ENHANCED PERFORMANCE OF SOLAR IRRADIANCE PREDICTION USING DEEP LEARNING AND DATA MINING TECHNIQUES

by

Najiya Omar

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
April 2023

To my father who gave me the greatest gift anyone could give another person, he believed in me...

To my late supervisor, Prof. Mohamed El-Hawary, who supported me, encouraged me, and is still guiding me from above...

Table of Contents

List of Tables

List of Figures

Abstract

The limited accessibility of solar irradiance data drives the need for robust Global Horizontal Irradiance (GHI) prediction models. To date, numerous scholars have carried out research looking for ways to enhance the performance of a Long Short-Term Memory (LSTM) model in terms of univariate and multivariate analyses. Although high-dimensional heterogeneous weather data are desirable for enhancing forecasting accuracy, LSTM performance deteriorates when changing from univariate to multivariate analyses. As previous research stops short of conducting detailed explorations on how interactions in high dimensional heterogeneous data represent critical elements in LSTM predictive model development, the present research aims to fill that gap. This work proposes two techniques to enhance predictive performance.

The first technique addresses implementation details regarding relevancy and redundancy measures, exploring how they may, respectively, be enhanced and mitigated. The proposed technique is a novel hybrid feature selection method built to optimize feature-selection using a framework based on Least Redundant/Highest-Relevant, named Weather Recursive Feature Elimination (WRFE). The WRFE approach uses feature importance to measure reductions in variance in Random Forest Regression (RFR) in addition to data perturbation in LSTM. The training set's optimal features demonstrate strong contributions to the prediction outcome, indicating the proposed WRFE's generalizability for hourly GHI prediction. However, high variability in irradiance conditions reduces the overall accuracy of the training subset.

To lessen the seasonality effect, the second proposed technique employs a deep stack of the clustering connected layer with hybrid LSTM models. This novel Seasonal Clustering Forecasting Technique (SCFT) is then compared with other forecasting strategies, revealing its superiority. The SCFT design is further validated using Köppen climate classification data and when measured against the Granger-Newbold and Diebold-Mariano tests. In this as well, the performance of the proposed SCFT shows significant stability and reliability.

List of Abbreviations used

H_O Extraterrestrial Global Solar Radiation.

RH_{mean} Average Relative Humidity.

R^2 Coeficient of Determination.

R_a Extraterrestrial Radiation.

S_O Sunshine Duration.

T_{max} Max Air Temperature.

T_{mean} Mean Air Temperature.

T_{mean} Average Air Temperature.

ρ_C Population Correlation Coeficient.

ρ_P Pearson Correlation.

ρ_S Spearman Correlation.

ANFIS Adaptive Network Fuzzy Inference System.

AR Linear Auto-Regressive Model.

ARD Automatic Relevance Determination.

ARIMA Autoregressive Integrated Moving Average.

ARMA Autoregressive Moving Average.

BDT Boosted Decision Tree.

BPNN Back-Propagation Neural Network.

CGHI Clearsky GHI.

CNN Convolutional Neural Network.

DHI Diffuse Horizontal Irradiance.

DL Deep Learning.

DNI Direct Normal Irradiance.

DP Dew Point.

DRWNN Diagonal Recurrent Wavelet Neural Network.

FFNN Feed Forward Neural Network.

FL Federated Learning.

GBRT Gradient Boosted Regression Trees.

GCPV Grid-Connected Photovoltaic.

GFM Generalized Fuzzy model.

GHI Global Horizontal Irradiance.

HMM Hidden Markov Model.

ISOs Independent System Operators.

LMNN Levenberg Marquardt Neural Network.

LR Linear Regression.

LSTM Long Short-Term Memory.

M Calendar Month Number.

MAE Maximum Absolute Error.

MAPE Mean Absolute Percentage Errors.

MBE Mean Bias Error.

MLP Multi-Layer Perception.

MLR Multiple Linear Regression.

MTSF Multivariate Time-Series Forecasting.

NAR Nonlinear Auto-Regressive Model.

NGA Niching Genetic Algorithm.

nMAE Normalized Mean Absolute Error.

NWP Numerical Weather Prediction.

P Pressure.

PHANN Physical Hybrid Artificial Neural Network.

PW Precipitable water.

RBFNN Radial Basis Function Neural Network.

RENN Recurrent Elman Neural Network.

RFR Random Forest Regression.

RH Relative Humidity.

RMSE Root Mean Square Error.

RNN Recurrent Neural Networks.

SA Surface Albedo.

SCFT Seasonal Clustering Forecasting Technique.

SVM Support Vector Machine.

SZA Solar Zenith Angle.

T Temperature.

UTSF Univariate Time-Series Forecasting.

VIF Variance Inflation Factors.

WD Wind Direction.

WNN Wavelet Neural Network.

WRBPN Wavelet Transformation.

WRFE Weather Recursive Feature Elimination.

WS Wind Speed.

Acknowledgements

I would like to firstly start off by expressing my sincere gratitude to everyone who has helped and supported me throughout my Ph.D. journey. It has been a very dificult yet memorable journey with many ups and downs but overall has shaped me for the future. Without the encouraging words and heartfelt support from supervisors, family, friends, and colleagues, I would not have been able to preserve and continue my research to reach the point I am at today.

I would like to express my sincere thanks to my thesis supervisors, Dr. Timothy Little and Dr. Hamed Aly for stepping in and aiding me to continue and complete my Ph.D. degree. Your supports and guidance were always very comforting. I would never forget to extend a thank you to the members of the respectable committee Dr. Jason Gu and Dr. William Phillips, that have allowed me to clarify any questions that I had.

The loss of my late professor Dr. El-Hawary was a very heavy moment during my Ph.D. journey. His unexpected death hit quite hard both mentally and emotionally. His valuable guidance will forever be missed, as well as the unwavering support he provided me within the beginning steps of the journey. Without his encouraging words to guide me into my Ph.D. studies, I would not be where I am today.

The support I received from the faculty of engineering staff and the department of electrical and computer engineering was immense and was very much needed during the time of grief I faced. Firstly, I would like to acknowledge a very special woman I have met during my studies, Heather Hillyard. Her warm heart and comforting words were always delightful. She never missed checking up on me and if I needed to talk.

Within the department of Electrical and Computer Engineering, I would like to send a huge thanks, but would also like to say a few names: Thank you to the Department Head, Dr. Jean-François Bousquet, the Grad Advisor at the time Dr. Dmitry Trukhachev, and Dr. Hamed Aly, for showing endless guidance during my transition to the new supervision.

Last but not least, I would like to acknowledge my family and friends: My husband

Imhamed, and my children, especially my eldest daughter Mayar for their support and patience as I complete my degree, and their never-ending encouraging words. My parents, although them being thousands of miles away, have never stopped sending their love and their prayers. Lastly but most importantly, I would like to thank God for his guidance throughout these past few years and has never left me lost, confused and overwhelmed when I needed him most.

Chapter 1

Introduction

1.1 Background and Problem Statement

The exponential rate of growth registered recently in PV systems is prompting an increased need for more integration of PV power generation into main grid systems in countries around the world. Additionally, remote microgrids, which typically operate on diesel generators, are being integrated with Grid-Connected Photovoltaic (GCPV) systems to reduce power generation costs. Obtaining full benefits from renewable power generation is functionally achievable only if scheduling and coordination occurs between the grid system and the renewable source. Solar energy is by nature both uncontrollable and intermittent, which means solar sources typically provide varying outputs and peak randomly, bringing a number of challenges to the table. The main ones are: exacerbating issues related to grid management and sustainability of continuous production and consumption balance; power quality, including stability issues of voltage and frequency; and dificulty of power scheduling, and regulation; ensuring a steady supply within a specified range for electricity companies and Independent System Operators (ISOs). The issues related to the integration of PV systems into microgrids could be managed by holding large reserves to guarantee continuous and dependable operations. However, large reserves come with increased costs related to operations, transmission, repairs, and increased fuel consumption, all of which leads, counterintuitively, to the release of more carbon emissions [1].

Considering the aforementioned problems, the research focus in recent years has shifted to developing more accurate PV power output forecasting models. The output power production of the designed PV system could be then estimated based on local solar radiation and weather data. Solar irradiation and ambient temperature have a major effect on three critical parameters of PV panels [2]. Therefore, solar irradiance predictive modeling is crucial to determine the size of PV arrays, design grid-connected PV systems, and eventually develop robust PV power forecasting. An important initial step for PV power prediction systems is GHI forecasting[3][4][5]. In general, GHI forecasting can be classified into the following three categories:

- Forecasting horizons
- Forecasting techniques
- Input features

However, if the researchers or operators lack access to stations that measure solar irradiance and/or lack the financial resources to purchase solar irradiance devices, the necessary data may be unobtainable for them [6] [7]. Several renewable energy stakeholders, including solar power producers, utilities, and Independent System Operators (ISOs), are keen to have highly accurate solar power forecasts. In fact, in some jurisdictions, it is a legal requirement for power producers to provide accurate power forecasts to their customers as a part of their power purchase agreement [8].

Historical data are being applied to Deep Learning (DL) algorithms as a means to determine stochastic dependency between past and future data. These kinds of models have been shown to outperform statistical models with regard to forecasting of solar irradiance. The design workflow of forecasting application involves four main steps: 1) Data pre-processing, which is used to clean, inspect, analyze, and aggregate data sets; 2) feature selection, which involves choosing optimal features in combination, aiming for peak performance in data utilization; 3) technique selection, which involves choosing a best DL algorithm based on predictive performance; and 4) model design, which aims to optimize the model's performance by tuning the hyperparameter values. Figure 1.1 illustrates the four-step design process. Much of the current literature focuses on applying DL techniques to solar power and solar irradiance forecasting and modeling, with the LSTM model emerging as the most prevalent forecaster of solar irradiance [9][10][11][12].

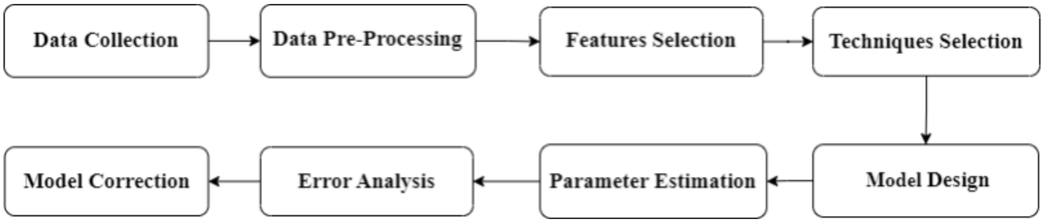


Figure 1.1: Design Process of Forecasting Application

1.2 Objectives and Research Questions

The LSTM model has become more prevalent as a solar irradiance forecaster and its performance is still under investigation, which this research will seek to do. The research is concerned with developing robust models for forecasting hourly solar irradiance. In the first phase of this work, the behavior of LSTM models will be 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) will be utilized as input features for hour-ahead solar irradiance forecasting models. In the work, a comparison is made with regard to Univariate Time-Series Forecasting (UTSF) and Multivariate Time-Series Forecasting (MTSF). Secondly, as a means to better understand relationships between model components, forecasting models typically use numerous inputs of high dimensionality 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 as shown in Figure 1.2.

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

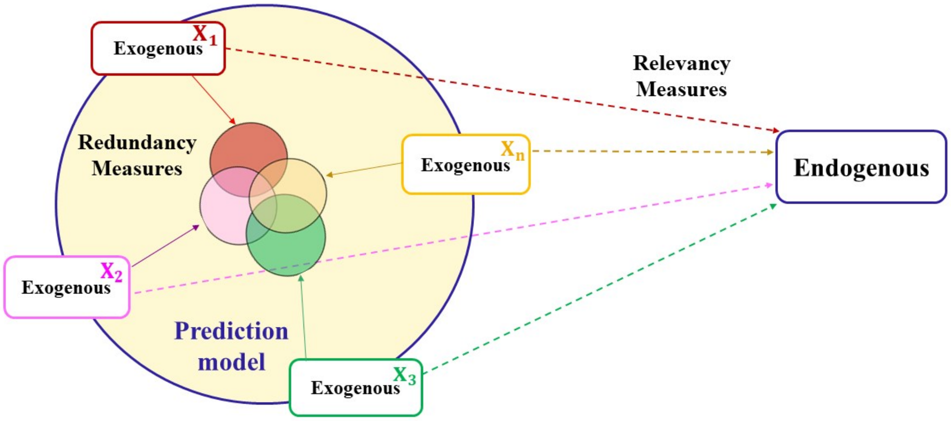


Figure 1.2: Redundancy and Relevancy measures

variables are not yet fully researched in the literature [13][14]. Even so, they represent critical elements for developing robust predictive models. In the second phase, the focus is on the measures and attributes of redundancy and relevancy and will investigate how these can be mitigated and enhanced, respectively, to develop more accurate forecast models. Interestingly, the results of the initial investigations show that the LSTM performance deteriorated when additional features were added. Therefore, a series of questions are presented that the work 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 [13][15][16][17]. 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 coeficient to find linear associations between the exogenous variables. However, it could be argued that linear association is not enough to make a fully informed decision about redundant variables. Therefore, the work 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 coeficient 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?

It is noticed that when a univariate LSTM forecasting model shifted to a multivariate forecasting model, the predictive performance is affected and depends on geographical and meteorological conditions [18]. Thus, considering seasonality in LSTM model is necessary. More studies indicate that the changes of seasonality require further exploration of seasonality patterns in the weather data through LSTM models which are capable of obtaining nonlinearity patterns rooted in the exogenous and endogenous variables[19] [20]. Eficiency wise, comparing LSTM model to DL techniques in the domain of solar irradiance forecasting proved the ability of LSTM model to learn from nonlinearity behavior in solar irradiance data with a long range of temporal dependencies [9][10][11][12]. Second, the studies suggested that LSTM model should be a potential approach for the computational complexity of multivariate prediction due to the different weather phenomena. Further, the literature mentions that LSTM-based forecasting accuracy could be enhanced through the application of large datasets, which have been shown to boost decision-making capabilities. High-dimensional heterogeneous data may, however, fall prey to issues concerning data quality. To overcome these, the process of data mining may be applied to look for patterns, trends, anomalies, and correlations in large datasets. In particular, a clustering algorithm such as k-means could be used for classifying data point as either rainy, cloudy or sunny clusters, as a way to decrease uncertainty in solar irradiance forecasting.

1.3 Contribution

The main contributions of the research are published in the following conference and journal papers:

- The research paper [18] investigated the behavior of LSTM models under variety of geographical and meteorological conditions and according to historical solar irradiance and meteorological data.
- The research paper [21] analysed the stability performance of the correlation analysis for a one-year dataset and a ten-year timeframe.
- The research paper [21] proposed the novel WRFE method for optimizing feature selection schema according to a Least-Redundant/Highest- Relevant framework..
- The research paper [22] investigated seasonality-based hourly predictions of GHI and performing experiments utilizing seasonal forecasting 3D LSTM models that have been developed for pattern-recognition of the four seasons.
- The research paper [22] proposed a deep stack of the clustering connected layer with hybrid LSTM models to improve accuracy in forecasting. The result is the proposed SCFT.
- The research paper [22] validated the generalizability of the proposed SCFT by using it in regions that feature different climatic conditions than the original test region.

1.4 Publications Associated with the Thesis

The work explained in this chapter contains materials that were published at the following conference and journal publications that are associated with this thesis:

- Seasonal Clustering Forecasting Technique for Intelligent Hourly Solar Irradiance Systems. IEEE Transactions on Industrial Informatics, June 4, 2022 [22].
- Optimized Feature Selection Based on a Least-Redundant and Highest-Relevant Framework for a Solar Irradiance Forecasting Model. IEEE Access, May 13, 2022 [21].
- LSTM and RBFNN Based Univariate and Multivariate Forecasting of Day-ahead Solar Irradiance for Atlantic Region in Canada and Mediterranean Region in Libya. International Conference on Energy, Electrical and Power Engineering (CEEPE-IEEE), Chongqing, China, April 1, 2021 [18].
- Grid-Connected Photovoltaic System: System Overview and Sizing Principles. International Journal of Electrical and Computer Engineering (IJECER), December 2020 [2].

1.5 Thesis Organization

- Chapter 2 includes the research that has been conducted already on Solar irradiance forecasting systems, and enhancements that have been achieved in stages of feature and technique selection, as well as the recommendations that have been stated for solar irradiance forecasting application.
- Chapter 3 presents workflow of research methodology, explains data collection and sources, and performs a comprehensive exploratory data analysis
- Chapter 4 includes the investigation of the behavior of LSTM in references to the input feature constructure, presents the implementation and design of the proposed models WRFE and SCFT, discuses the results of the proposed models, and performs confirmatory data analysis for validation.
- Chapter 5 is the conclusion, limitations, and directions for future works

Chapter 2

Literature Review

The focus is to explore the important aspects of the existing body of literature and narrow the conducted research in terms of the forecasting techniques and forecasting horizon. DL algorithms use historical data to learn the stochastic dependency between the past and the future. Much research has been concluded and stated that these models outperform statistical models in solar irradiance forecasting. Therefore, the application of DL techniques for modeling and forecasting of day-ahead solar irradiance is targeted in my research. Due to the varying assumptions and range of inputs in DL forecasting techniques, it is dificult to compare to extant models due to the variability in the ways that they have been studied. Regardless, a review of the past and current literature is important for understanding the most useful models and the benefits of LSTM over previously used models. GHI forecasting approaches may be categorized according to the input data, the forecasting techniques, and the forecasting horizons.

2.1 Algorithm Selection

Several different GHI forecasting approaches have been proposed in the relevant literature. These typically use geographical location, data quality, and weather conditions as references. As such, they can be easily categorized as either statistical, tradi-tional Machine Learning, and physical methods such as ANN, Autoregressive Inte-grated Moving Average (ARIMA), Autoregressive Moving Average (ARMA), Feder-ated Learning (FL), Numerical Weather Prediction (NWP), with each one adopting its own forecasting strategy, as illustrated in Figure 2.1.

2.1.1 MLP Model-based Solar Irradiance Forecasting

Early models, originally developed by [6], used Back-Propagation Neural Network (BPNN) to predict solar radiation as a function of available weather data and other

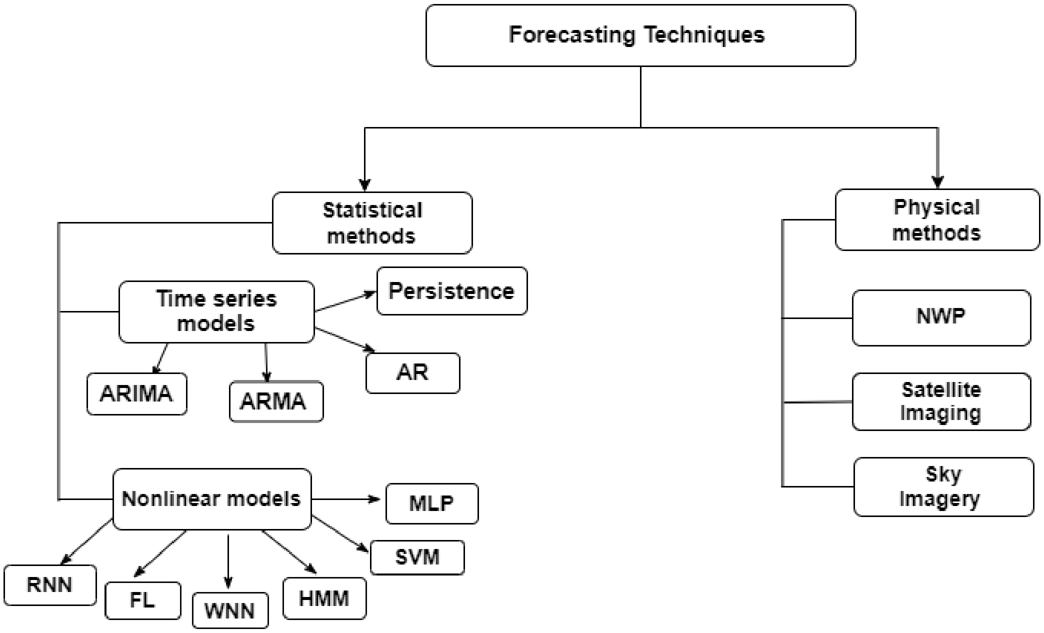


Figure 2.1: Forecasting Techniques

environmental variables. A collective total of 23 years of weather data sets were available from four sites in the southeastern USA, and these data sets were separated into 11 years for the training data set and 12 years for the testing data set. These weather data included daily minimum air temperature, precipitation, daily clear sky radiation, and daylength. The predicted daily solar radiation values were compared with the observed daily solar radiation values for these 12 years. The performance of the model was evaluated based on Root Mean Square Error (RMSE) and the Coeficient of Determination (R^2) between predicted and observed solar radiation of the test data. The performance for each yearly testing data varied from 2.29 to 3.64 for RMSE and 0.52 to 0.74 for R^2. They recommended that future research is required to validate this approach at locations with higher latitudes.

Other research on BPNN by [23] proposed a BPNN for modeling monthly mean daily values of global solar radiation from 41 data collection stations in the kingdom of Saudi Arabia, divided into 31 neural network training locations and 10 testing locations. The proposed model utilized latitude, longitude, altitude, and the sunshine duration for the prediction of solar radiation values. The results of the proposed model indicated relatively good performance between the predicted values and the observed ones in terms of the Mean Absolute Percentage Errors (MAPE) around 19.1%.

Research on BPNN has also been extended by [24]. They estimated global radiation values for different locations in the Sultanate of Oman. Meteorological data from six weather stations in the Sultanate of Oman were obtained for the years 1987 to 1992. Eight input features entered into the BPNN including the location, month, mean temperature, mean pressure, mean relative humidity, mean wind speed, mean duration of sunshine, mean evaporation. The results demonstrated that the proposed BPNN-based model can estimate the global radiation value for the given data set with an accuracy of 93%. To further evaluate the generalization capability of the proposed model, they developed a model to estimate global radiation based on historical data from the prior 12 months for a location in the Sultanate of Oman that has global radiation measurement instrumentation where the model achieved a prediction accuracy of 95%. The model was subsequently used to predict global radiation data for a different location in the Sultanate of Oman where no direct measurement instrumentation for global radiation was available.

