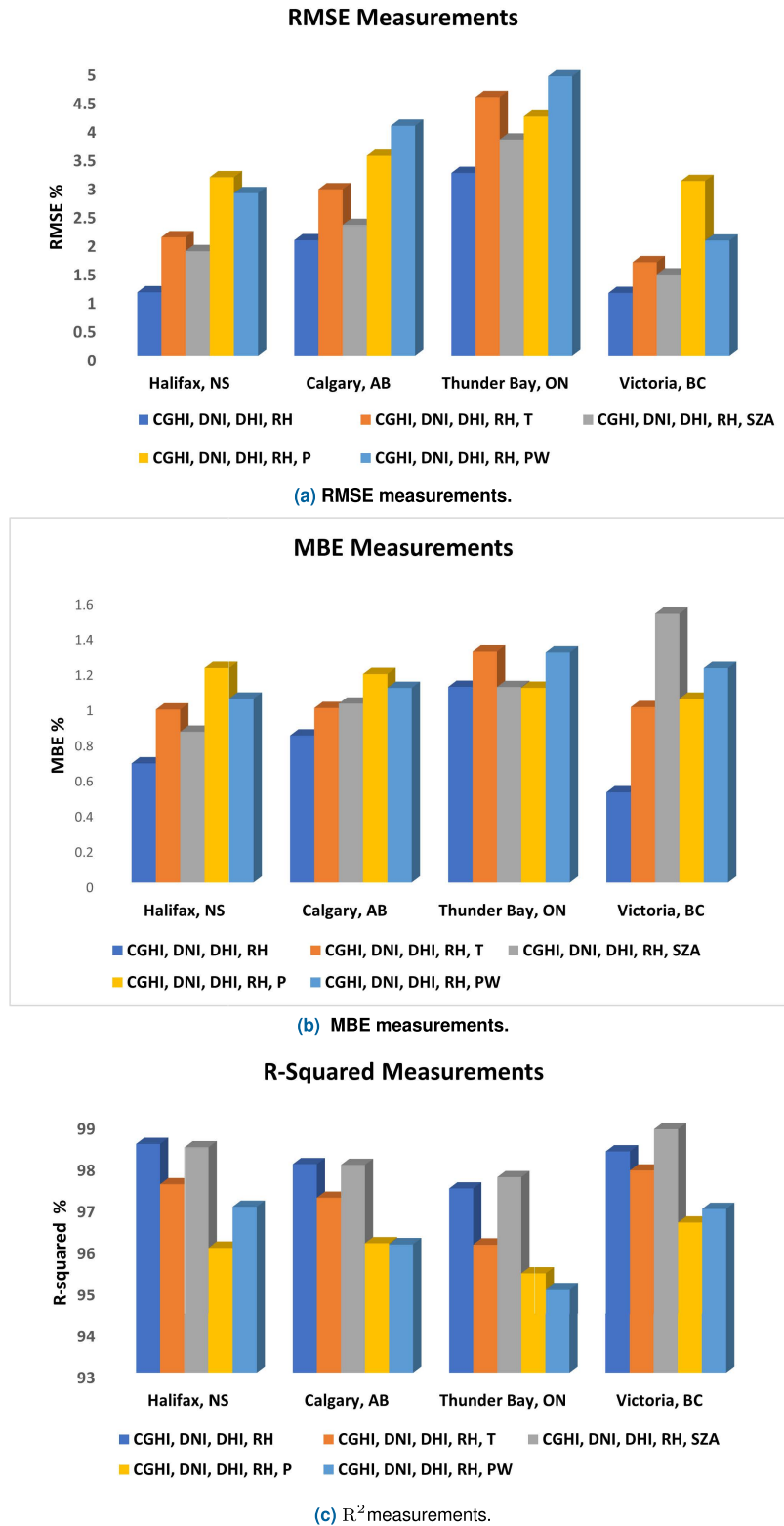


FIGURE 8. Performance comparison of proposed WRFE with several sets of input features for different locations in Canada.

results are not contradicting the findings of the proposed WRFE, as shown in Tables 14 and 15. The result of the

WRFE for the optimal subset is verified by Gray’s relational analysis.

TABLE 15. Evaluation of proposed WRFE using Gray’s relational analysis.

Input feature set	Grey Relational Coefficient			Grey Relational grade	Rank
	RMSE	MBE	R2		
CGHI, DNI, DHI, RH	1	1	0.333333	0.777778	1
CGHI, DNI, DHI, RH, T	0.510985	0.469427	0.833277	0.604563	3
CGHI, DNI, DHI, RH, SZA	0.583049	0.60085	1	0.727966	2
CGHI, DNI, DHI, RH, P	0.333333	0.333333	0.648231	0.438299	5
CGHI, DNI, DHI, RH, PW	0.372035	0.423817	0.756535	0.517463	4

TABLE 16. Performance evaluation of proposed WRFE vs other feature selection approaches.

Ref.	Location	Feature selection Technique	Optimal features	Forecasting Model	RMSE(%)
Ref[30]	Turkey	VIF analysis	$M, R_a, T_{mean}, RH_{mean}$	ANN,ANFIS MLR,Bahel equation	1.650
Ref[31]	southeast of Spain	ARD	DOY, K_t^{ref}	ANN	5.2
Ref[32]	Argentina	NGA	temperature,humidity pressure,sunshine hours	LR	2.3
Ref[33]	India	Pearson correlation Subsets Evaluator	temperature humidity,pressure sunshine hour, wind speed	HMM with GFM	7.9124
Ref[34]	south of Algeria	Subsets Evaluator	$H_0, S_0, T_{max}, T_{mean}$	BDT, ANN, LR	4.5233
Proposed WRFE	Western Canada	WRFE	CGHI, DNI, DHI, RH	LSTM	1.0927

B. COMPARISON OF DIFFERENT FEATURE SELECTION TECHNIQUES

In this section, we compare the performance of the proposed WRFE with other feature selection methods, including the Variance Inflation Factors (VIF) analysis proposed in [30], the Automatic Relevance Determination (ARD) method proposed in [31], the Niching Genetic Algorithms (NGA) proposed in [32], the Pearson correlation analysis followed by the subset evaluator proposed in [33], and the subset evaluator proposed in [34]. ANN and fuzzy logic models are used in [30]. Specifically, ANN, ANFIS, MLR models, and four empirical equations are applied to estimate solar radiation in Turkey. The meteorological data (month number, extraterrestrial radiation, air temperature, relative humidity, sunshine duration, and daylight hours) were measured in 163 stations over 20 years by the Turkish State Meteorological Service (MGM). To determine the multi-collinearity of independent variables, VIF was used. The results indicate that R_a, M, T_{mean} , and RH_{mean} can be powerful input features when estimating solar irradiance in ANN, ANFIS, and MLR models. As well, we dissect and compare different

input variable combinations using MAE, MARE, RMSE, overall index of model performance (OI), and R^2 . We found that ANN outperforms the ANFIS and MLR models and the empirical equations estimating Turkey’s solar irradiance when RMSE is at 1.65 %. The accuracy of ANNs are supported by [31], where ARD is used to select network inputs. The dataset features 36 months of global radiation data (daily) measured at twelve stations in Spain. The authors aimed to estimate daily global irradiation for complex terrain. Estimated values from the ANN model were compared to measured ones, giving an MBE of 0.2 % and an RMSE of 6.0 %. The daily clearness index and DOY are also proven to be relevant input variables. Individual station performance was around [5.0–7.5] %. To further validate the model, it could be applied to other topographically complex areas. The authors in [32] propose solving the variable selection problem by using two applications of NGA to estimate solar radiation. This strategy selects relevant input variables by employing different parameters of genetic algorithm. The technique estimated daily Global Solar Radiation in northern Argentina by applying linear regression