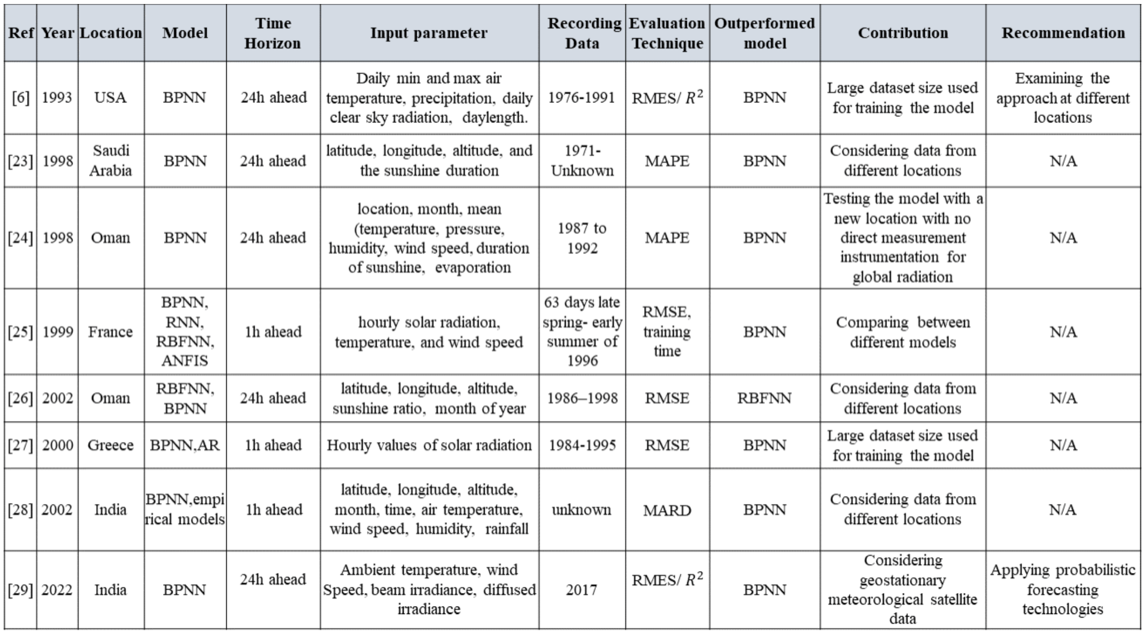
Though prior research has therefore demonstrated some support for the usefulness of BPNN, later work by [25] introduced a comparative study between BPNN and Levenberg Marquardt Neural Network (LMNN), Recurrent Elman Neural Network (RENN), and Radial Basis Function Neural Network (RBFNN) alongside the Adaptive Network Fuzzy Inference System (ANFIS) for the forecasting of mean hourly global solar radiation. The data used throughout the study are mean hourly solar radiation values on a horizontal level, in W/m^2, measured on the French island of Corsica. The data cover a period of 63 days in total (i.e. 1512 h) during late spring and early summer of 1996. A comparison between the various models in terms of RMSE and training time indicated that the LMNN model outperforms other techniques for predicting hourly global solar radiation.

Despite the apparent usefulness of the LMNN model, further research on BPNN demonstrated mixed results. For example, [26] discuss DL techniques for estimating solar radiation by first estimating the clearness index which is the ratio of the average daily solar radiation, and the daily maximum radiation. RBFNN and BPNN models were investigated using long-term data from eight stations in Oman over ten years (1986–1998). The input parameters were latitude, longitude, altitude, sunshine ratio, and month of the year. The output parameter is the clearness index. The estimated solar radiation was obtained by multiplying the estimated clearness index by the daily maximum radiation. The authors demonstrated that both the RBF and MLP models performed well based on RMSE between the observed and estimated solar radiations. However, the RBF model is favored since it requires less computing time.

Other research on BPNN by [27] proposed a BPNN and a Linear Auto-Regressive Model (AR) for solar radiation forecasting. From the Institute of Meteorology and Physics of the Atmospheric Environment of the National Observatory in Greece, hourly values of solar radiation for twelve years (1984-1995) for various months of the year were used for training and testing the network. Nine years (1984-1992) were used for training the neural network and three years (1993, 1994 and 1995) for testing process. Their forecasting models were able to simulate the future values of total solar radiation time series based on their past values. The results show that BPNN approach leads to better predictions than the AR model with RMSE between the measured and the estimated values of 4.9% lower than the AR model.

Given these conflicting results in the eficacy of the BPNN model, other research on BPNN has considered alternative models. [28], for example, presented a BPNN model for estimation of hourly values of solar global radiation. Solar radiation data from 13 stations throughout India around the year were used for training and testing the model. The solar radiation data from 11 locations – six from South India and five from North India – were used for training the BPNN and data from the remaining two locations – one each from South India and North India – were used for testing the estimated values. The nine input parameters were considered to estimate the radiation for each city, including latitude, longitude, altitude, month, time, air temperature, wind speed, relative humidity, and rainfall. The authors note that to improve the performance of the network, it was necessary to divide the data. The entire training set was divided based on region (South India and North India) and seasons (summer, rainy and winter). The results of the BPNN model were compared with other empirical regression models proposed in the literature. The solar radiation estimations by BPNN were superior to the other models, with Maximum Absolute Error (MAE) values of 0.028, 0.06, and 0.032 W/m^2 for the summer, winter, and rainy seasons, respectively. Future works of applying probabilistic forecasting technologies are recommended by [29]. Table 2.1 is a summary of several research works

Table 2.1: MLP Model-based Solar Irradiance Forecasting



Ref	Year	Location	Model	Time Horizon	Input parameter	Recording Data	Evaluation Technique	Outperformed model	Contribution	Recommendation
[6]	1993	USA	BPNN	24h ahead	Daily min and max air temperature, precipitation, daily clear sky radiation, daylength.	1976-1991	RMES/ R^2	BPNN	Large dataset size used for training the model	Examining the approach at different locations
[23]	1998	Saudi Arabia	BPNN	24h ahead	latitude, longitude, altitude, and the sunshine duration	1971-Unknown	MAPE	BPNN	Considering data from different locations	N/A
[24]	1998	Oman	BPNN	24h ahead	location, month, mean (temperature, pressure, humidity, wind speed, duration of sunshine, evaporation	1987 to 1992	MAPE	BPNN	Testing the model with a new location with no direct measurement instrumentation for global radiation	N/A
[25]	1999	France	BPNN, RNN, RBFNN, ANFIS	1h ahead	hourly solar radiation, temperature, and wind speed	63 days late spring- early summer of 1996	RMSE, training time	BPNN	Comparing between different models	N/A
[26]	2002	Oman	RBFNN, BPNN	24h ahead	latitude, longitude, altitude, sunshine ratio, month of year	1986–1998	RMSE	RBFNN	Considering data from different locations	N/A
[27]	2000	Greece	BPNN,AR	1h ahead	Hourly values of solar radiation	1984-1995	RMSE	BPNN	Large dataset size used for training the model	N/A
[28]	2002	India	BPNN,empirical models	1h ahead	latitude, longitude, altitude, month, time, air temperature, wind speed, humidity, rainfall	unknown	MARD	BPNN	Considering data from different locations	N/A
[29]	2022	India	BPNN	24h ahead	Ambient temperature, wind Speed, beam irradiance, diffused irradiance	2017	RMES/ R^2	BPNN	Considering geostationary meteorological satellite data	Applying probabilistic forecasting technologies

on developing Solar Irradiance Forecasting based on MLP Models.

2.1.2 Hybrid Model-based Solar Irradiance Forecasting

This work supports prior findings by [30] who proposed a multistage BPNN to predict daily solar radiation. Meteorological data at Omaezaki, Japan in 1988–1993 were used as input data to forecast solar irradiance in 1994. In the first stage, meteorological data of the atmospheric pressure of the previous day were adopted as input data of BPNN and the average atmospheric pressure was forecasted as the output. The second stage was aimed to forecast the irradiance level of the next day, where the forecasted average atmospheric pressure of the next day, as well was meteorological data of the atmospheric pressure of the previous day, were used as inputs for the model. The irradiance level is divided into three classes based upon clearness index. In the final stage, three models of BPNN were designed depending on the high,middle, and low irradiance level. Seven meteorological data from the previous day were input to each of the three networks, and the irradiance of the next day was forecasted in accordance with the irradiance level. Irradiance forecasts by the multi-stage and single-stage neural networks were compared with measured irradiance. The results show that the Mean Bias Error (MBE) by the multi-stage BPNN was about 20% while that by the single-stage BPNN about 30%.

Though other models have become prevalent more recently, the use of BPNN remained in the literature until as recently as 2010, [31] developed a BPNN to forecast the daily solar irradiance. The proposed model accepts mean daily irradiance, mean daily air temperature, and the day of the month as input parameters, outputting a day ahead solar irradiance. Historical data for solar irradiance and air temperature from July 2008 to May 2009 and from November 2009 to January 2010 has been collected in Trieste, Italy. The measurements of prediction performance were based on RMSE and MBE. The results indicate that the proposed model performs well, while the correlation coeficient is in the range 98–99% for sunny days and 94–96% for cloudy days. As a means of validating the system, the authors compare the results of the forecasted solar irradiance and the energy produced by the GCPV plant installed on the local building in Trieste shows the goodness of the proposed model.

Indeed, a significant amount of work has focused on similar models and comparisons. Work by [32], for example, used BPNN models for modelling solar resources in Turkey. Meteorological data for the prior three years (2000–2002) from 17 stations throughout Turkey were used as training and testing data. Meteorological and geographical data including latitude, longitude, altitude, month, mean sunshine duration and mean temperature served as the input while the output was solar radiation. The authors examined the proposed model with different numbers of hidden layers and neuron numbers of the hidden layers using four different training algorithms in the proposed forecasting model including the SCG, the Pola–Ribiere conjugate gradient (CGP) and the LM algorithms. The results indicate that the RMSE value is around 6.73% compared to [30] who found values closer to 12.5% and [25] who found values of 19.1%.

Despite early research suggesting the usefulness of BPNN, there have been competing models that have also demonstrated promise for the prediction of solar radiation. For example, [33] utilized RBFNN technique for the estimation of monthly mean daily values of solar radiation, comparing performance to the BPNN, a classical regression model, and Angstrom regression. The authors used solar radiation data similar to that applied in [23]. The MAPE for 10 locations in the kingdom of Saudi

Arabia were used for testing the models. The results indicate that RBFNN networks outperformed the performance of BPNN for global solar radiation modeling with the values of MAPE 15.2%, and 19.1% for RBFNN and BPNN, respectively.

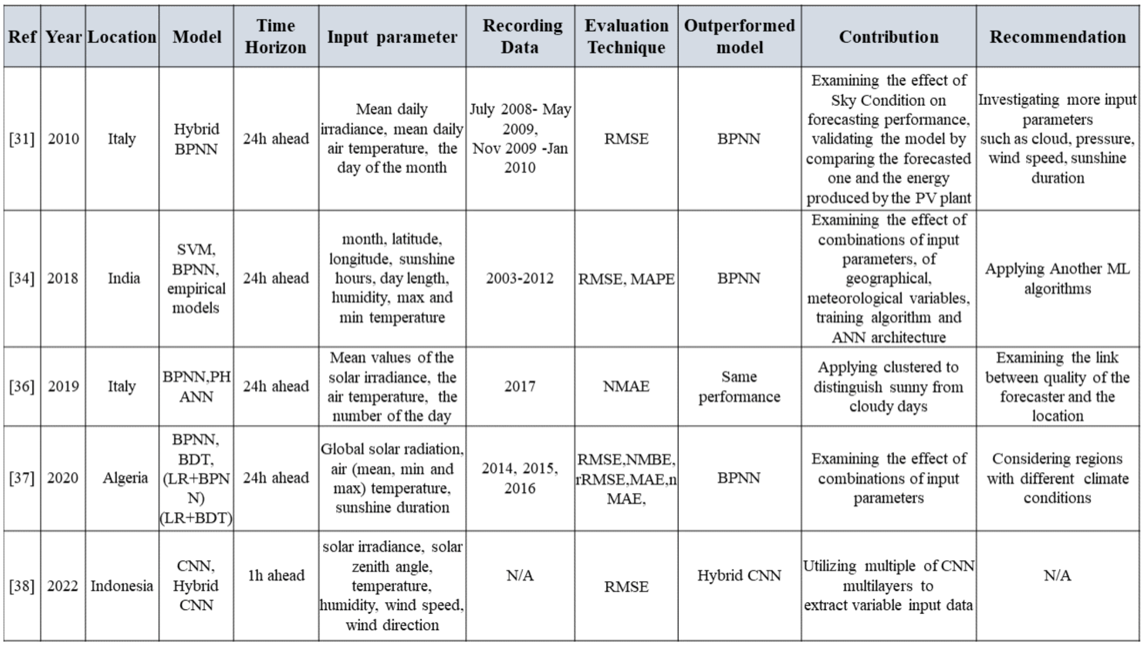
Differences in results, and conflicting findings in support of the use of BPNN, has more recently lead to the use of Artificial Neural Networks. Most notably, [34] evaluated the accuracy of Support Vector Machine (SVM), Artificial Neural Network (ANN) and empirical models for the estimation of monthly mean daily Global Solar Radiation (GSR) where different combinations of input parameters were examined. Data used in this study was provided by the India Meteorological Department (IMD) in eight stations from 2003- 2012. The parameters include month, latitude, longitude, bright sunshine hours, day length, relative humidity, maximum and minimum temperature. After applying sensitivity analysis, they indicated that the model using bright sunshine hours and day length inputs performs good in radiation prediction with the least RMSE followed by the model with minimum and maximum temperature as inputs. Regression error is further minimized if the geographical parameters are added, with relative humidity as the least influencing input parameter. They found a higher correlation coeficient for ANN (0.9968) and SVM (0.9912) when compared to empirical models. Per recommendations by [34], the MAPE of different ANN models changes with the influence of geographical, meteorological variables, training algorithm and the architecture configuration of ANN. Therefore, the appropriate selection of input parameters is important for predicting solar radiation with better accuracy. Further, to improve the prediction accuracy, future research should focus on global solar radiation prediction using k-NN, regression tree, boosting, and random forest.

A recent review on ANNs across 24 articles by [35] found that ANN models have proven to be a powerful technique for predicting solar radiation in different climatic conditions. This is due to the reliability of ANNs for accepting many input parameters as compared to empirical. In addition, they conclude that ANN-based prediction offers greater accuracy as compared to empirical models. More recent research on the eficacy of ANNs, however, has provided mixed results. First, [36] compared the forecasting performance of BPNN and Physical Hybrid Artificial Neural Network (PHANN) which were trained on the same dataset. The dataset collected at the

SolarTechLab of the Politecnico of Milan, Italy and contained historical hourly solar irradiances including the climatic parameters and PV system parameters covering an entire year, and was clustered to distinguish sunny from cloudy days and separately train the ANN. The available dataset for 2017 was composed of 268 days, and further divided into two sub-datasets, depending on whether the mean daily forecast irradiation was greater or lower than 150 W/m^2 (i.e. Sunny days and Cloudy days). The input of the first NN forecaster were the mean values of the solar irradiance, the air temperature, and the number of the day. The model was developed using 240 days of the original dataset, while 28 days were used for the testing of the model. The output layer has 24 output nodes representing the produced hourly power of the next day. The second forecaster has a hybrid approach, including as an input the daily weather forecast and the Clear Sky Radiation Model (CSRM). This work shows that no one model outperforms the other under all possible conditions with an almost constant Normalized Mean Absolute Error (nMAE).

These conflicting outcomes for ANNs have led to a more recent shift to consider the usefulness of Multi-Layer Perception (MLP), examined by [37]. The authors present a comparison between different prediction models for solar radiation application, MLP, Boosted Decision Tree (BDT), and a new combination of these mod-els with Linear Regression (LR) for the prediction of daily global solar irradiation (DGSR). The performance of the proposed models was validated using a real dataset measured at the Applied Research Unit for Renewable Energies (URAER) located in the south of Algeria. The database contains DGSR, extraterrestrial global solar radiation, (mean, min and max) air temperature and sunshine duration for three years (2014, 2015 and 2016). Different input combinations were analysed to select the relevant input parameters for DGSR prediction. They found that maximum sunshine duration significantly improves the performances of the models. The best prediction output is achieved when the inputs features include global solar radiation, mean air temperature, max air temperature and sunshine duration as the error between the measured and predicted values is relatively small error. The results achieved show that the MLP model preforms better than the other models in terms of statistical indicators such as R^2, RMSE, nMBE, rRMSE, MAE, and nMAE. The comparison results showed that the MLP model achieves high accuracy compared to the LR, BDT

Table 2.2: Hybrid Model-based Solar Irradiance Forecasting



Ref	Year	Location	Model	Time Horizon	Input parameter	Recording Data	Evaluation Technique	Outperformed model	Contribution	Recommendation
[31]	2010	Italy	Hybrid BPNN	24h ahead	Mean daily irradiance, mean daily air temperature, the day of the month	July 2008- May 2009, Nov 2009 -Jan 2010	RMSE	BPNN	Examining the effect of Sky Condition on forecasting performance, validating the model by comparing the forecasted one and the energy produced by the PV plant	Investigating more input parameters such as cloud, pressure, wind speed, sunshine duration
[34]	2018	India	SVM, BPNN, empirical models	24h ahead	month, latitude, longitude, sunshine hours, day length, humidity, max and min temperature	2003-2012	RMSE, MAPE	BPNN	Examining the effect of combinations of input parameters, of geographical, meteorological variables, training algorithm and ANN architecture	Applying Another ML algorithms
[36]	2019	Italy	BPNN,PHANN	24h ahead	Mean values of the solar irradiance, the air temperature, the number of the day	2017	NMAE	Same performance	Applying clustered to distinguish sunny from cloudy days	Examining the link between quality of the forecaster and the location
[37]	2020	Algeria	BPNN, BDT, (LR+BPNN) (LR+BDT)	24h ahead	Global solar radiation, air (mean, min and max) temperature, sunshine duration	2014, 2015, 2016	RMSE,NMBE, rRMSE,MAE,nMAE,	BPNN	Examining the effect of combinations of input parameters	Considering regions with different climate conditions
[38]	2022	Indonesia	CNN, Hybrid CNN	1h ahead	solar irradiance, solar zenith angle, temperature, humidity, wind speed, wind direction	N/A	RMSE	Hybrid CNN	Utilizing multiple of CNN multilayers to extract variable input data	N/A

and hybrid LR-MLP, LR-BDT models. The results proved that the MLP model is a suitable approach for solar radiation forecasting. It is recommended that the effect of applying hybrid model should be examined [38]. Table 2.2 is a summary of several research works on developing Solar Irradiance Forecasting based on Hybrid Models.

2.1.3 Fuzzy logic and empirical Model-based Solar Irradiance Forecasting

In a similar way that questions about BPNN led to the consideration of alternative models, some researchers turned to fuzzy logic models as a possible viable alternative. A review of this work was conducted by [39]. The authors present a review of solar radiation prediction using different ANN techniques. The ANN models are found to outperform the Angstrom, conventional, linear, nonlinear, and fuzzy logic models in predicting solar radiation. The performance of ANN models is improved with the impact of geographical, meteorological variables, training algorithms and ANN architecture configuration. The geographical and meteorological parameters such as sunshine duration, maximum ambient temperature, relative humidity, latitude, longitude, day of the year, daily clear sky global radiation, total cloud cover, temperature, clearness index, altitude, months, average temperature, average cloudiness, average wind velocity, atmospheric pressure, reference clearness index, mean diffuse radiation, mean beam radiation, month, extraterrestrial radiation, evaporation, soil temperature were used as input variables to ANN models for solar radiation prediction. Sunshine hours and air temperature are found to be effective inputs for the model with correlation coeficient of 0.97. The effective input was be selected using Niching Genetic Algorithm (NGA) and Automatic Relevance Determination (ARD) methodology.

One of the first such references to fuzzy logic models was by [40]. The authors present models for global and diffuse solar irradiances for five sites in Malaysia. The global solar irradiance is modeled using linear, nonlinear, fuzzy logic, and ANN models, while the diffuse solar irradiances is modeled using linear, nonlinear, and ANN models, drawing on data for solar irradiations (1975–2004) taken from the five sites in Malaysia. The input parameters included latitude, longitude, day number, and sunshine ratio. Three statistical values were used to evaluate the proposed models based on MAPE, RMSE, MBE. The results showed that the ANN models are superior compared with the other models with 5.38% for the MAPE while the MAPE of 8.13%, 6.93%, and 6.71%, for the linear, nonlinear, and fuzzy logic models, respectively. The results for the diffuse solar energy showed that the MAPE of the ANN model is 1.53%, while the MAPE of the linear and nonlinear models are 4.35% and 3.74%, respectively.

[41] provided support for the eficacy of ANNs and fuzzy logic models. Here, the ANN, the ANFIS, and NGA models, and four empirical equations are applied for estimation of the solar radiation in Turkey. The meteorological data consist of month number, extraterrestrial radiation, average air temperature, average relative humidity, average sunshine duration, and daylight hours. This data with monthly solar irradiance measured by the Turkish State Meteorological Service (MGM) at 163 stations for 20 years are used in in developing the models. Variance Inflation Factors (VIF) was applied to measure the existence of multi-collinearity among independent variables and based on these results, calendar month number, extraterrestrial radiation, average air temperature, and average relative humidity are determined to be

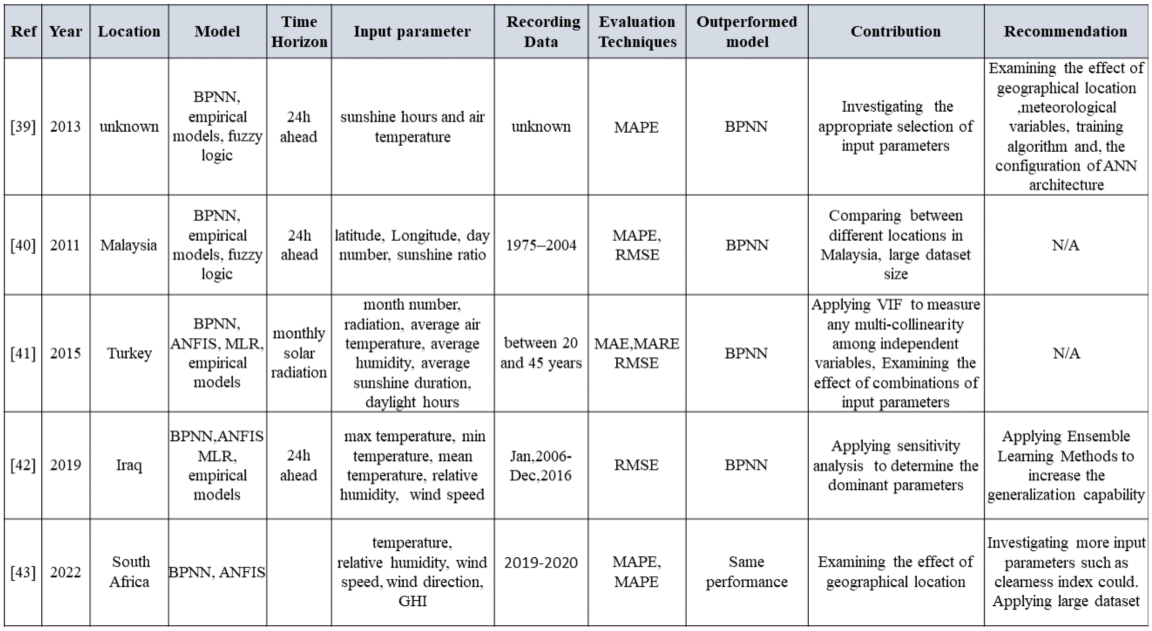
powerful input features for the ANN, ANFIS, and MLP models for estimation of so-lar irradiance. In addition, various combinations of input variables are dissected and compered based on MAE, RMSE, overall index of model performance (OI), and R^2 . The results show that the ANN model performs better than the ANFIS and MLP models and the empirical equations in estimating solar irradiance in Turkey.

Finally, [42] explored two ANN-based models including Feed Forward Neural Network (FFNN) and ANFIS, three temperature-based empirical models including Meza–Varas, Hargreaves–Samani, and Chen, and MLP for daily global solar irradiance prediction in Iraq. For this purpose, daily meteorological data of maximum, minimum and mean temperature, relative humidity, and wind speed were obtained from 2006 to 2016 from four major cities in Iraq representing, north, west, south, and east regions. Sensitivity analysis was conducted to determine the leading features for the forecast. The results showed that of maximum, minimum and mean temperature, and relative humidity are the dominant features. In addition, a comparison between two ensemble approaches, neural average ensemble and simple average ensemble, were applied to improve the performance of the single models. The general idea of the ensemble technique was employed to improve the predictive performance by combining the outputs of the single models. The results of this research indicated that while temperature-based empirical models and MLP model could be employed to achieve reliable results, ANN based models are superior in performance to other models. It is suggested that regions with different climate conditions should be considered in the future investigations[43].Additionally, suggesting improvements in daily global solar irradiance forecast could be achieved by model ensemble as the concept of the ensemble structure is to increase the generalization capability of the models to improve the prediction performance. Table 2.3 is a summary of several research works on developing Solar Irradiance Forecasting based on Fuzzy logic and empirical Models.

2.1.4 RNN Model-based Solar Irradiance Forecasting

The last relevant model to explore before considering current directions in predicting solar irradiation is that of Recurrent Neural Networks (RNN). The first such reference to this model of relevance is that of [44]. This paper presented a forecast-ing solar irradiance model using recurrent back-propagation network (RBPN) and

Table 2.3: Fuzzy logic and empirical Model-based Solar Irradiance Forecasting



Ref	Year	Location	Model	Time Horizon	Input parameter	Recording Data	Evaluation Techniques	Outperformed model	Contribution	Recommendation
[39]	2013	unknown	BPNN, empirical models, fuzzy logic	24h ahead	sunshine hours and air temperature	unknown	MAPE	BPNN	Investigating the appropriate selection of input parameters	Examining the effect of geographical location ,meteorological variables, training algorithm and, the configuration of ANN architecture
[40]	2011	Malaysia	BPNN, empirical models, fuzzy logic	24h ahead	latitude, Longitude, day number, sunshine ratio	1975–2004	MAPE, RMSE	BPNN	Comparing between different locations in Malaysia, large dataset size	N/A
[41]	2015	Turkey	BPNN, ANFIS, MLR, empirical models	monthly solar radiation	month number, radiation, average air temperature, average humidity, average sunshine duration, daylight hours	between 20 and 45 years	MAE,MARE RMSE	BPNN	Applying VIF to measure any multi-collinearity among independent variables, Examining the effect of combinations of input parameters	N/A
[42]	2019	Iraq	BPNN,ANFIS MLR, empirical models	24h ahead	max temperature, min temperature, mean temperature, relative humidity, wind speed	Jan,2006-Dec,2016	RMSE	BPNN	Applying sensitivity analysis to determine the dominant parameters	Applying Ensemble Learning Methods to increase the generalization capability
[43]	2022	South Africa	BPNN, ANFIS		temperature, relative humidity, wind speed, wind direction, GHI	2019-2020	MAPE, MAPE	Same performance	Examining the effect of geographical location	Investigating more input parameters such as clearness index could. Applying large dataset

combines RBPN with Wavelet Transformation (WRBPN). A forecast of daily solar irradiance was carried out based on daily records of irradiance by Baosan Meteorological Station in Shanghai from 1995 to 2000. The WRBPN model demonstrates remarkable improvement in the accuracy of the forecast for the daily solar irradiance of a year compared with models not combining wavelet transformation. The forecast-ing performance with wavelet analysis reports for 7.83%, which is one fourth of the forecast performance without wavelet analysis. Future investigation like the selection of proper mother functions of wavelet, and the optimal updating of weights and biases could enhance the model’s performance.

This was further examined in 2011 by [45] who proposed a predictive model that is based on RNN trained with the Levenberg-Marquardt backpropagation learning algorithm to forecast the solar irradiance. The solar irradiance data was collected at the Zero Energy Center at the University of the District of Columbia at Washington D. C. between June 2011 and July 2011. The model is designed to predict future values of solar irradiance, based on the previous solar irradiance. The proposed RNN showed excellent predictions based on the MSE analysis, error autocorrelation function analysis, regression analysis, and time series response.

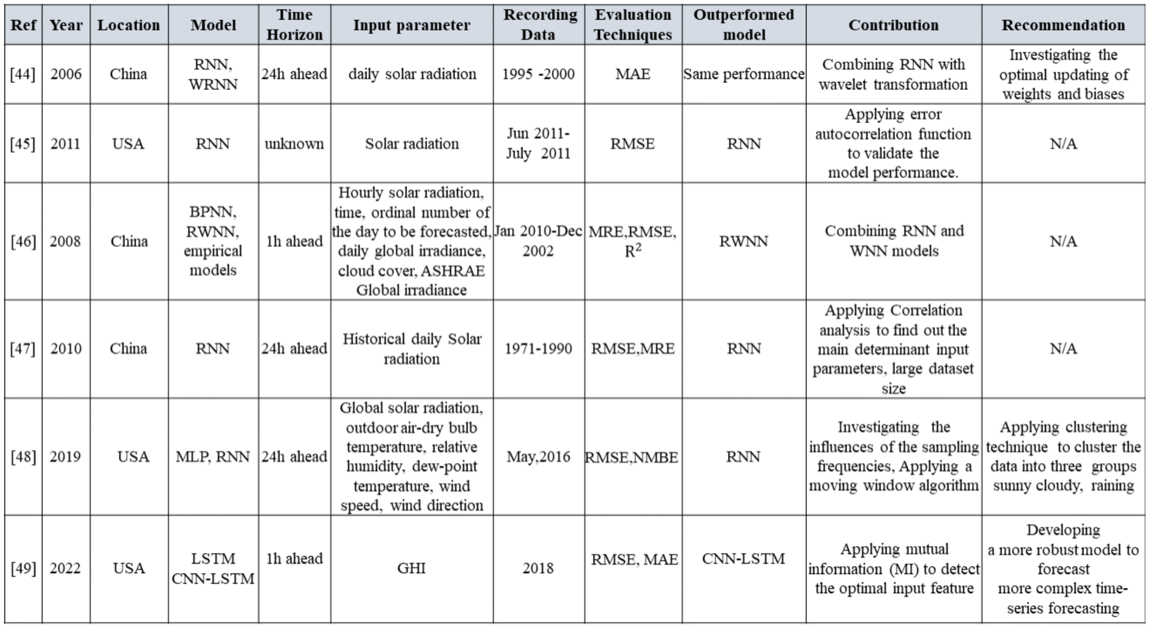
In [46], combining RNN with Wavelet Neural Network (WNN) , a Diagonal Recurrent Wavelet Neural Network (DRWNN) diagonal recurrent wavelet neural network (DRWNN) is proposed for forecasting hourly global solar irradiance. Historical hourly global solar irradiance at Baoshan Meteorological Observatory in Shanghai from 2001 to 2002 is used as the input feature to the DRWNN model. The input vector of the DRWNN has 9 inputs including the hour, ordinal number of the day to be forecasted, defuzzificated cloud cover on the day to be forecasted, and the hourly records of global solar irradiance of the hours 14, 15, 28, and 29 h before the hour to be forecasted. The existing ASHRAE (2005) model gives relatively basic information of the tendency of hourly solar radiation, so global irradiance of the hour to be forecasted, predicted with ASHRAE model is also input to the model, along with the daily global irradiance as released by Meteorological Observatories. The comparisons between the real records of hourly global irradiance and the predicted ones by three models of Collares-Pereira and Rabl, BP network, and DRWNN found that the forecasts by DRWNN coincide with the real records quite well and the forecasts by Collares-Pereira and Rabl, and BP network do not perform as well as by DRWNN in regards to RMSE, and R^2.

More recent research by [47] uses RNN to forecast the daily total solar irradiance. The input features in the proposed model are based on the data sequence of the historical daily records of irradiance by Longhua Meteorological Station in Shanghai from 1971to 1990. These features include solar elevation angle, solar azimuth angle, solar hour angle, sun declination, and geographic location, sunshine durations, and the day number of a year, the forecast time, air mass, clouds. The correlation analysis was used to calculate the correlation coeficients between the features and solar irradiance where the proposed model with high correlated features demonstrates remarkable improvements in the accuracy of the forecasting daily solar irradiance compared with that without applying correlation analysis. Table 2.4 is a summary of several research works on developing Solar Irradiance Forecasting based on RNN Models.

2.1.5 LSTM Model-based Solar Irradiance Forecasting

Finally, a recent paper [15] provide compelling support for the eficacy of RNN. More specifically, however, this research suggests that a new model, LSTM may be the most beneficial. Indeed, in [48] an ANN model and a RNN model are developed for

Table 2.4: RNN Model-based Solar Irradiance Forecasting



Ref	Year	Location	Model	Time Horizon	Input parameter	Recording Data	Evaluation Techniques	Outperformed model	Contribution	Recommendation
[44]	2006	China	RNN, WRNN	24h ahead	daily solar radiation	1995 -2000	MAE	Same performance	Combining RNN with wavelet transformation	Investigating the optimal updating of weights and biases
[45]	2011	USA	RNN	unknown	Solar radiation	Jun 2011- July 2011	RMSE	RNN	Applying error autocorrelation function to validate the model performance.	N/A
[46]	2008	China	BPNN, RWNN, empirical models	1h ahead	Hourly solar radiation, time, ordinal number of the day to be forecasted, daily global irradiance, cloud cover, ASHRAE Global irradiance	Jan 2010-Dec 2002	MRE,RMSE, R^2	RWNN	Combining RNN and WNN models	N/A
[47]	2010	China	RNN	24h ahead	Historical daily Solar radiation	1971-1990	RMSE,MRE	RNN	Applying Correlation analysis to find out the main determinant input parameters, large dataset size	N/A
[48]	2019	USA	MLP, RNN	24h ahead	Global solar radiation, outdoor air-dry bulb temperature, relative humidity, dew-point temperature, wind speed, wind direction	May,2016	RMSE,NMBE	RNN	Investigating the influences of the sampling frequencies, Applying a moving window algorithm	Applying clustering technique to cluster the data into three groups sunny cloudy, raining
[49]	2022	USA	LSTM CNN-LSTM	1h ahead	GHI	2018	RMSE, MAE	CNN-LSTM	Applying mutual information (MI) to detect the optimal input feature	Developing a more robust model to forecast more complex time-series forecasting

the solar irradiance forecast. The meteorological data from a local weather station in Alabama were used for the training and testing processes, which included global solar irradiance, air-dry bulb temperature, relative humidity, dew-point temperature, wind speed, and wind direction. Various scenarios, including different sampling frequencies and moving window algorithms, are included for evaluation of the overall performance. In addition, 10 min, 30 min, and 60 min sampling frequencies were applied to investigate their influences on the overall performance. The results suggest that compared with the ANN model, the solar irradiance forecast using the RNN model has a higher forecasting accuracy. They concluded that raining data with a higher sampling frequency; by increasing the sampling frequency of the training data 60 min to 10 min can improve the forecast performance for both ANN and RNN. Additionally, applying a moving window algorithm can also fairly improve the prediction accuracy. They indicated that cloud cover could cause a significant impact on the forecast accuracy specifically for complex time series data forecasting [49].

The conclusions of the research [15] stated that the ability to predict day ahead solar irradiance forecasting was compared for a LSTM, FFNN, a persistence model, and a Nonlinear Auto-Regressive Model (NAR) model. Six datasets from four different countries with diverse climatic conditions and geographical features were used. The weather variables collected in each location from 2008 to 2018 and six experiments came from weather stations in Germany, U.S.A, Switzerland, and South Korea. To select the most influenced meteorological parameters, the authors used a Pearson correlation coeficient to capture the correlation between solar irradiance and each of these meteorological parameters. As a result, they selected three exogenous fea-tures as input features (dry bulb temperature, dew point temperature, and relative humidity), as well as two additional categorical features (the hour of the day and the month of the year). In case of NAR model, two additional endogenous features the solar irradiance for the previous two days were added as input features. Experimental results provided compelling evidence that the proposed approach of LSTM is superior to that of FFNN and NAR in terms of RMSE.

Importantly, we believe that LSTM model demonstrates promise in its effectiveness, though the literature is sparse, and the current model has a few notable areas for improvement. Based on the stated suggestions from the literature, the areas for future study that would be investigated are outlined. The LSTM model has become more prevalent as a solar irradiance forecaster and its performance has not been fully investigated, which this work will seek to do. Secondly, the association of exogenous (i.e. geographical and meteorological variables) and endogenous variables (i.e. historical solar irradiance) as input features to lead to better overall prediction results will be investigated. Third, it has been suggested that in order to increase the accuracy of the DL based forecasters techniques, large datasets are preferable to enhance decision-making capabilities, however, this high dimensional, heterogeneous data could be suffering from data quality issues as the database of historical records of solar irradiance tends to be quite large. One way of resolving this issue is applying clustering algorithms to classify each data point into three clusters including sunny, cloudy, and rainy weather to reduce the uncertainty associated with forecasting solar irradiance. A growing body of literature has informed LSTM, though we review only those most germane to the current research, for purposes of brevity.

The first study of relevance to consider is [50]. The authors propose LSTM to pre-dict hourly solar irradiation forecasting along with techniques like Gradient Boosted

Regression Trees (GBRT) and FFNN. It was found that LSTM models are suitable to minimize MAE, with relatively low RMSE. An exploration of multi-location-based models demonstrated significant performance improvement over single location-based models which lends further support to the eficacy of LSTM. It is important to note here that in instances where GHI values of the neighboring locations were taken along with the target location, the future GHI of a target location is dependent on prior GHIs of neighboring locations.

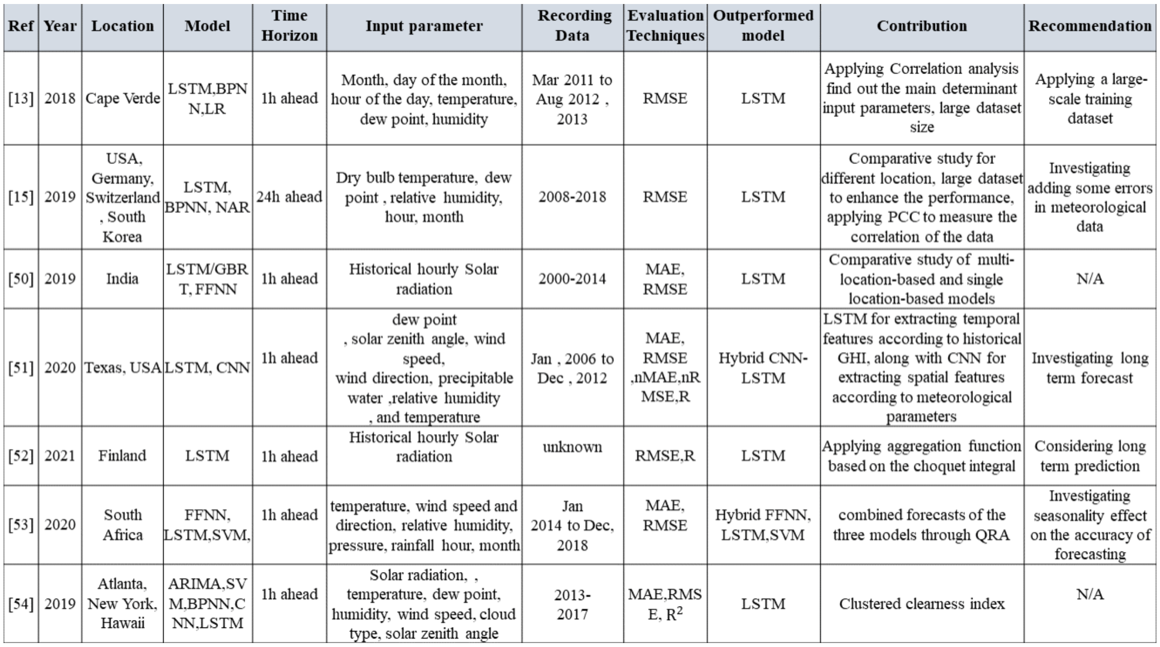
Prior work on LSTM has found similar results. [13] for instance, compare four different models for hourly day-ahead solar irradiance forecasting by using LSTM, persistence algorithm, linear least square regression and BPNN. The proposed LSTM model outperformed the other models in terms of RMSE and shows less overfitting and better generalization capability. They indicated that the performance of the forecasting improved by applying a large-scale training dataset, which provides support for suggestions in the literature that the size of the dataset leads to marked improvements in forecasting. While here, the proposed LSTM forecast model is trained to discover the dependence of hourly weather forecasts between consecutive hours of the same day, [51] proposed a hybrid CNN-LSTM model with spatiotemporal correlations to enhance the accuracy of hourly solar irradiance forecast. The proposed model implements a Convolutional Neural Network (CNN) to extract spatial features based on meteorological parameters and an LSTM model to extract temporal features based on historical GHI time series data. This effectively combines the temporal and spa-tial correlations of the data to obtain accurate predictions of hourly GHI. The GHI values of the previous 10 h are taken as the inputs of the LSTM network to predict the GHI 1 h in advance. The proposed LSTM forecast model is trained to discover the dependence of hourly historical GHI sequences.

In [52], the authors employ LSTM and an aggregation function based on the choquet integral as a way to forecast hourly solar irradiation. The choquet integral improves the model by using a fuzzy measure for capturing interactions occurring between the input features. It is worth noting that other related works utilized either ensemble techniques such as weighted average or relied solely on individual forecasting models. As neither of these strategies took into consideration any interactions that may happen between aggregated values, the forecasting reliability was invariably worsened. The researchers in [53] looked at the forecasting accuracy of FFNN, LSTM, and support vector regression (SVR) models in hourly GHI forecasting. Their results showed that, under all tested conditions, the FFNN model outperformed the LSTM and SVR models. They found that the combined forecasts of the three models through quantile regression averaging (QRA) significantly improves forecasting accuracy. Moreover, the researchers also discovered that solar irradiance seasonality had a major effect on the accuracy of forecasting. These findings support those of [54], who introduced an LSTM based strategy for one-hour-in-advance predictions ofGHI. One of the main conclusions of these researchers was that most of the forecasts that were inaccurate fell on days that were partially or mainly overcast. To overcome this inaccuracy, the researchers introduced a clearness index as LSTM model input data. The clearness index classifies weather type using k-means in the data processing step. More specifically, the k-means took into account the day's total GHI, along with the equivalent clearness index. The classification was then simplified into three main categories, namely cloudy, mixed (partially cloudy), and sunny. Alternative meteorological data may also be used instead of, or in addition to, the clearness index in order to identify sky conditions and weather type. In related work, and based on the findings in [48], Pan et al. [55] explored day-ahead hourly forecasting for solar generation using an ensemble model and combined cluster analysis. However, because they focused mainly on solar generation data clustering in order to find a weather regime, they essentially ignored any weather-related data that may explain nonlinearity behavior in solar irradiance [55]. Much of the current literature focuses on applying DL techniques to hour-ahead solar irradiance forecasting and modeling, with the LSTM model emerging as the most prevalent forecaster of solar irradiance. A review of the past and current literature has suggested that the benefits of RNN and LSTM over previously used models [15][45][47][48][50]. Table 2.5 is a summary of several research works on developing Solar Irradiance Forecasting based on LSTM Models.