to data obtained from 14 weather stations. From an average of 64 of 329 initial variables, the results show an R^2 of 0.926 and an RMSE of $2.36 \text{ MJ}/\text{m}^2$, using sunshine hours and the most relevant variables (pressure, humidity, temperature). In [33], the authors propose combining GFM and continuous density HMM to estimate solar radiation based on meteorological data from 2009-2011. From the total 915 days, data from the first 750 days is used to training the novel paradigm, while the remaining data is used to validate the proposed model. After analyzing estimations from 15 meteorological parameter combinations, the authors found sunshine duration to be the main parameter in solar radiation estimation, followed by temperature, relative humidity, atmospheric pressure and wind speed. The R-value and RMSE for the best performing meteorological parameter combinations in the framework are 0.9921 and 7.9124 %, respectively. Conflicting experimental outcomes have prompted a shift to reconsider ANNs' usefulness. In [34], several authors compare various solar radiation prediction models, including BDT, ANN, and combinations of these models using LR. The aim is to test predictions for daily global solar irradiation, with performance being validated by a dataset from Algeria's Applied Research Unit for Renewable Energies. The dataset includes global solar radiation, sunshine and air temperature and sunshine duration during 2014-2016. The authors analysed a range of input combinations to find the most relevant input parameters to include in their predictive models. Of the tested parameters, maximum sunshine duration was found to best improve the models' performance. Further, they achieved the best prediction output using input features that included H_0 , S_0 , T_{max} , and T_{mean} , since errors occurring between predicted and measured values are generally quite small. With regard to statistical indicators like RMSE, rRMSE, R^2 , nMBE, MAE and nMAE, the ANN model was shown to perform the best of all the models (e.g., LR, BDT, and hybrid LR-MLP and LR-BDT), achieving a high accuracy of RMSE = 4.5233 %. Regions that have different climate conditions than those tested could be the focus of future work. As seen in Table 16, the proposed WRFE with CGHI, DNI, DHI, and RH as input features yields the lowest RMSE values. The proposed forecasting approach shows lower forecasting errors than the other methods, even with highly fluctuating solar irradiance profiles. However, there is a slight deterioration in the LSTM model performance results obtained using the training dataset for regions with different climate conditions. The ability of enhancing the usage in RMSE and MBE alone will not be a proper indicator of the model's performance. Hence, the t-statistics criteria usage should be in place with these two indicators to receive a proper evaluation of the model's performance [35]. As shown in Table 14, the performance of the models shows a verified result where the t-statistic values (obtained from Equation 12) of the four models are less than the critical t-values.

$$t = \left[\frac{(N-1)MBE^2}{RMSE^2 - MBE^2} \right]^{\frac{1}{2}} \quad (12)$$

IV. CONCLUSION

To date, interactions and nonlinearities that potentially exist between variables are not yet fully researched in the literature. Even so, they represent critical elements for developing robust predictive models. In this work, we focused on redundancy and relevancy, investigating how these can be mitigated and enhanced, respectively, to develop a more robust forecasting model for hourly solar irradiance. Monotonic and non-monotonic associations were probed by applying Spearman's rho and Hoeffding's D correlation analysis in combined usage for locating groups that have major non-monotonic endogenous variable changes. We found that while variables might be non-predictive in isolation, they can be highly predictive in combination with others. This finding led us to perform a subset evaluation to determine relevance using the proposed novel hybrid feature reduction method, Weather Recursive Feature Elimination (WRFE). Our aim was to optimize feature selection according to a Least-Redundant/Highest-Relevant framework, with feature importance measuring RFR impurity and LSTM data perturbation. The simulation results of hourly predictions for GHI demonstrate that the resulting optimal features of the training subset make the greatest contributions to the prediction target. Overall, the outcomes of these investigations indicate the superiority of the proposed WRFE method when compared to other developed models with regard to RMSE. In addition, our study shows that the high variability of irradiance conditions lowers the reliability of the training subset, as most deviations in the correlation coefficients occurred in the yearly dataset. This may warrant further investigation.

REFERENCES

- [1] N. Omar, H. Aly, and T. Little, "LSTM and RBFNN based univariate and multivariate forecasting of day-ahead solar irradiance for Atlantic region in Canada and Mediterranean region in Libya," in *Proc. 4th Int. Conf. Energy, Electr. Power Eng. (CEEPE)*, Apr. 2021, pp. 1130–1135.
- [2] M. Husein and I.-Y. Chung, "Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach," *Energies*, vol. 12, no. 10, p. 1856, 2019.
- [3] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, Apr. 2018.
- [4] Q. Ashfaq, A. Ulasyar, H. S. Zad, A. Khattak, and K. Imran, "Hour-ahead global horizontal irradiance forecasting using long short term memory network," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, Nov. 2020, pp. 1–6.
- [5] X. Huang, J. Shi, B. Gao, Y. Tai, Z. Chen, and J. Zhang, "Forecasting hourly solar irradiance using hybrid wavelet transformation and Elman model in smart grid," *IEEE Access*, vol. 7, pp. 139909–139923, 2019.
- [6] H. S. Dhiman, D. Deb, and J. M. Guerrero, "Hybrid machine intelligent SVR variants for wind forecasting and ramp events," *Renew. Sustain. Energy Rev.*, vol. 108, pp. 369–379, Jul. 2019.
- [7] H. S. Dhiman and D. Deb, "Machine intelligent and deep learning techniques for large training data in short-term wind speed and ramp event forecasting," *Int. Trans. Electr. Energy Syst.*, vol. 31, no. 9, p. e12818, 2021.
- [8] H. S. Dhiman, D. Deb, S. Mueen, and A. Abraham, "Machine intelligent forecasting based penalty cost minimization in hybrid wind-battery farms," *Int. Trans. Electr. Energy Syst.*, vol. 31, no. 9, p. e13010, 2021.
- [9] J. Heng, J. Wang, L. Xiao, and H. Lu, "Research and application of a combined model based on frequent pattern growth algorithm and multi-objective optimization for solar radiation forecasting," *Appl. Energy*, vol. 208, pp. 845–866, Dec. 2017.