2.2 Features Selection

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.,

Table 2.5: LSTM Model-based Solar Irradiance Forecasting



Ref	Year	Location	Model	Time Horizon	Input parameter	Recording Data	Evaluation Techniques	Outperformed model	Contribution	Recommendation
[13]	2018	Cape Verde	LSTM,BPNN,LR	1h ahead	Month, day of the month, hour of the day, temperature, dew point, humidity	Mar 2011 to Aug 2012 , 2013	RMSE	LSTM	Applying Correlation analysis find out the main determinant input parameters, large dataset size	Applying a large-scale training dataset
[15]	2019	USA, Germany, Switzerland , South Korea	LSTM, BPNN, NAR	24h ahead	Dry bulb temperature, dew point , relative humidity, hour, month	2008-2018	RMSE	LSTM	Comparative study for different location, large dataset to enhance the performance, applying PCC to measure the correlation of the data	Investigating adding some errors in meteorological data
[50]	2019	India	LSTM/GBRT, FFNN	1h ahead	Historical hourly Solar radiation	2000-2014	MAE, RMSE	LSTM	Comparative study of multi-location-based and single location-based models	N/A
[51]	2020	Texas, USA	LSTM, CNN	1h ahead	dew point , solar zenith angle, wind speed, wind direction, precipitable water ,relative humidity , and temperature	Jan , 2006 to Dec , 2012	MAE, RMSE ,nMAE,nRMSE,R	Hybrid CNN-LSTM	LSTM for extracting temporal features according to historical GHI, along with CNN for extracting spatial features according to meteorological parameters	Investigating long term forecast
[52]	2021	Finland	LSTM	1h ahead	Historical hourly Solar radiation	unknown	RMSE,R	LSTM	Applying aggregation function based on the choquet integral	Considering long term prediction
[53]	2020	South Africa	FFNN, LSTM,SVM,	1h ahead	temperature, wind speed and direction, relative humidity, pressure, rainfall hour, month	Jan 2014 to Dec, 2018	MAE, RMSE	Hybrid FFNN, LSTM,SVM	combined forecasts of the three models through QRA	Investigating seasonality effect on the accuracy of forecasting
[54]	2019	Atlanta, New York, Hawaii	ARIMA,SVM,BPNN,CNN,LSTM	1h ahead	Solar radiation, , temperature, dew point, humidity, wind speed, cloud type, solar zenith angle	2013-2017	MAE,RMSE, R^2	LSTM	Clustered clearness index	N/A

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.

Endogenous and exogenous variables usually get evaluated numerous times throughout a predictive model's selection trial. A few of the more common feature selection strategies are depicted in Figure 2.2. Feature selection is done using wrapper, filter, and embedded techniques. The wrapper approach, also known as the greedy search algorithm, applies an ML algorithm to determine the best feature set for the algorithm. However, the wrapper approach is lacking in generalization, which means that the best subset selected would not likely be the best when applied to other ML algorithms. As well, the wrapper method's search procedure is quite costly from a computational perspective. The filter method employs a number of different statistical measures, such as correlation analysis, univariate statistical analysis, mutual information, ANOVA, chi-square distribution, in order to determine the relevancy and/or redundancy of certain variables. The primary advantages of using the fil-ter approach are the fast screening time, low computational cost, and the accurate measurement of monotonic and linear degrees of variable pairs. For the embedded approach, feature selection occurs at the design stage, when the ML algorithm is being trained. Like the filter method, the embedded approach has the benefit of accuracy in results and fast screening times, but like the wrapper method, they are computationally expensive. The following section presents an overview of the main feature selection methods specific to climatological, atmospheric and meteorological data in application to solar irradiance. The associations of exogenous (meteorological and geographical) variables and endogenous (historical solar irradiance) variables as features that result in improved prediction results will be examined.

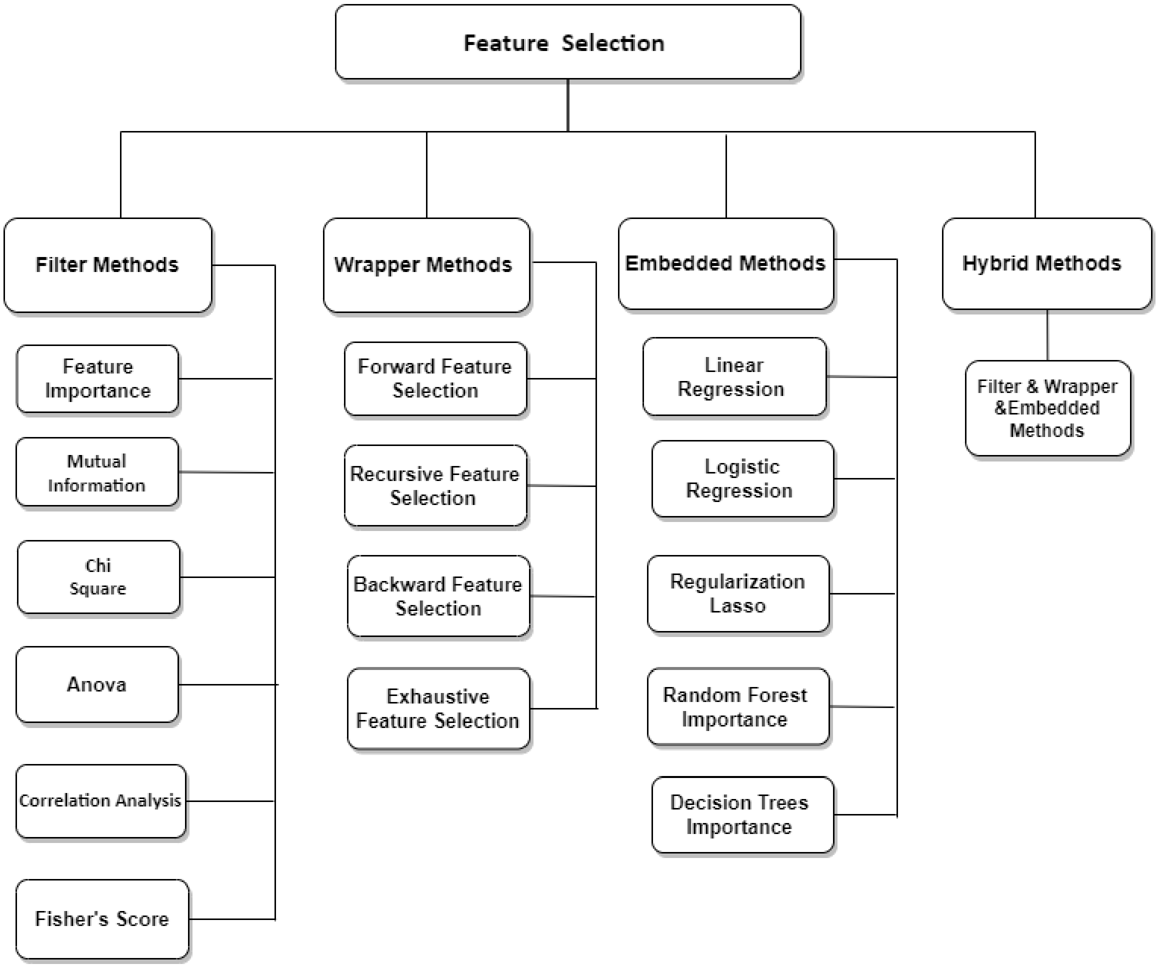


Figure 2.2: Feature Selection Methods

Most feature selection techniques in the literature are based on correlation analysis for redundancy and relevancy measures and are performed to determine whether a linear relationship exists among the input features [13] [15] [16] [17]. Indeed, a significant amount of work has focused on other statistics and DL techniques are extended to extract the optimal feature subset for GHI forecasting model. The techniques include Subsets Evaluator, VIF,ARD,NGA are proposed in [37] [41] [56] [57] [58]. ANN and fuzzy logic models are used in [41], Specifically, ANN, ANFIS, Multiple Linear Regression (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. To determine the multi-collinearity of independent variables, VIF was used. The results indicate that Calendar Month Number (M) Extraterrestrial Radiation (R_a),Average Air Temperature (T_{mean}),Average Relative Humidity (RH_{mean}) can be powerful input features when estimating solar irradiance in ANN, ANFIS, and MLR models.

As well, the authors dissect and compare different input variable combinations using MAE,MAE, MARE, RMSE, overall index of model performance (OI), and R^2. They 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 is supported by [56], 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 [57] proposed 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 WJ/m^2, using sunshine hours and the most relevant variables (pressure, humidity, temperature). In [58], the authors proposed combining Generalized Fuzzy model (GFM) and continuous density Hidden Markov Model (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^2 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 [37], several authors compared 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 Extraterrestrial Global Solar Radiation (H_O), Sunshine Duration (S_O), Max Air Temperature (T_{max}),Mean Air Temperature (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.

2.3 Summary

This chapter began with a comprehensive literature review of several forecasting techniques and feature selection methods. Next, various strategies for GHI forecasting models (MLP, fuzzy logic, RNN, and LSTM) were reviewed.

Chapter 3

Research Methodology and Data Collection

This chapter contains a portion of the works that are published in research papers [18] [21] [22].

3.1 Introduction

The research methodology includes three phases that aim to enhance the forecast-ing accuracy of the solar irradiance model shown in Figure 3.1. The first phase is primarily focused on investigating the behavior of LSTM model and the influence of the input features (e.g., univariate, or multivariate) on the model's performance. The main outcomes of this work show that the performance of the LSTM forecasting model deteriorates when changed from univariate to multivariate analysis. This deterioration provides motivation to comprehensively analyse the way that features are being evaluated during the selection of a predictive model for GHI. The second phase of the research methodology is devoted to proposing a novel hybrid feature selection method that optimizes feature selection using a Least-Redundant/Highest-Relevant framework. The proposed WRFE utilizes feature importance for measuring variance reduction in RFR and as data perturbation in LSTM . The key results show that the proposed optimal features of the training subset make the greatest contributions to the prediction hourly GHI. The results also prove that the high variability of irradiance data due to geographical and meteorological conditions lowers the reliability of the training subset. Therefore, a primary investigation is conducted in the third phase to analyse seasonality-based hourly predictions for GHI. The investigation finds that seasonality affects the accuracy of predictions due to high levels of autumn- and winter-related weather phenomena and climate uncertainty. Accordingly, the SCFT based on an LSTM hybrid strategy and stacked layers of weather clusters is proposed. The main results show that the proposed SCFT forecasting approach yields improvements in the learning tasks when training the LSTM model by enhancing the model's

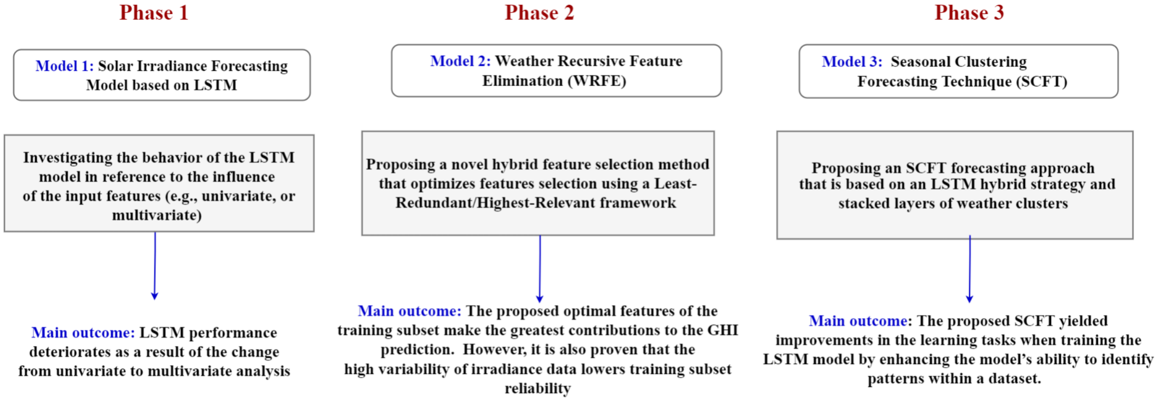

Phase 1
Phase 2
Phase 3
Model 1: Solar Irradiance Forecasting Model based on LSTM
Model 2: Weather Recursive Feature Elimination (WRFE)
Model 3: Seasonal Clustering Forecasting Technique (SCFT)
Investigating the behavior of the LSTM model in reference to the influence of the input features (e.g., univariate, or multivariate)
Proposing a novel hybrid feature selection method that optimizes features selection using a Least-Redundant/Highest-Relevant framework
Proposing an SCFT forecasting approach that is based on an LSTM hybrid strategy and stacked layers of weather clusters
Main outcome: LSTM performance deteriorates as a result of the change from univariate to multivariate analysis
Main outcome: The proposed optimal features of the training subset make the greatest contributions to the GHI prediction. However, it is also proven that the high variability of irradiance data lowers training subset reliability
Main outcome: The proposed SCFT yielded improvements in the learning tasks when training the LSTM model by enhancing the model's ability to identify patterns within a dataset.

Figure 3.1: Workflow of Research Methodology

ability to identify patterns within a dataset. This forms the basis for dataset clusters and making predictions of the hourly GHI with suficient accuracy even for regions with highly fluctuating climates.

3.2 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 3.1 [59].

$$\rho_P(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\,_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{3.1}$$

where x_i is the x-variable value, $\bar{x}$ is the mean of the x-variable, y_i is the y-variable value, $\bar{y}$ is the mean of the y-variable, n is the number of observations.

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. Pearson correlations can be highly susceptible to normality and linearity assumptions and can also easily per-ceive 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 3.2. 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 [59].

$$\rho_S(X, Y) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3.2}$$

where d_i is the difference between the two ranks of each observation. n is the number of observations.

Hoeffding's D is applied as a non-parametric rank-based measure for determining non-linear associations, as presented in Equation 3.3. 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 [60].

$$D(X, Y) = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)} \tag{3.3}$$

where D_n is the rank of variable n in two different samples.n is the number of observations.

3.3 Data Collection and Analysis

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) [61]. The data were downloaded using the NSRDB data viewer for Halifax, Nova Scotia, Canada, at coordinates 44.88 N and 63.51 W. In the first model, the data covered four years (2014-2017) on an hourly basis, were for Dew Point (DP), Temperature (T), Relative Humidity (RH), Pressure (P), Precipitable water (PW) , Diffuse Horizontal Irradiance (DHI), GHI, and Solar Zenith Angle (SZA). The four designated years were used for training, while the year 2018 was used for testing. The Correlation coeficients indicate redundant attributes because they relay more or less similar information, resulting in multicollinearity. Such correlated features may cause overfitting. Therefore, the Pearson Correlation (ρ_P) has been applied for measure any linear associations occurring in the explanatory/predictors variables. As shown in Figure 3.2, a heat map method of bivariate analysis has been utilized for visualizing correlation coeficients. In the figure, the ρ_P coeficients whose values are somewhere between (+1 and-1). The map illustrates that DP has the strongest positive linear relationship with PW and T of up to ρ_P = 0.82 and ρ_P = 0.95 respectively, in comparison to the other data variables. As well, the figure indicates that PW has a positive relationship of ρ_P = 0.71 with T. Moreover, in the same figure, the correlation analysis results indicate that SZA has a negative liner relationship (ρ_P = -0.73 and ρ_P = -0.77) with DHI and GHI, respectively, while RH and GHI both have a moderate one (ρ_P = -0.57). Our suggestion would be to remove one of the variables in any explanatory variable pairs where (ρ_P>±0.5). However, because GHI represents a response variable, this should be considered when eliminating any redun-dant variables. In other words, a stronger association between GHI and redundant variables may mean the variable could be more relevant for the model. Based on this assumption, and as shown in the graph, SZA and T were retained because they show the strongest association of the redundant variables.

The input features for the forecasting model were then formulated using the five retained features (i.e., after eliminating DHI, DP and PW). In Table 3.1, the descriptive statistics information for the training dataset using RH, T, SZA, GHI, and P is presented. The data have been standardized to enable feature scaling. The table also shows five features that have been scaled to a mean of 0, with a standard deviation of 1. A further assumption is that every variable's contribution in the analysis is equal, having no bias.

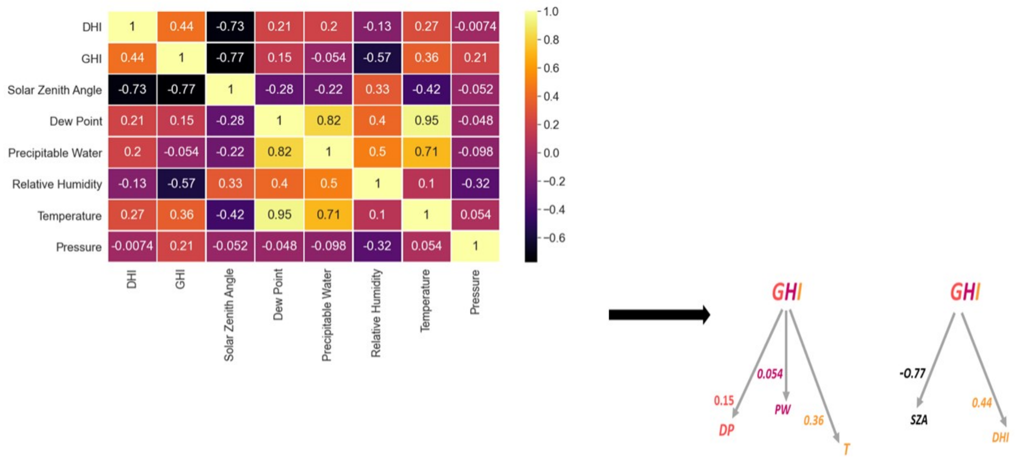


Figure 3.2: Pearson Correlation

Table 3.1: Descriptive Statistics Information for Model 1 Dataset

	count	mean	std	min	25%	50%	75%	max
GHI	1.46E+03	4.91E-17	1.00E+00	-1.56E+00	-8.39E-01	-1.58E-01	7.88E-01	2.26E+00
SZA	1.46E+03	5.63E-16	1.00E+00	-3.20E+00	-6.47E-01	2.25E-01	8.20E-01	1.70E+00
RH	1.46E+03	-6.94E-16	1.00E+00	-2.25E+00	-8.35E-01	-1.59E-02	8.21E-01	1.97E+00
T	1.46E+03	9.76E-17	1.00E+00	-2.30E+00	-8.63E-01	-1.28E-01	9.47E-01	1.91E+00
P	1.46E+03	-1.83E-16	1.00E+00	-2.06E+00	-7.98E-01	1.01E-01	9.35E-01	1.42E+00

3.4 Comprehensive Data Analysis

The second model utilizes data from the same source of the first model with the same geographic location using the NSRDB data viewer for Halifax, Nova Scotia. The data were collected for hourly periods using large data range with time-frame of 2000 - 2018. The dataset in second model includes more meteorological data such as Direct Normal Irradiance (DNI), Clearsky GHI (CGHI), Surface Albedo (SA),Wind Speed (WS), and Wind Direction (WD), in addition to the meteorological data included in the first model (DP, T, RH,P, PW, DHI,GHI, SZA). Table 3.2 presents the descriptive statistics information for the dataset that includes around 82,118 observations.

3.4.1 Redundancy Measures

Redundant attributes are usual measured by Pearson’s correlation coeficient. Figure 3.3 shows a heat map technique employed to visualize the correlation coeficients. As

Table 3.2: Descriptive Statistics for Model2 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

can be seen, the coeficient 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 insuficient to render a full decision on the redun-dant 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 3.3 and 3.4, it is seen the Pearson and Spearman correlation coeficients for evaluating bivariate analysis and for measuring linear or monotonic relationships in the vari-able pairs. The coeficient 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 the endogenous variable, GHI, is excluded 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 coeficients and their respective p-values. A hypothesis test will be performed to determine any significances in the correlation coeficient and to gauge if the sample data's linear/monotonic relationship may be suficiently 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 Coeficient (ρ_C),

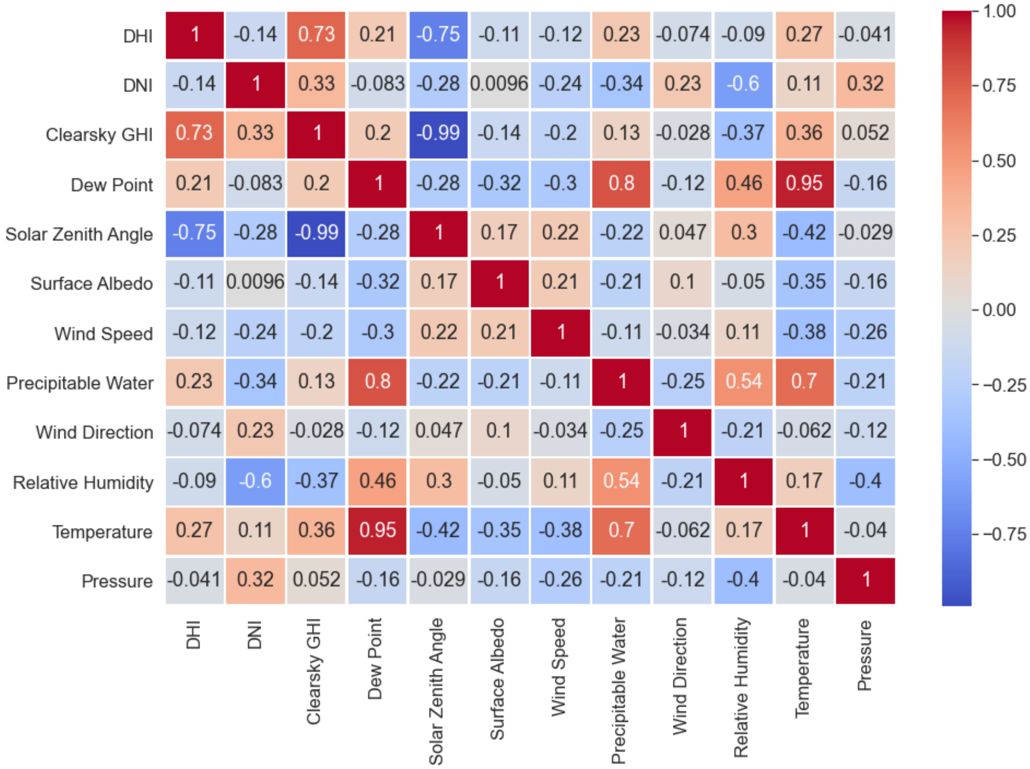


Figure 3.3: Heatmap of Pearson Correlation

we need to see if there is 95% confidence (at a 0.05 level of significance), in which case:

H0:$\boldsymbol{\rho_C}$**= 0** ("If the population correlation coeficient equals 0, no association is detected").

H1:$\boldsymbol{\rho_C}$**= 0** ("If the population correlation coeficient does not equal 0, a nonzero correlation may exist").

As shown in Tables 3.3 and 3.4, the Pearson and Spearman correlation coeficients 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 coeficients being -0.74 and -0.81, respectively, and p ¡ .0001. Therefore, between SZA and CGHI, using the Pearson and Spearman correlation

Table 3.3: Pearson Correlation Coeficients for the Year 2000

Pearson Correlation Coeficients, 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	1.000	-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	

coeficients, there is a robust association of -0.99 and -0.995, respectively.

Figure 3.4 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 coeficient. 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 3.3 gives p values to test associations between DNI and SA, CGHI and WD, and P and SZA. The 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.

Normality Assumption Tests

Table 3.5 provides descriptive statistics analyses of DHI datapoints. In order to determine whether the DHI data are normally distributed, numerical techniques by looking

Table 3.4: Spearman Correlation Coeficients for the Year 2000

Spearman Correlation Coeficients, 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	1.000	-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	

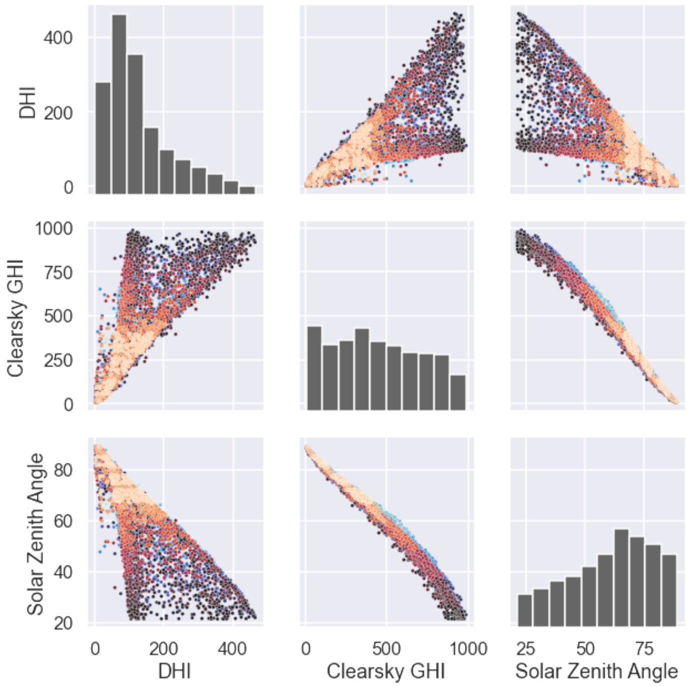


Figure 3.4: Pairwise Analysis in Scatterplot

Table 3.5: 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

at kurtosis and skewness values to gauge normality according to criteria proposed in [62] can be applied. 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, it is determined 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 [63] 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 3.6. 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 3.5 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 3.7 presents both the Spearman and Pearson correlation coeficients in descending rank for DHI and a range of variables.

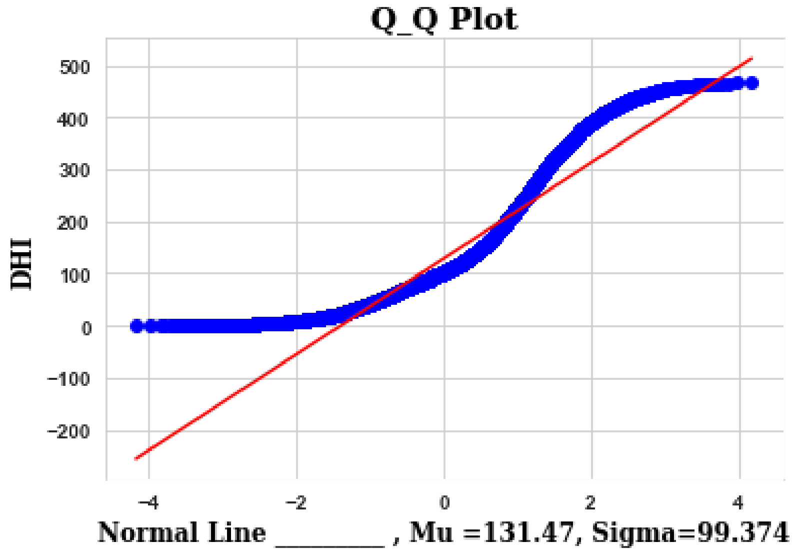


Figure 3.5: QQ Plot for DHI

Table 3.6: 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

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 coeficients Spearman Correlation (ρ_S) 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, it is needed 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 3.8 provides a list of highly correlated variables which could potentially be redundant. As can be seen, there is a positive monotonic relationship

Table 3.7: Nonparametric Measure of Association with DHI for a 95 % Confidence Interval

Variable	Pearson Rank	Spearman Rank	Pearson Coeficient	Pearson P-Value	Spearman Coeficient	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 3.8: List of High Correlated Variables

First Feature	Second Feature	Spearman Coeficients
DHI	CGHI	0.79
DHI	SZA	-0.81
CGHI	SZA	-0.99
DNI	RH	-0.60
DP	T	0.94
DP	PW	0.83
T	PW	0.72
RH	PW	0.56

between CGHI and DHI, where (ρ_S) = 0.79. Additionally, it can be seen that there is a negative monotonic relationship between SZA and DHI, where (ρ_S) = -0.81. There is also a strongly negative association between SZA and CGHI, where (ρ_S) = -0.99. In the same table, a negative moderate association exists between DNI and RH, where (ρ_S) = -0.6. Further, in comparison to other data, DP shows the strongest positive monotonic and linear relationships to T (up to (ρ_S) = 0.94), and a positive, monotonic, and curvilinear relationship (up to (ρ_S) = 0.83) to PW. There is also a moderate association of (ρ_S) = 0.72 between PW and T.

3.4.2 Effect of Sample Size in Correlation Analysis of Weather Data

To test stability visually, a comparison of the correlation analysis of a one-year dataset is made. 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, a relatively stable magnitude of correlations both in large and small data samples is observed. Moreover, the majority of the correlation coeficients within the dataset appeared entirely stable in relation to the dataset size, as shown in Table 3.4 and Table 3.9, 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 coeficients, changing, for instance, from a significantly weak association to a significantly very weak or null association. The correlation coeficients 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 coeficients with ρ_S values greater than 0.4, these deviations are not considered problematic. On the other hand, as most deviations in the correlation coeficients occurred in the yearly dataset (correlation coeficients differ from year to year), they may warrant further investigation.

3.4.3 Relevance Measures

Evaluation of irrelevancy for irrelevant attributes is commonly done using Spearman’s ranking of correlation coeficients. This can be performed by measuring monotonic associations between endogenous and exogenous variables [64][65]. 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

Table 3.9: Spearman Correlation Coeficients for the Years 2000-2010

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

not be identified when only using Spearman's. As demonstrated in [66], 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 3.10 presents a comparison of Hoeffding's D and Spearman's correlation coeficients. As can be seen, CGHI, SZA, DNI, DHI, RH, and T are all deemed relevant attributes to GHI, which is the tar-get. 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 3.11. The completed results of the validation are given in the Appendix A.

Although Spearman's rank correlation coeficients on different data sizes can be employed, Gilpin [67] found that with increases in sample size, the Kendall correlation coeficient is more practical. Croux and Dehon [68] 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 eficiency. The authors in [69] [70] mention that the Kendall correlation has a computation complexity of $O(n^2)$ in comparison to the $O(n \cdot \log n))$ 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. An intensive screening of the features using filter methods that rely on the data's statistical characteristics is performed, such as parametric and non-parametric tests. Equation 3.4 illustrates a way to capture the dependency degree between ith exogenous variables (x) and endogenous variable (y) [71]. Strong dependence shows a high degree of mutual information I, which indicates greater knowledge of joint distribution p(x, y) than marginal distribution p(x)p(y). The normalized mu-tual information (NMIFS) method proposed in [72] as a measure of irrelevancy is applied. For both methods, Figure 3.6 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.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.4}$$

Table 3.10: Nonparametric Relevance Measure for a 95% Confidence Interval(Year 2000

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Coeficient	Spearman p-value	Hoeffding Coeficient	Hoeffding p-value	Kendall Coeficient	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

Table 3.11: Nonparametric Relevance Measure for a 95% Confidence Interval (Years 2000-2010)

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Coeficient	Spearman p-value	Hoeffding Coeficient	Hoeffding p-value	Kendall Coeficient	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

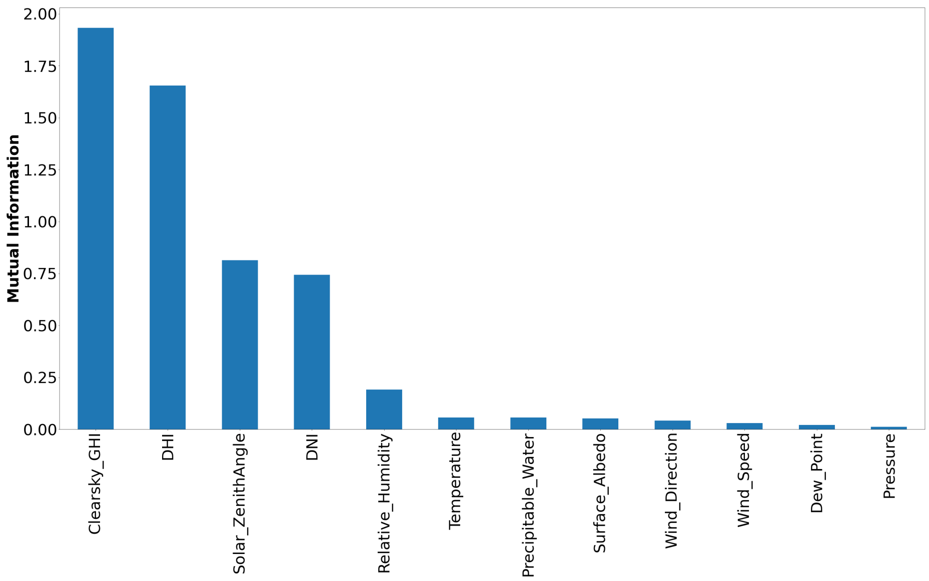


Figure 3.6: Degree of Mutual Information between Exogenous Variables and GHI

3.4.4 Mathematical Model of Long Short-Term Memory

For traditional ANNs, fixed-size vectors may be utilized for primary inputs. However, in these cases, the model's usage may be restricted when variable sized sequence data are involved. RNN represents an improvement over traditional ANNs. As noted in [18], RNNs represent sequence-based models, and as such are capable of obtaining data correlations for a variety of time points. This is because the memory cells in RNNs are updated for every time point, making them dependent on variable-length input/output. Hence, even simple RNNs have enhanced short-term memory capacity. Overall, RNNs' most important feature is their so-called hidden state that recalls information on sequences for each time step. Numerous factors of weight matrix W are involved in computing the gradient of h as shown in equation 3.5.

$$h_t = \tanh \quad w \quad \begin{matrix} & ! \; ! \\ h_{t-1} & \\ X_t & \end{matrix} \qquad (3.5)$$

It is found that RNNs are prone to explosion gradient and vanishing gradient problems [73]. Gradient explosion refers to the exponentially fast increase in the gradient, while gradient vanishing is the exponentially fast reduction in gradient, both of which restrict RNN models' capacity in learning long-term temporal correlations. One possible solution for the gradient vanishing problem may be found in LSTM

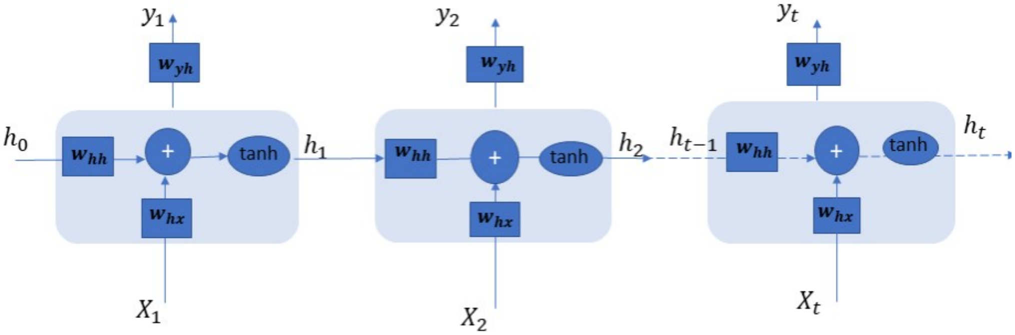


Figure 3.7: RNN Architecture

architecture, as proposed in the study by [74], This approach intends to mitigate both the vanishing and exploding gradients problems in order to capture long-term dependency for the models' inputs and outputs. In addition to RNN cells, building a cell state that includes long-term memory information. In storing patterns, numerous gates are formed, including the Output (o), Input (i), and Forget (f) gates as shown in figure 3.8.

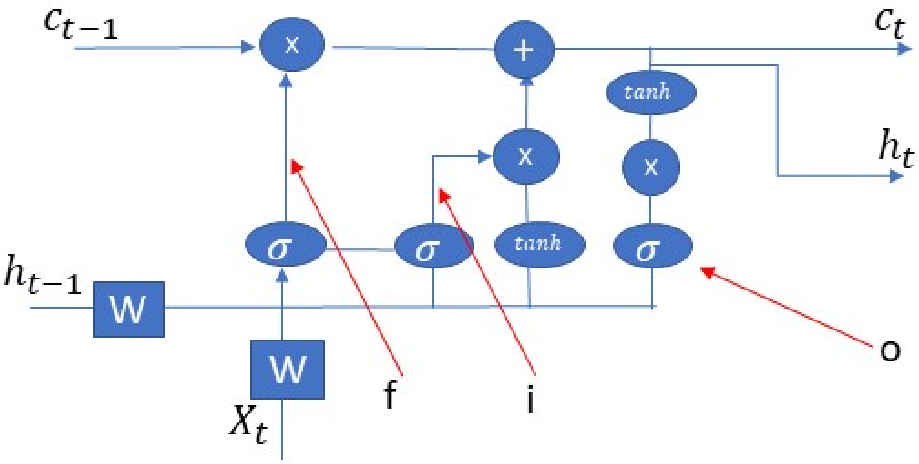


Figure 3.8: LSTM Architecture

The gates are connected via a sigmoid dance layer which serves as a filter (e.g., for output, input, and forget). The outputs of these gates are formulated in the equation3.6.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \tag{3.6}$$

The gates will be utilized for updating the cell state C_t, as formulated in the equation 3.7.

$$c_t = f \odot c_{t-1} + i \odot g \tag{3.7}$$

After being updated, the cell state will then be applied in computing the hidden state as 3.8.

$$h_t = o \odot \tanh(c_t) \tag{3.8}$$

3.5 Summary

This chapter discussed the Research Methodology steps in detail and included feature screening and comprehensive data analysis. The chapter also looked at the LSTM model's mathematical formation

Chapter 4

Models Design and Results

This chapter contains materials that are published in research papers [18] [21] [22]

4.1 Model 1: Solar Irradiance Forecasting Model based on LSTM

LSTM model is being increasingly applied in solar irradiance forecasting, but the performance of LSTM is still relatively unknown. The present section explores how meteorological and geographical (i.e., exogenous) and past records of solar irradiance (i.e., endogenous) variables may be incorporated as input features in day-ahead solar irradiance forecasting models that use DL models. In this study, the results for the LSTM model are compared to those for the RBFNN in relation to both MTSF and UTSF. A comparison of MTSF and UTSF is done for daily solar irradiance readings at yearly intervals. UTSF looks only at GHI for single time-dependent variables when predicting day-ahead solar irradiance. So, this study will consider past data readings for GHI and to understand any correlations that may exist and extract any potential patterns. MTSF, on the other hand, presents multiple variables, so the predicted GHI relies on previous values as well as weather and other meteorological variables. The dependency can be measured using statistical analysis and may be applied to forecasting GHI values in the future.

4.1.1 Experimental Setup

To forecast daily solar readings for Halifax, Nova Scotia , four experiments are conducted using LSTM and RBFNN models designed to reflect two scenarios. These were based on the form of input data (MTSF or UTSF) used in forecasting and modeling time series. From a mathematical perspective, the UTSF strategy may be expressed as a function of past GHI values, as formulated in equation 4.1. The MTSF strategy may also be expressed as a function of past GHI values, along with other weather and meteorological values, as given in equation 4.2.

Table 4.1: Hyperparameter Values

Hyperparameter	LSTM	RBFNN
Learning rate	0.001	0.001
Batch size	7	7
Optimizer	Adam	Adam
No. of Epochs	140	140
Input shape	3-D	2-D
No. of hidden layer	5	5
No. of units in each hidden layer	120	300
No. of units in output layer	1	1
Dropout rate	0.1	0.1

$$H_{t+1} = f\ (H_{t-1}, H_{t-2}, H_{t-3}, \ldots\ldots\ldots\ldots\ldots H_t) \tag{4.1}$$

$$H_{t+1} = f\ (H_{t-1}, H_{t-2}, H_{t-3}, \ldots\ldots H_t, SZA_t, RH_t, T_t, P_t) \tag{4.2}$$

where: $(H_{t-1}, H_{t-2}, H_{t-3}, \ldots\ldots H_t)$ indicates historical solar irradiance H values for time $(t - 1, t - 2, \ldots t)$ and H_{t+1} denotes predicted solar irradiance for time $t + 1$. Note that these forecasting models have been trained using daily observation data from 2014 to 2017; they have also been tested using data from 2018. Thus, the training dataset comprises 1,460 days, while the testing dataset comprises 365 days. The forecasting models have been designed using Keras as applied to TensorFlow 2.0. Table 4.1 shows the hyperparameters values for the proposed LSTM and RBFNN models.

First Experiment:UTSF_LSTM

In this experiment, the LSTM input layer is a three-dimensional tensor. Hence, the 3-D tensor for the training set (N, L, D) denotes the observation number, the length of the sequence, and the number input features. Because this model has been built using a sequence length measuring 30 days, at every time T point, LSTM takes values from the previous 30 days of GHI prior to T into consideration. Thus, according to the resultant monthly pattern, LSTM captures both trends and seasonalities, based on which it predicts the subsequent GHI for time (T+1). Note that the training set has been divided as two subsets to accommodate both input and output training. The input training set structure looks at the adjacent 30 GHI values for time T,

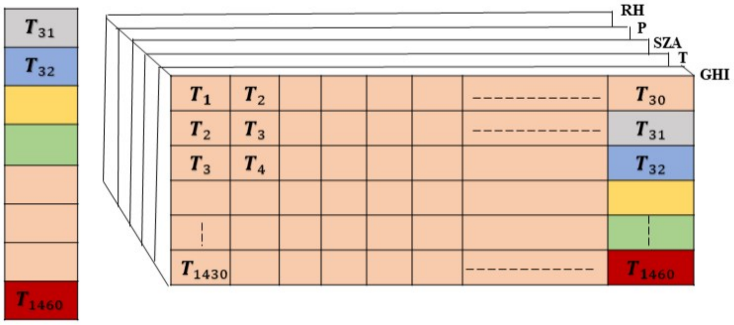


Figure 4.1: Input and Output Training Shape for LSTM Models

thereby creating a 3-D tensor (1430, 30, 1) as well as an output training set of (1430, 1). Here, the input for the test set creates the shape (365, 30, 1), which indicates daily GHI for 2018. The results explained in detail in section (4.1.2).

Second Experiment: MTSF_LSTM

In the second experiment, LSTM is used to find values of the previous month (30 days) for RH, TR, SZA, GHI, and P. Specifically, it will be used for capturing patterns of seasonality pattern in the weather data, and then will utilize these patterns for predicting subsequent GHI for time (T+1). The input training set formation considers adjacent 30 values for RH, TR, SZA, GHI, and P for time T. In so doing, it creates a 3-D tensor (1430, 30, 5) and the GHI output training set at size (1430, 1) as shown Figure 4.1.

Third Experiment: UTSF_RBFNN

A total of 1460 days of solar irradiance values are used as input for the RBFNN. During the training task, the model was trained by employing 1430 days from the original dataset. This created a vector of (1430, 1). Meanwhile, the output layer with 30 output nodes referred to average daily solar irradiance for the subsequent 30 days, as vector (30, 1). The model was trained for the four years (2014-2017), after which it was tested with the aim of predicting daily GHI for 2018.

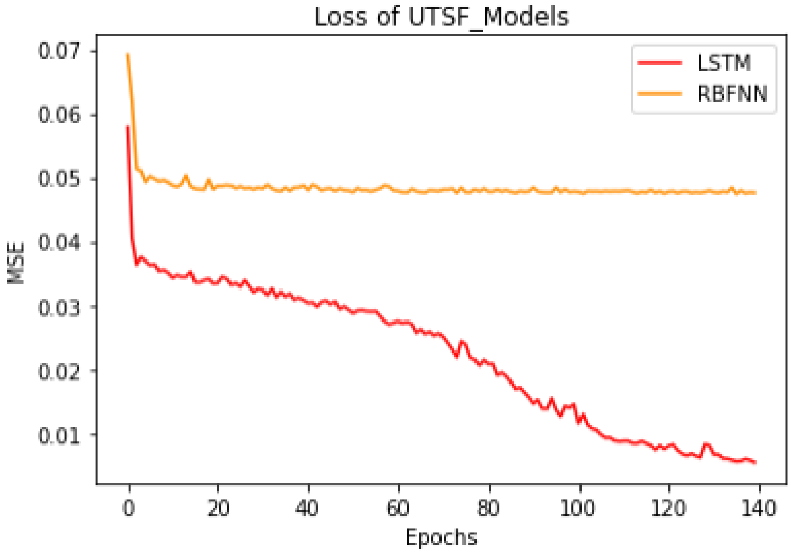


Figure 4.2: Loss of UTSF Models(First and Third Experiments)

Fourth Experiment: MTSF_RBFNN

In the fourth experiment, it is considered the values RH, TR, SZA, GHI, and P to create the matrix (1430, 5) as the input training set form. For the output training set, it is considered the value GHI at size (30, 1). The model was then tested on its ability to predict daily GHI for 2018.

The four models are trained based on batches of weekly historical values. Mini Batch Gradient descent is applied to update the weights every batches of 7 val-ues of GHI and weather data. Figure 4.2 and Figure 4.3 illustrate the learning curves with 140 epochs where the final loss values on the training that are around 0.0042, 0.0461, 0.0248, and 0.0352 for UTSF_LSTM, UTSF_RBFNN, MTSF_LSTM, and MTSF_RBFNN respectively.