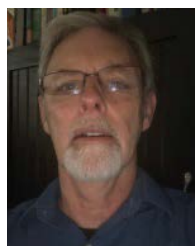
- [10] S. de Siqueira Santos, D. Y. Takahashi, A. Nakata, and A. Fujita, "A comparative study of statistical methods used to identify dependencies between gene expression signals," *Briefings Bioinf.*, vol. 15, no. 6, pp. 906–918, Nov. 2014.
- [11] W. Hoeffding, "A non-parametric test of independence," *Ann. Math. Statist.*, vol. 19, no. 4, pp. 546–557, Dec. 1948.
- [12] *National Renewable Energy Laboratory*, Nrel Nat. Radiat. Database, Washington, DC, USA, 2020.
- [13] H.-Y. Kim, "Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis," *Restorative Dentistry Endodontics*, vol. 38, no. 1, pp. 52–54, 2013.
- [14] *Sas Ondemand for Academics*. Accessed: 2021. [Online]. Available: <https://support.sas.com/ondemand/manuals/SASstudio.pdf>
- [15] A. Yadav and N. Kumar, "Solar resource estimation based on correlation matrix response for Indian geographical cities," *Int. J. Renew. Energy Res.*, vol. 6, no. 2, pp. 695–701, 2016.
- [16] M. J. Patetta. (2001). *Predictive Modeling Using Logistic Regression: Course Notes*. [Online]. Available: <https://support.sas.com/ondemand/manuals/SASstudio.pdf>
- [17] A. R. Gilpin, "Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis," *Educ. Psychol. Meas.*, vol. 53, no. 1, pp. 87–92, Mar. 1993.
- [18] C. Croux and C. Dehon, "Influence functions of the Spearman and Kendall correlation measures," *Stat. Methods Appl.*, vol. 19, no. 4, pp. 497–515, Nov. 2010.
- [19] H. Liu, F. Han, and C.-H. Zhang, "Transelliptical graphical models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 800–808.
- [20] D. Christensen, "Fast algorithms for the calculation of Kendall's τ ," *Comput. Statist.*, vol. 20, no. 1, pp. 51–62, 2005.
- [21] M. T. Cover and A. Joy Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Hoboken, NJ, USA: Wiley, 2006.
- [22] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.
- [23] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 99–106.
- [24] B. Jia, Z. Pan, Z. Hu, J. Pan, and D. Manocha, "Cloth manipulation using random-forest-based imitation learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 2086–2093, Apr. 2019.
- [25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Jun. 2011, pp. 1297–1304.
- [26] I. Kamwa, S. R. Samantaray, and G. Joós, "On the accuracy versus transparency trade-off of data-mining models for fast-response PMU-based catastrophe predictors," *IEEE Trans. Smart Grid*, vol. 3, no. 1, pp. 152–161, Mar. 2011.
- [27] K. Bandara, C. Bergmeir, and H. Hewamalage, "LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1586–1599, Apr. 2020.
- [28] D. Chen, J. Zhang, and S. Jiang, "Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and LSTM neural networks," *IEEE Access*, vol. 8, pp. 91181–91187, 2020.
- [29] J. L. Deng, "Introduction to grey system theory," *J. Grey Syst.*, vol. 1, no. 1, pp. 1–24, 1989.
- [30] H. Citakoglu, "Comparison of artificial intelligence techniques via empirical equations for prediction of solar radiation," *Comput. Electron. Agricult.*, vol. 118, pp. 28–37, Oct. 2015.
- [31] J. L. Bosch, G. López, and F. J. Batlles, "Daily solar irradiation estimation over a mountainous area using artificial neural networks," *Renew. Energy*, vol. 33, no. 7, pp. 1622–1628, Jul. 2008.
- [32] A. Will, J. Bustos, M. Bocco, J. Gotay, and C. Lamelas, "On the use of niching genetic algorithms for variable selection in solar radiation estimation," *Renew. Energy*, vol. 50, pp. 168–176, Feb. 2013.
- [33] S. Bhardwaj, V. Sharma, S. Srivastava, O. S. Sastry, B. Bandyopadhyay, S. S. Chandel, and J. R. P. Gupta, "Estimation of solar radiation using a combination of hidden Markov model and generalized fuzzy model," *Sol. Energy*, vol. 93, pp. 43–54, Jul. 2013.
- [34] A. Rabehi, M. Guermoui, and D. Lalmi, "Hybrid models for global solar radiation prediction: A case study," *Int. J. Ambient Energy*, vol. 41, no. 1, pp. 31–40, Jan. 2020.
- [35] M. M. Khan, M. J. Ahmad, and B. Jamil, "Development of models for the estimation of global solar radiation over selected stations in India," in *Energy, Transportation and Global Warming*. Springer, 2016, pp. 149–160.



NAJIYA OMAR (Graduate Student Member, IEEE) received the B.Sc. degree (Hons.) in electrical and computer engineering from Sirte University, Libya, and the M.A.Sc. degree in pattern recognition and feature selection techniques from Dalhousie University, Canada, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. Her research interests include artificial computation, photovoltaic forecasting, fault detection, big data analytic, optimization theory, and smart cities.



HAMED ALY (Senior Member, IEEE) received the B.Eng. and M.A.Sc. degrees (Hons.) in electrical engineering from Zagazig University, Egypt, in 1999 and 2005, respectively, and the Ph.D. degree from Dalhousie University, Canada, in 2012. He worked as a Postdoctoral Research Associate for one year and an Instructor for three years at Dalhousie University. He worked as an Assistant Professor at Acadia University. He is currently an Adjunct Assistant Professor at Dalhousie University. His research interests include smart grid, applications of artificial intelligence, energy management, and optimization.



TIMOTHY LITTLE is currently the Associate Dean of engineering and a full-time Professor with the Department of Electrical and Computer Engineering, Dalhousie University, Canada. His research interests include engineering education, wind energy and renewable generation, and electromechanical energy conversion.

...