4.1.2 Results and Discussion

For the forecasting models, R^2 and RMSE scores have been employed for performance verification. Low RMSE values indicate improved performance, and if the value R^2 tends to 1, this points to a strong relationship existing between response variables and predictors. In Figure 4.4b, it can seen performance metrics for all the mod-els. As shown, LSTM gave the best performance in day-ahead forecasting, but the

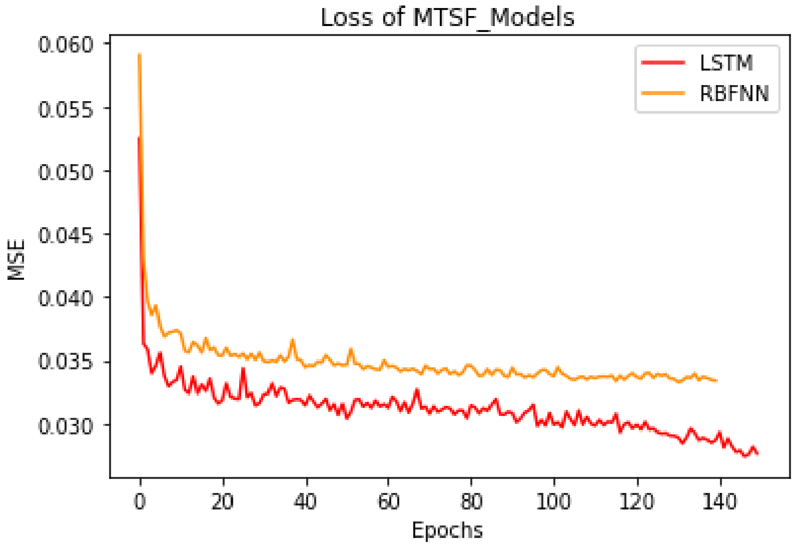


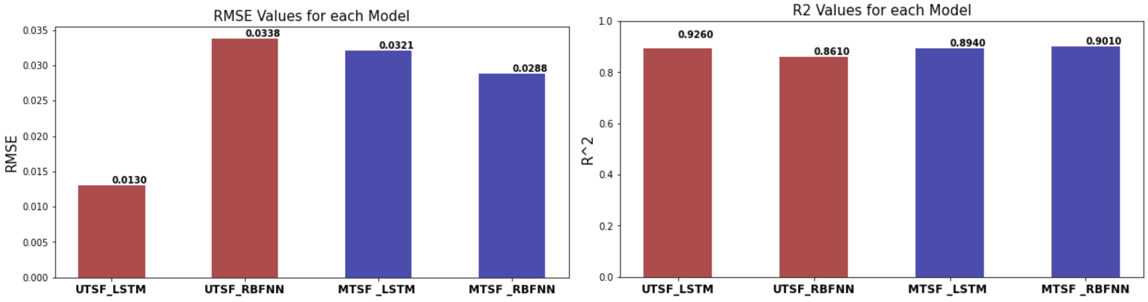
Figure 4.3: Loss of MTSF Models(Second and Fourth Experiments)

UTSF_LSTM model gave a better performance than the rest. This shows the capability of UTSF_LSTM for forecasting GHI through the capturing of nonlinearity patterns embedded in past GHI data. This ability is due to the memory in LSTM cells, which enables them to store information from earlier time steps. It also demonstrates that although the MTSF_LSTM model performed reasonably well, it was outperformed by the MTSF_RBFNN. These results are due to the non-linear nature of the latter model and its strong tolerance to input noise. As for the UTSF_LSTM model, it can be seen that the RMSE between observed and estimated solar irradiance was 0.013. Adding some more features to MTSF_LSTM bumps the RMSE up to 0.0321. In the UTSF_RBFNN model, the RMSE was 0.0338. However, when more features are added, it was found that the MTSF_RBFNN RMSE declined, landing at 0.0288.

As can be seen in Figure 4.5 and Figure 4.6 show the real daily GHI and forecasted daily GHI of the four models for year of 2018 in Halifax, NS.

4.1.3 Model Validation

In model development, validation is a crucial step in the process. The amount of solar irradiance registered in a specific location, as mentioned previously, can vary depending on factors such as time of day, latitude, and time of year. To validate



(a) RMSE values

(b) R^2 values

Figure 4.4: Performance of The Four Proposed Models

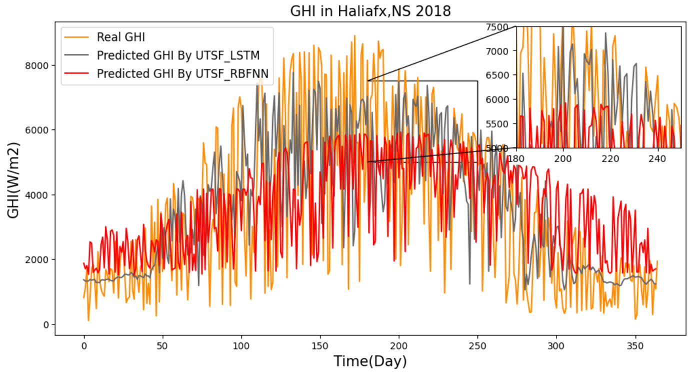


Figure 4.5: Measured GHI Vs. Predicted GHI of UTSF Models

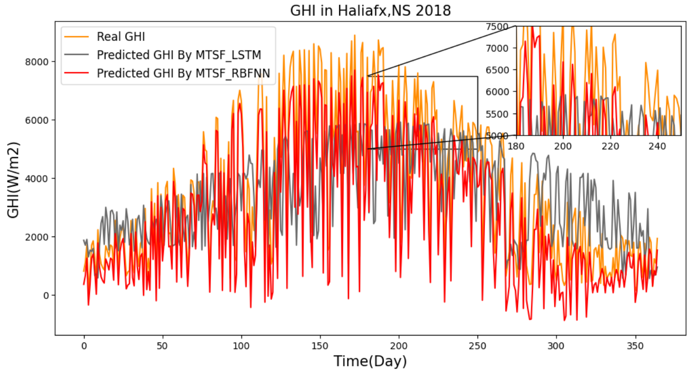


Figure 4.6: Measured GHI Vs. Predicted GHI of MTSF Models

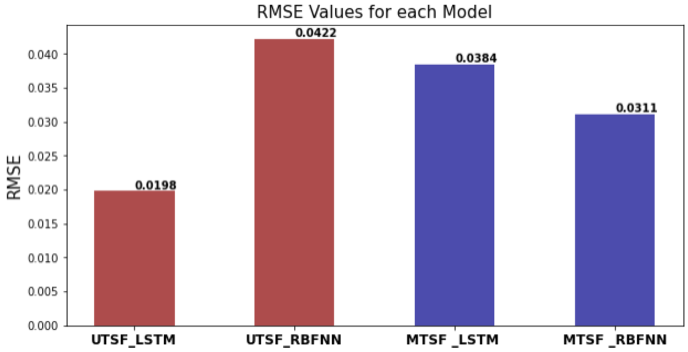


Figure 4.7: RMSE Values for Validation Data

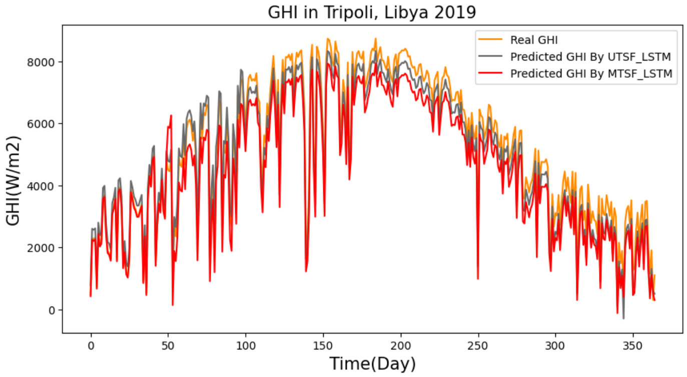


Figure 4.8: Measured GHI Vs. Predicted GHI of LSTM Models for Validation Data

the proposed models' performance, a dataset from a region characterized by substantially different climatic conditions was utilized. Daily Solar irradiation for 2019, sourced from U.S. NREL/NSRDB, has been used for validating the models. Data for Tripoli, Libya, were downloaded at coordinates (latitude 32.89 N and longitude 13.14 E). From Figure 4.7, the corresponding RMSE of the four proposed models can be seen. As shown in Figure 4.8, LSTM models display good generalizability and a stable performance when interaction with new data. This indicates that, in addition to memorizing training data, LSTM also had a fairly good comprehension of the patterning within the data. These results indicate that LSTM has an overall high performance that features good generalizability and few forecasting errors when faced with data from a variety of regions.

4.1.4 Summary and Recommendations

This study looked at four forecasting strategies, which were evaluated based on both their performance and their ability to create accurate forecasts for day-ahead solar irradiance. In the experiments, RBFNN and LSTM have been used to explore how exogenous and endogenous variables impact forecasting performance. The results clearly indicated that the developed UTSF_LSTM model showed the best capabilities in learning long-term patterns through the techniques of capturing consecutive hours as well as through long-term learned dependency (e.g., seasonality) behavior. On the other hand, RBFNN gave a better performance utilizing exogenous instead of only endogenous variables in the MTSF_RBFNN model. Overall, our results clearly demonstrate the four models' generalizability to give quite accurate daily predictions for solar irradiance. The results have also been validated with data from a region that features different climatic conditions from those originally tested. Overall, the outcome of these investigations clearly indicates the superiority of the proposed UTSF_LSTM method when compared to the UTSF-RBFNN, MTSF_RBFNN, or MTSF_LSTM developed models with regard to R^2 and RMSE. Suggestions for future work include:

- Further testing of the four models' stability using big data,

- Optimizing feature selection techniques as this would allow the models to capture different associations and learn additional nonlinearity behaviors related to solar irradiance and meteorological variables.

4.2 Model 2: Optimized Feature Selection Technique

Exogenous and endogenous variables are typically evaluated several times during the selection trial of a predictive model for 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 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 RFR and as data perturbation in LSTM.

4.2.1 Weather Recursive Feature Elimination (WRFE)

After implementing comprehensive data analysis in section (3.4) to the response and explanatory variables, it can be concluded that six features (CGHI, DHI, SZA, DNI, RH and T) appear to have the closest relation with the target (GHI). Even so, it is needed to consider our previous finding, which was that CGHI and SZA (along with DP and T) are redundant attributes. Here, DP can be eliminated, 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, it is needed 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, it is simply needed to look at the mutual information and correlation coeficients 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, a subset evaluation to determine relevance is needed to perform. This can be done using a hybrid feature reduction method that utilizes 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.

4.2.2 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 phase I: data perturbation for the LSTM model as shown in 4.9, and phase II: impurity measure (variance reduction) for the RFR model as shown in 4.10. 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 4.11 demonstrates the flowchart of the proposed WRFE, with algorithm 1 showing the pseudocode of the proposed procedure.

4.2.3 Model Implementation

Accordingly, the feature importance for the Halifax, NS, dataset for the year 2000 is measured. The dataset includes the six above-mentioned features. As the first step of algorithm 1, an LSTM model as establishing in the Model 1 [18] is designed

Algorithm 1 Proposed WRFE Method

```
Input: a data set of n features $M(F_1, F_2, \ldots, F_n)$
Output: Optimal feature subset $M_{best}$
Phase I- Modeling LSTM
Design LSTM utilizing the n features
Calculate RMSE1 (RMSE for full model performance)
for $i \leftarrow 1, n$ do
    Eliminating least importance feature $f_i$
    perturbing the n features
    $f_1 + d_1, f_2 + d_2, \ldots, f_n + d_n$
    Calculating perturbed prediction error RMSE2 from the perturbed dataset $M(F_{pn})$
    Calculating the drop in the performance
    $E = |RMSE2| - |RMSE1|$
    if $E <$ threshold then
        remove
$f_i$  else
        keep $f_i$
        $M_{best} \leftarrow M_{best} + f_i$
    end if
end for
Phase II: Modeling RFR
Design RFR utilizing the n features
Calculate RMSE1 (RMSE for full model performance)
for $i \leftarrow 1, n$ do
    Eliminating least importance feature $f_i$
    Measuring impurity of the n features (variance reduction)
    $var(\{f_1, f_2, \ldots, f_n\}) = \sum_{i=1}^{n} \frac{|f_i|}{|M(F_1, F_2, \ldots, F_n)|} var(f_i)$
    Calculating prediction error RMSE2 from the new dataset $M(F_{nn})$
    Calculating the drop in the performance
    $E = |RMSE2| - |RMSE1|$
    if $E <$ threshold then
        remove
$f_i$  else
        keep $f_i$
        $M_{best} \leftarrow M_{best} + f_i$
    end if
end for
```

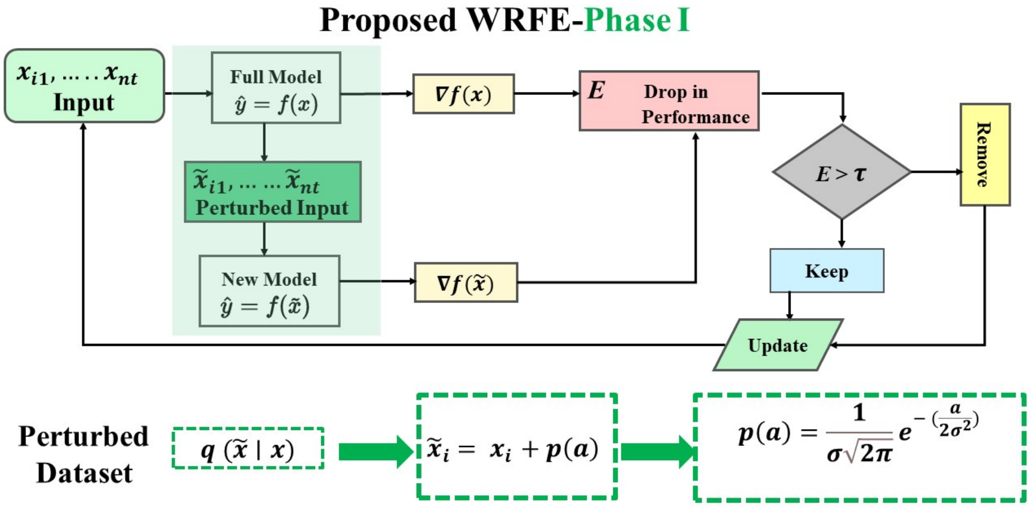


Figure 4.9: Proposed WRFE-phase I

and trained. Next, as stated in [75], each feature is perturbed by adding the random Gaussian distribution noise (mean $\mu = 0$, standard deviation $\sigma = 1$), with probability function p(a) as defined in Equation 4.3, and then the perturbed prediction is calculated.

$$p(a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{a}{2\sigma^2}} \tag{4.3}$$

The Euclidean distance (d) between original feature (x_i) and perturbed feature ($\tilde{x}_i$), as defined in 4.4

$$d(x_i, \tilde{x}_i) = \left(\sqrt{\sum_{i=1}^{n} (x_i - \tilde{x}_i)^2} \right) = 0 \tag{4.4}$$

When that is completed, the perturbation effects in gradients are measured. This is done by calculating RMSE for the perturbed and original forecasts. In the case, the calculation is given via the gradient values that 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 4.2). Also an RFR model is designed and trained in order to 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

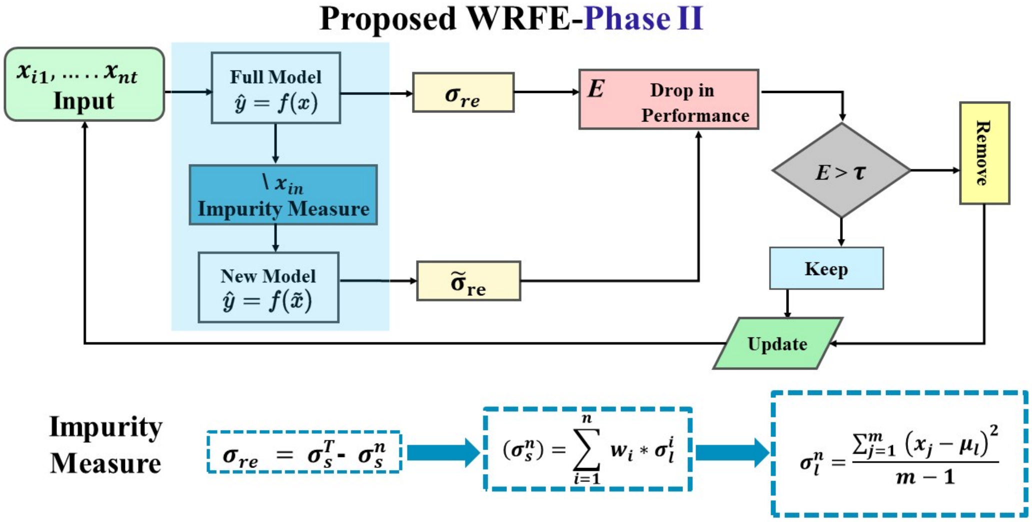


Figure 4.10: Proposed WRFE-phase II

the variance reduction criteria [76]. The designed forecast model is an ensemble of T decision trees, each comprising split and leaf nodes, as inspired by those proposed in [77][78]. Each split node (s) consists of a feature F_n and a threshold τ. The variance for every single leaf node (l_n) that is related to a particular split node is calculated as given in Equation 4.5:

$$\sigma_l^n = \frac{\sum_{j=1}^{m} (x_j - \mu_l)^2}{m-1} \tag{4.5}$$

Where n denotes the number of leaf nodes, m is the number of data points.

The variance reduction is computed by subtracting the variance of each (S) from the weighted average variance of the combined variance of the leaf nodes as given in Equation 4.6:

$$w(\sigma_s^n) = \sum_{i=1}^{n} w_i \ast \sigma_l^i \tag{4.6}$$

where w_i denotes the weight applied to l_n 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 Table 4.3). Subsequently, input data is projected 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

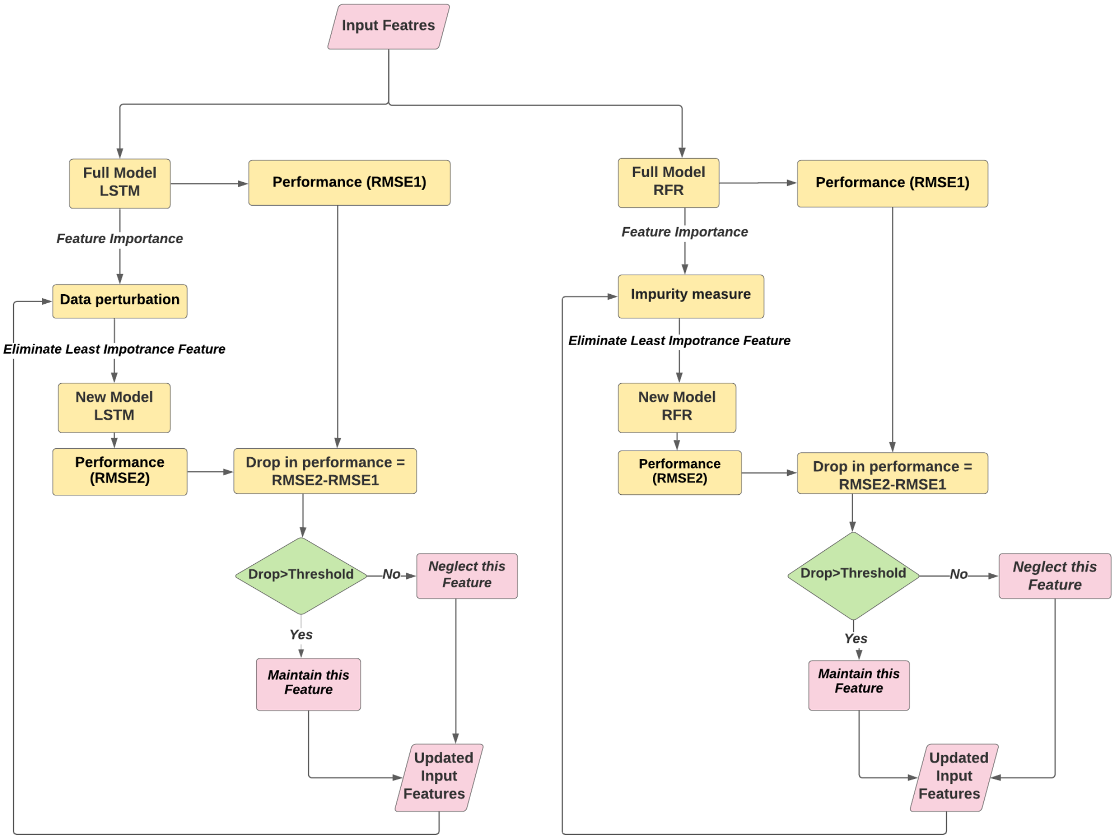


Figure 4.11: Flowchart of Proposed WRFE

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 4.4 shows the hyperparameter values for the proposed LSTM and RFR models, while Figure 4.12 and 4.13 demonstrate the inspection of feature importance according to data perturbation and variance reduction.

4.2.4 Forecasting Results and Analysis

LSTM models are employed to discover seasonality pattern of the previous 24 hours for the respective input features. These patterns are then utilized for predicting

Table 4.2: 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 4.3: 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

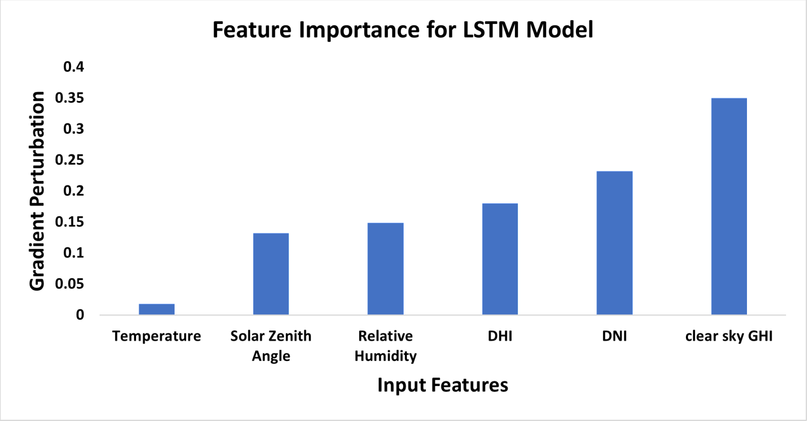


Figure 4.12: Feature Importance for LSTM

Table 4.4: 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

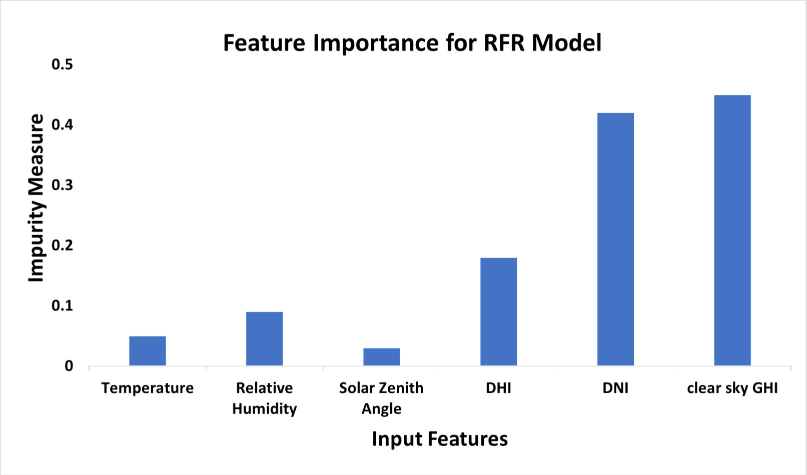


Figure 4.13: Feature Importance for RFR

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 and MBE values are utilized for performance verification of the designed models that belong to weather data for different locations in Canada. The ability of enhancing the usage in RMSE and MBE (obtained from Equations 4.7 and 4.8) 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 [79]. As shown in Table 4.5, the performance of the models shows a verified result where the t-statistic values (obtained from Equation 4.9) of the four models are less than the critical t-values.

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} (GHI_{i,\ pre} - GHI_{i,\ meas})^2 \right)^{\frac{1}{2}} \tag{4.7}$$

$$MBE = \frac{1}{N} \sum_{i=1}^{N} (GHI_{i,\ pre} - GHI_{i,\ meas}) \tag{4.8}$$

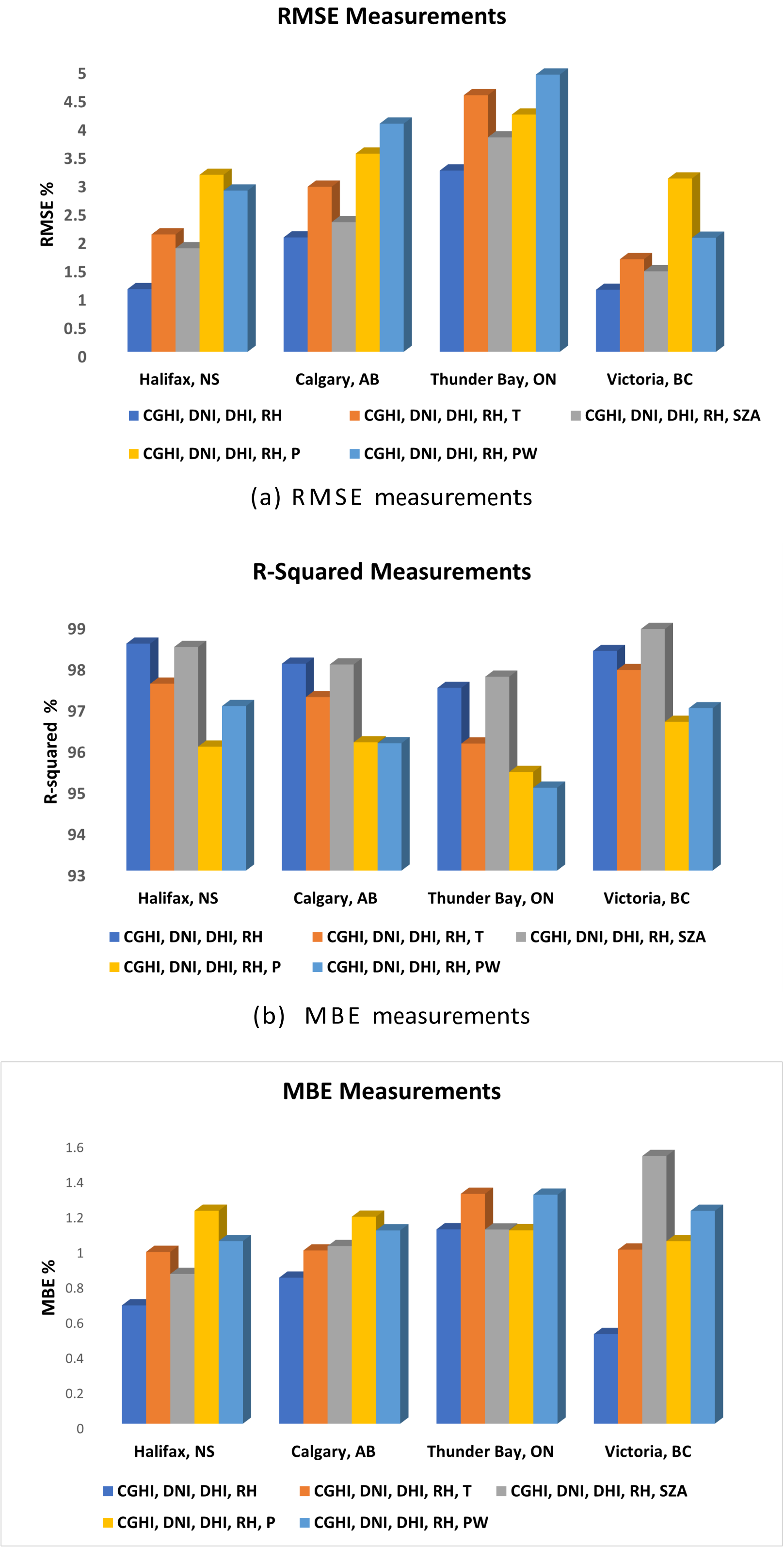
$$t = \left(\frac{(N-1)MBE^2}{RMSE^2 - MBE^2} \right)^{\frac{1}{2}} \tag{4.9}$$

From Figure 4.14, as can seen 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). After filtering out the variables using combined usage of different correlation techniques, the final optimal variables are selected during the selection trial based on mathematical characterises including impurity measures and perturbation theory. When separately adding exogenous variables to the LSTM models, as can seen that the RMSE between observed and estimated GHI is affected. Table 4.5 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 [19] [20]. These changes warrant future investigation of seasonality patterns in the weather data through capturing nonlinearity patterns embedded in the exogenous and endogenous variables.

4.2.5 Model Validation

Comparison of Different Feature Selection Techniques

In this section,the performance of the proposed WRFE is compared with other feature selection methods that explained in details (see section 2.2), including the VIF analysis proposed in [41], the ARD method proposed in [56], the NGA proposed in [57], the Pearson correlation analysis followed by the subset evaluator proposed in [58], and the subset evaluator proposed in [37]. As seen in Table 4.6, the proposed WRFE with CGHI, DNI, DHI, and RH as input features yield the lowest RMSE values. The proposed forecasting approach shows lower forecasting errors than the



(a) RMSE measurements

(b) MBE measurements

(c) R^2 measurements

Figure 4.14: Performance Comparison of Proposed WRFE with Several Sets of Input Features for Different Locations in Canada

Table 4.5: Performance Comparison of the Proposed WRFE for Different Locations in Canada.

Locatin	Input featurs	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	

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.

4.2.6 Summary and Recommendations

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. This work 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 nonmonotonic endogenous variable changes. It has been found that while variables might be non-predictive in isolation, they can be highly predictive in combination

Table 4.6: Performance Evaluation of Proposed WRFE vs Other Feature Selection Approaches

Ref.	Location	Feature selection Technique	Optimal features	Forecasting Model	RMSE(%)
Ref[41]	Turkey	VIF analysis	M, R_a, T_{mean} , RH_{mean}	ANN,ANFIS MLR	1.650
Ref[56]	Southeast of Spain	ARD	DOY, K_t^{ref}	ANN	5.2
Ref[57]	Argentina	NGA	temperature,humidity pressure,sunshine hours	LR	2.3
Ref[58]	India	Pearson correlation Subsets Evaluator	temperature humidity,pressure sunshine hour, wind speed	HMM with GFM	7.9124
Ref[37]	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

with others. For example, RH showed a weak association to be considered as a relevant attribute to GHI prediction model, however, it showed an improved predictive performance when combining with other attributes. This finding led us to perform a subset evaluation to determine relevance using the proposed novel hybrid feature reduction method, WRFE. Our aim was to optimize feature selection according to a Least-Redundant/Highest-Relevant framework, with feature importance measur-ing 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, the study shows that the high variability of irradiance conditions lowers the reliability of the training subset, as most deviations in the correlation coeficients occurred in the yearly dataset. From the observations of the historical data, it was noticed that GHI shows clear seasonal patterns through seasonal differences in solar irradiance. This may warrant further investigation of seasonality effects.

4.3 Model 3: Seasonal Clustering Technique for Hourly Solar Irradiance Forecasting

The main purpose of the study is to reach a forecasting accuracy level through layering and stacking clusters of weather data to reduce seasonality - related uncertainty. To enhance forecasting accuracy, high dimensional heterogeneous weather data should be added to training datasets; yet, weather data is deemed to have seasonality - related uncertainty. Since utilizing LSTM model does not achieve the required results, LSTM forecasting model performance deteriorates as a result of the change from univariate to multivariate analyses. Therefore, a primary investigation was conducted to analyze seasonality-based hourly predictions for GHI. The investigation found that seasonality affects accuracy of predictions due to high levels of autumn- and winter-related weather phenomena and climate uncertainty. Accordingly, SCFT based on an LSTM hybrid strategy and stacked layers of weather clusters was proposed.

4.3.1 Utilizing LSTM Models in Seasonality-Based Hourly Solar Irradiance Forecasting

Experimental Setup

Four experiments are carried out to forecast hourly solar irradiance readings for the 2000-2018 time-frame in Halifax, NS. The experiments are conducted using LSTM models that are designed with the ability to recognize patterns within the four seasons of the year (Spring, Summer, Autumn, and Winter). The forecasting models are trained and tested with hourly observation data for 2000-2010 and 2011-2018 respectively. In the first model (M1), the training dataset has 8,928 hourly observations for the recorded data points. The observations are from 8 am to 4 pm during Winter (December, January and February) and include the optimal feature set. The testing dataset has 6,550 hourly observations. Table 4.7 presents both the training and testing datasets used in the four seasonal models. The forecasting models are developed using Keras in TensorFlow 2.0. Table 4.8 presents the hyperparameter values of the newly developed LSTM models. The Keras Tuner library in TensorFlow is used to tune hyperparameters and find the optimal set for the application(part of the Code shown in Appendix B).

Table 4.7: Training and Testing Datasets Utilized in the Four Seasonal Models

Prediction Models based on Seasonality	Seasonality Dataset	Training Dataset Data from (2000-2010)	Testing Dataset Data from (2011-2018)
Model (M1)	Data recorded from 8 am- 4 pm in Winter seasonal (Dec, Jan, Feb)	8928 observations	6550 observations
Model (M2)	Data recorded from 7 am- 6 pm in Spring seasonal (Mar, Apr, May)	13167 observations	9631 observations
Model (M3)	Data recorded from 5 am- 7 pm in Summer seasonal (Jun, Jul, Aug)	14651 observations	10726 observations
Model (M4)	Data recorded from 7 am- 5 pm in Autumn seasonal (Sep, Oct, Nov)	10661 observations	7808 observations

Table 4.8: Hyperparameter Values for LSTM Model

Hyperparameter	Values
Learning Rate	0.001
Batch Size	24
Optimizer	Adam
No. of Epochs	100
Input Shape	3-D
No. of Hidden layer	3
No. of Units in each Hidden Layer	50
No. of Units in Output Layer	1
Dropout Rate	0.2

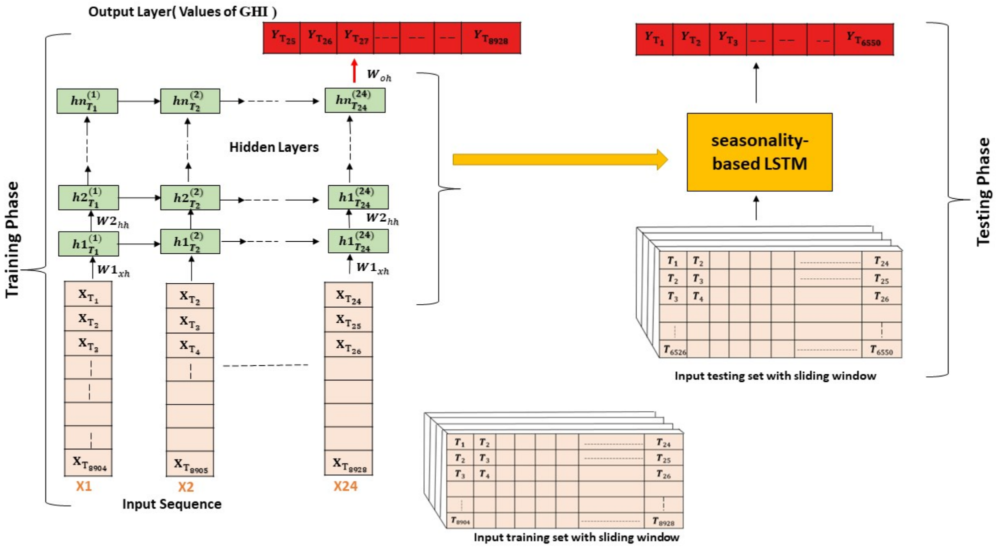


Figure 4.15: The Training and Testing Phases of The Proposed Seasonal Models

Architecture of the Seasonal Models

The timesteps used in the models are sequencing lengths of 24 hours, which is applied at each time-point, T. Further, LSTM considers values for the optimal feature set taken from the previous timestep before T. In so doing, LSTM is able to capture variations in daily readings, which it will then use to capture seasonality patterns in weather data to predict subsequent GHI for time (T+1). The training set is divided into two subsets (input and output) as shown in Figure 4.15. Input training considers adjacent 24 GHI values reported for the optimal feature set at time T. In so doing, it creates a 3-D tensor (N, T, D) for the observation number (N) as well as for the sequence length (T) and the number of input features (D), giving (8904, 24, 5), (13143,24,5), (14627,24,5) and (10637,24,5) for M1, M2, M3 and M4, respectively. The GHI output training set at size gives (8904,1), (13143,1), (14627,1), and (10637,1) for M1, M2, M3 and M4, respectively.

4.3.2 Seasonal Clustering Forecasting Technique (SCFT)

Design of k-means Clustering Algorithm

Clustering analysis as a concept demarcates and clusters data points according to similar characteristics. The proposed data aims to find weather types and patterns in a specific location. Here, a k-means clustering algorithm has been employed, with the elbow method being used to find the optimal number of clusters (k). This represents an optimization problem, and the objective function, as given in Equation 4.10, requires the sum of squared errors (SSE) to be minimized among observations of every cluster as well as every cluster's centroid [80]. The process then iterates through k, indicating the k value with the respective SSE. Then, with increases in cluster numbers, there is a slow flattening of the curve at the same time as SSE exhibits a rapid decrease. Note that the resulting optimal range of cluster numbers for k is 5 to 15, as these denote inflection points for the curve and determine the k value. In the present application, a reasonable number is 5 clusters as clearly shown in Figure 4.16. Let $D = (x_1, x_2, \cdots, x_n)$, x_i ⍰ R^d, The optimization problem is formulated as follows:

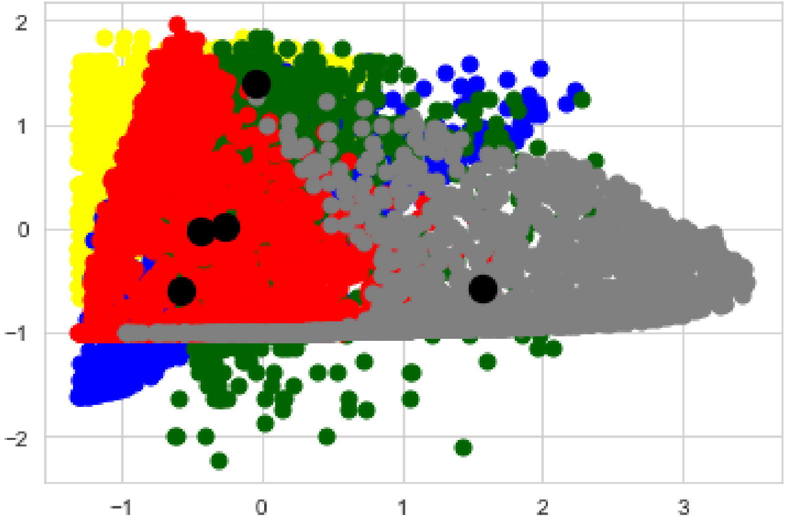


Figure 4.16: Five Clusters of Data Points with Respective Centres

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{i,j} \left\| x^{(i)} - \mu^{(j)} \right\|^2 \tag{4.10}$$

Minimize J subject to

$$\begin{cases} \sum_{i=1}^{n} a_{i,j} \geq \tau_j, & j = 1, \ldots, k \\ \sum_{j=1}^{k} a_{ij} = 1 \text{ for } i = 1, \ldots, n \\ a_{i,j} \geq 0, i = 1, \ldots, n, j = 1, \ldots, k \end{cases}$$

where $\sum_{j=1}^{k} \tau_j \leq n$, k is the number of clusters, μ_i is the cluster centroid, and $\mu_j \in R^d$, $\| \cdot \|$ is the Euclidian norm.

A Python Seaborn library pair plot is used for plotting multiple pairwise bivariate distributions in multivariate datasets. As shown in Figure 4.17, the 13×13 plot matrix illustrates 13 standardized features that have been paired. Note that the diagonal elements indicate univariate plots for kernel density estimation (KDE) distributions that correspond to individual features. These denote a continuous probability density curve of all the data points for every individual feature in the 5 clusters. In contrast, the off-diagonal elements represent bivariate plots for scatter distributions that correspond to two individual features. Scatter plots display that there is some distinct grouping of the point, which indicates that there is some clustering occurring. Figure

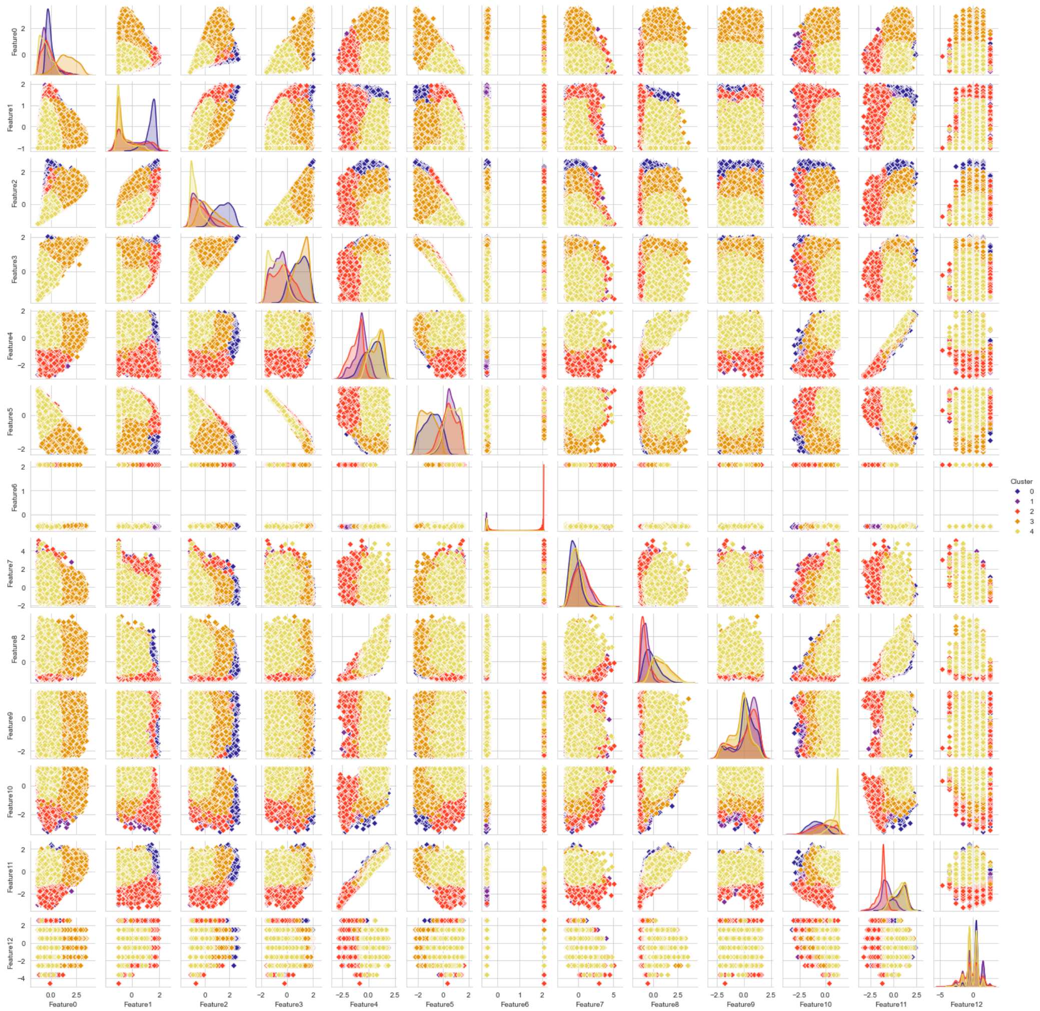


Figure 4.17: Standardized Data Point Pair Plot, with Associated Clusters Colored the Same in each Data Point

4.18 is a zoomed-in portion of Figure 4.17 showing clear clustering in the 2nd (DNI), 3rd (GHI), 4th (CGHI), and 6th (SZA) features.

Climatological Criteria-based Threshold Selection

Based on the analysis, critical conclusions may be drawn regarding the unique characteristics of each cluster. The variations within the 5 clusters act as a natural dataset divider for weather types and groups. In the proposed application for forecasting the performance of PV systems, the behaviour of historical meteorological characteristics based on sky conditions will be examined, with three weather types/groups

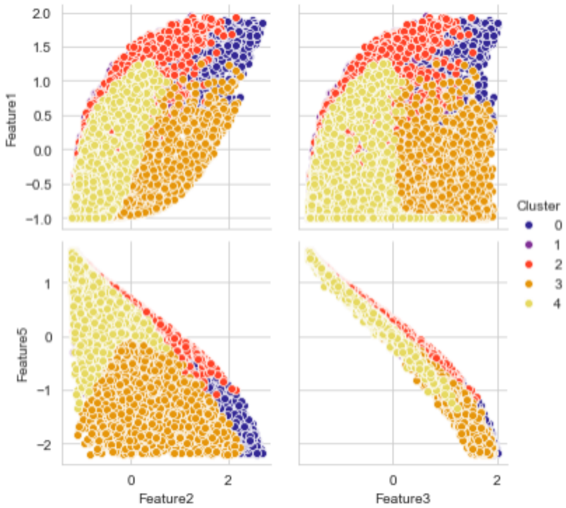


Figure 4.18: Zoomed-in Portion of The Plot Matrix Showing Clustering in DNI, GHI, CGHI, and SZA

being chosen (sunny, cloudy, and rainy). The threshold cut-off criteria of the correlated standardized data points were chosen according to meteorological and weather aspects that fundamentally affect atmospheric conditions. Consistent with this, beyond Earth's atmosphere, and considering the mean solar distance and beam irradiance (i.e., solar constant; W/m^2, irradiance decreases when sunlight passes through Earth's atmosphere. The decrease is caused by absorption and reflection activities. Hence, typical total irradiance at Earth's surface is around 700 to 1,300 W/m^2 at solar noon and in cloudless conditions. Season, altitude, and latitude will also affect this reading [81]. In considering the above, for the first criteria (cloudless skies), there is a decrease in DHI irradiance and an increase in GHI. This means that clusters representing sunny observations occur when GHI exceeds 600 W/m2 and DHI is below 100 W/m^2. Atmospheric pressure is also a useful weather pattern indicator because of the subsidence phenomenon. According to this phenomenon, low-pressure systems generally denote warmer temperatures as well as precipitation (mainly rain) events, whereas high-pressure systems typically denote cooler temperatures with clear skies and no precipitation (i.e., low humidity). In our second criteria, dataset components such as temperature, precipitation, relative humidity, and air pressure are used to denote weather groups that are rainy or cloudy. As Table 4.9 illustrates, the minimum, middle, and maximum values of the dataset indicators are presented, along

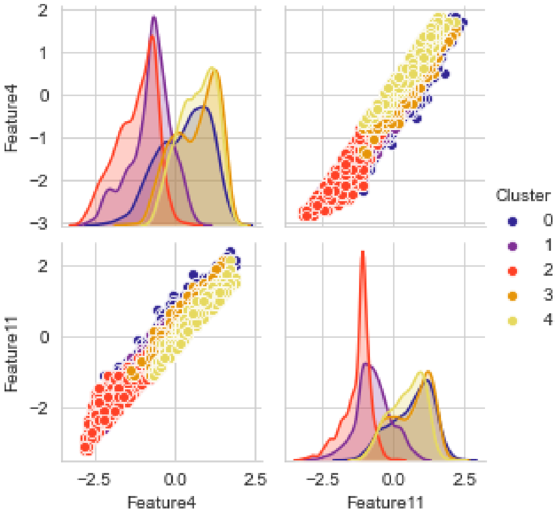


Figure 4.19: Linear Relationship Between DP and T

with corresponding standardized values. Note that each group's cut-off points have also been determined based on some of these values. With close investigation, cluster 4 (index 3) data point provides the best representation of sunny day observations with higher GHI and lower DHI. As well, it offers meteorological readings for other features typical of sunny weather (e.g., low precipitation, fewer clouds, low air pressure, etc.). Cluster 1 and cluster 2 as showing samples typical of low air pressure and temperature. Verification also shows features such as high cloud type, which indicates high planetary albedo consistent with evidence of cloudy weather. Clus-ter 3 and cluster 5 indicate that lower air pressure along with higher precipitation and dew point typically characterize rainy weather. Rain is usually accompanied by clouds and large masses of water vapor, with high humidity referring to the amount of water vapor in the air. Cluster 2 verifies this fact, with high relative humidity being higher in the cluster 2. Subplots (4,11) is also considered, in which the data points are generated depicting the 5th (DP) and 12th (T) features, while visually analyzing the standardized data points for these two features (DP and T) only. It is clear that the distribution, as shown in Figure 4.19(some of the numerical representations of the clustering results shown in Appendix C), provides valuable insight that there is a linear relationship between the two mentioned features.

Table 4.9: Indicators of Cut-Off Points

Feature	DHI	DNI	GHI	CGHI	DP	SZA	SA	WS	PW	WD	RH	T	P
count	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04	4.7407e+04
mean	1.5704e-15	3.6624e-15	-2.5461e-17	-1.3616e-16	1.9011e-16	4.5881e-16	-2.3888e-15	-1.4396e-16	1.3349e-16	3.3251e-16	-4.1198e-16	4.3763e-15	-1.6606e-15
std	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00	1.0000E+00
min	-1.3110e+00	-9.9036e-01	-1.2030e+00	-1.6358e+00	-2.9629e+00	-2.2071e+00	-5.5146e-01	-1.8775e+00	-1.4814e+00	-2.2669e+00	-3.4412e+00	-3.4916e+00	-5.6610e+00
25%	-7.0498e-01	-9.6661e-01	-8.3761e-01	-8.4149e-01	-6.7904e-01	-7.5304e-01	-5.0007e-01	-7.3718e-01	-8.2491e-01	-5.9502e-01	-7.4013e-01	-8.5529e-01	-5.1182e-01
50%	-2.7350e-01	-3.6693e-01	-2.7567e-01	-8.4074e-02	4.2171e-02	1.6684e-01	-4.7952e-01	-1.2901e-01	-1.9695e-01	1.2416e-01	1.1716e-01	2.3497e-02	5.1801e-01
75%	4.2506e-01	9.7490e-01	6.4778e-01	8.4700e-01	8.8359e-01	8.0447e-01	-4.3498e-01	5.5518e-01	6.4059e-01	7.8494e-01	9.2976e-01	9.0229e-01	5.1801e-01
max	3.4659e+00	1.9961e+00	2.7619e+00	2.0810e+00	1.9654e+00	1.5933e+00	2.0590e+00	5.2685e+00	4.1372e+00	1.6263e+00	1.2178e+00	2.4088e+00	2.5777e+00

SCFT Methodology

SCFT proposed in this work has been developed to apply a hybrid approach to the LSTM models. First, the SCFT optimizes input feature selections for the models and then clusters the relevant meteorological data points according to sky conditions for 1-hour-ahead GHI forecasting. Next, the SCFT filters the clustering data points so that they include optimal features only, as illustrated in Figure 4.20. The training and testing methods used in the SCFT are expressed in the algorithm 2:

- Step 1: Obtain the highest individual relevant features by measuring monotonic associations between endogenous and exogenous variables.
- Step 2: Concatenate the training dataset for the four seasonal models (M1, M2, M3, M4).
- Step 3: Project the concatenated seasonal data points for selected features as well as for the remaining features in order to conduct a cluster analysis.
- Step 4: Investigate historical weather data-point behavior according to sky conditions, using as weather groups the three clusters of sunny (MC1), cloudy (MC2), and rainy (MC3).
- Step 5: Filter the clustered data points associated with the optimal features to design the stack layer.
- Step 6: Train the proposed LSTM models using the clustered training datasets (MC1, MC2, and MC3) and concatenate the resulting output predictions.
- Step 7: Test LSTM1, LSTM2, and LSTM3 against the concatenation of the M1, M2, M3, and M4 datasets in order to forecast hourly GHI values.

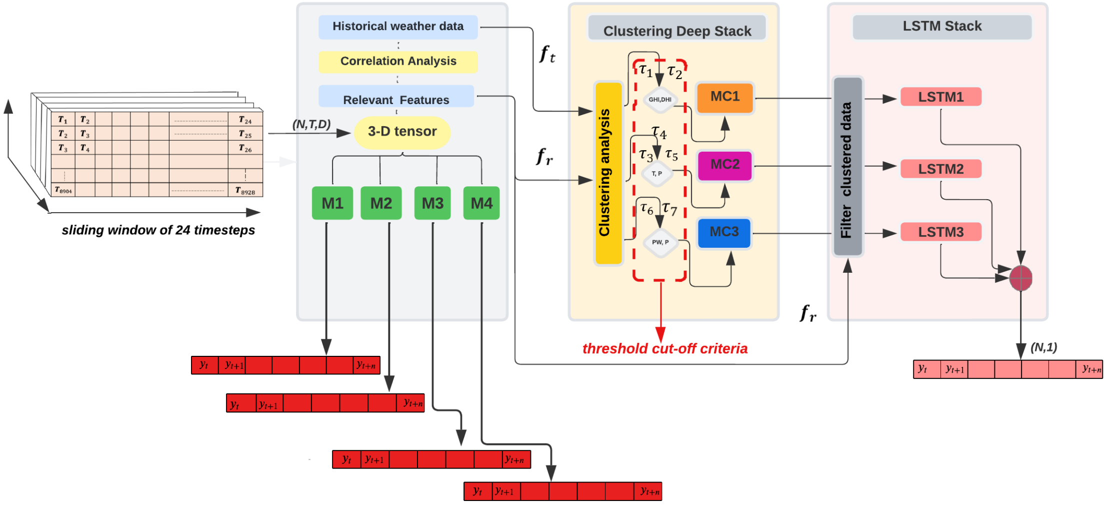


Figure 4.20: Detail Workflow of SCFT

Algorithm 2 Proposed SCFT Method

```
Input: historical weather data for t features $f_t$
Obtain highest relevant features $f_r$
Compute standardized datapoints for $f_t(n)$ and $f_r(l)$
Classify the data to Seasonality Datasets
Splitting them into Input training, output training, and testing sets.
Reshape the Seasonal datasets for M1, M2, M3, M4 as shown in Fig1
for i ← 1, N = 4 do
    Train $M_i$ with respective datasets
    Compute the performance for each model
end for
Concatenate training dataset for $f_r$ (M1), $f_r$ (M2),$f_r$ (M3), $f_r$ (M4)
Read $f_t$
Structure clustering datasets MC1,MC2, MC3
Setting of threshold cut-off criteria τ for specify weather indicators
for j ← 1, n do
    if GHI ≤ $\tau_1$ & DHI ≤ $\tau_2$ & (j ⍰ $f_r$) then
        Add j to MC1
    else if T ≤ $\tau_3$ & P ≤ $\tau_4$ & P ≤ $\tau_5$ & (j ⍰ $f_r$) then
        Add j to MC2
    else if PW ≤ $\tau_6$ & P ≤ $\tau_7$ & (j ⍰ $f_r$) then
        Add j to MC3
    end if
end for
yt1: Train MC1 based LSTM1
yt2: Train MC2 based LSTM2
yt3: Train MC3 based LSTM3
yt: Concatenate yt1, yt2,yt3
```

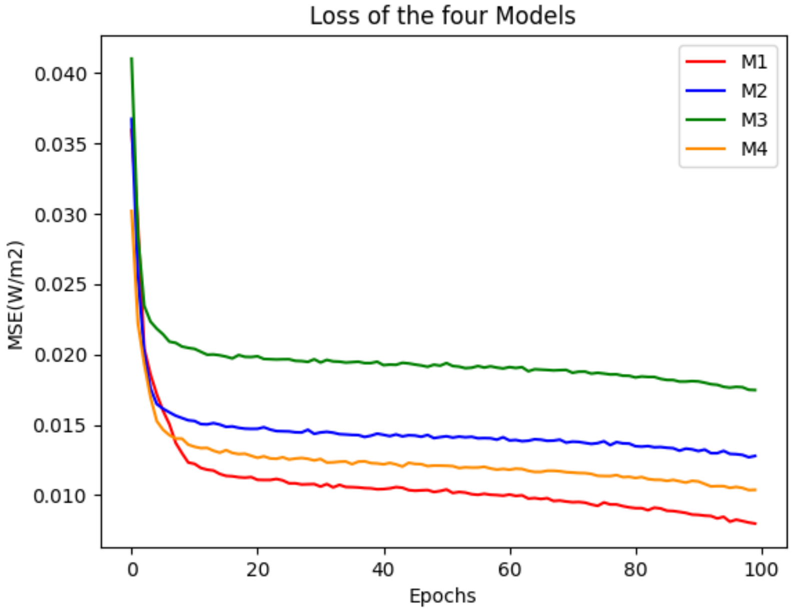
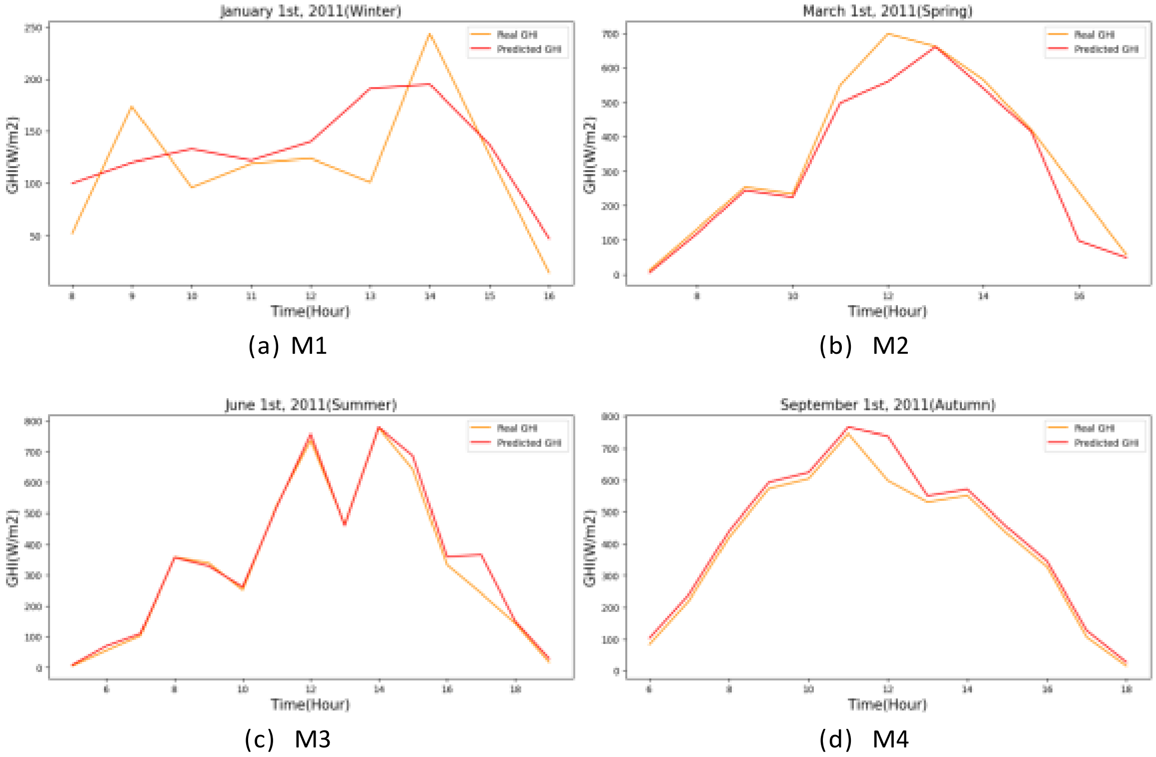


Figure 4.21: Loss of Proposed Seasonal Models

4.3.3 Forecasting Results and Analysis

The four proposed seasonal models are trained using hourly historical value batches. In this process, Mini Batch Gradient descent is applied for updating the weights for each batch of 24 values for weather data and GHI. Figure 4.21 depicts the learning curves for 100 epochs, showing the final loss values for the training task. The val-ues are approximately 0.0084, 0.0129, 0.0174, and 0.0101 for M1, M2, M3 and M4, respectively.

As shown in Figures 4.22 (b) and (c), the predicted GHI values and the actual values are quite close for Spring and Summer. Closer observation reveals only a slight deviation for Spring, as depicted in Figure 4.22 (b). However, for Winter and Autumn as shown in Figures 4.22 (a) and (d), the LSTM model over-predicts the mid-day hourly GHI values, though in the early and late hours of the day, the values are again in close agreement with actual values for Autumn. For Winter, the values are in close agreement only for the late hours of the day. Looking at these results, it can be speculated that the deviation from predicted hourly GHI in Autumn and Winter may be caused by climatic uncertainty as shown in Figures4.23. This is because ML results heavily depend on training datasets, and prediction is usually very dificult during this time of year due to the wide variety of weather phenomena.



(a) M1 (b) M2

(c) M3 (d) M4

Figure 4.22: Predicted Hourly GHI for The Four Seasonal Models

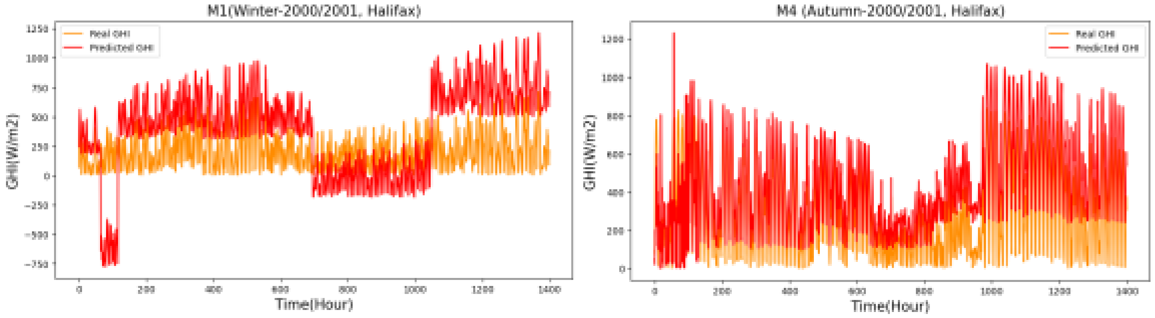
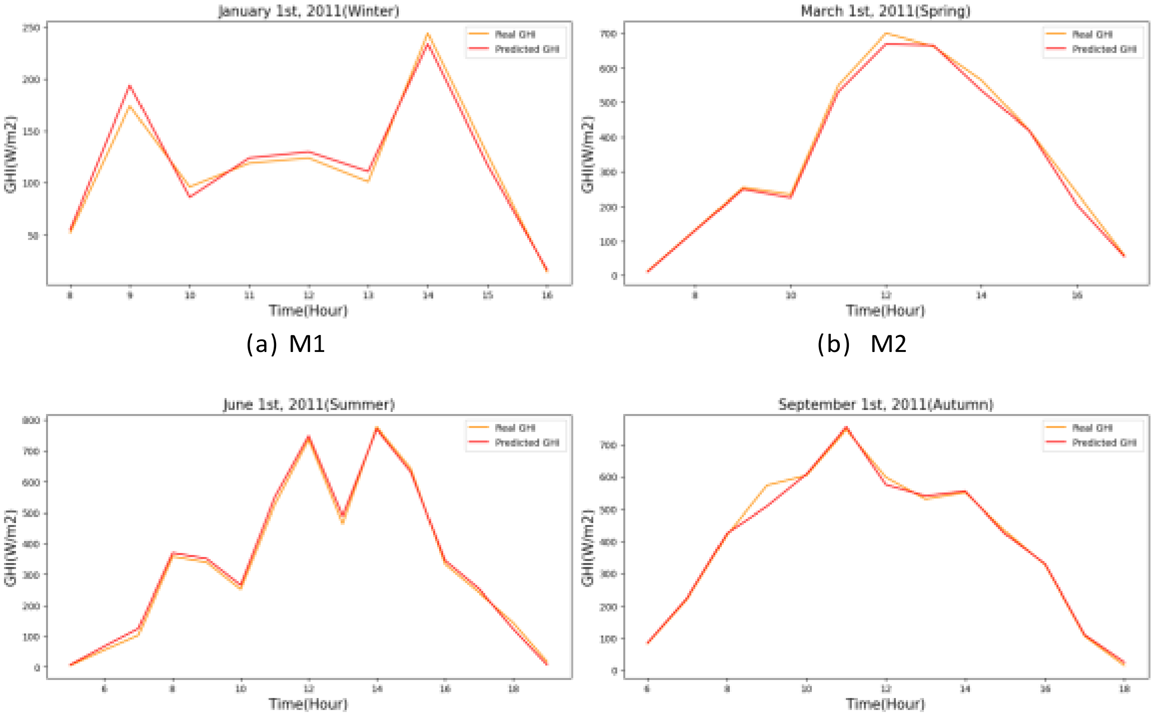


Figure 4.23: Predicted Hourly GHI for Autumn and Winter Seasons

The proposed SCFT has excellent performance in relation to M1, with an RMSE of 13.06 W/m^2, as illustrated in Figure 4.24c. For Spring, the proposed SCFT has both stable and good performance, with an RMSE of approximately 15.53 W/m^2. However, as mentioned previously, the forecasting performance for the LSTM models is substantially reduced for rainy and cloudy days, whereas the proposed SCFT's performance is better, as depicted in Figures 4.24a and 4.24d, with the RMSE values of approximately 20.04 W/m2 and 18.71 W/m^2. The results indicate that the models have good forecasting performance. The results are especially impressive for M2 and M3, while M1 are M4 less impressive. Hourly GHI, as predicted in the proposed seasonal LSTM models, along with the actual hourly data for the first days of the months January, March, June and September, 2011, for Halifax, NS, are presented in Figures 4.22 (a)–(d), and after data clustering are presented in Figures 4.24(a)–(d). Figure 4.25 shows the effects of the proposed algorithm to reduce uncertainty associated with forecasting solar irradiance in comparison to Figure 4.23.



(a) M1

(b) M2

(c) M3

(d) M4

Figure 4.24: Predicted Hourly GHI for The Proposed SCFT

4.3.4 Model Validation

Forecasting Model Performance Using Data from Other Regions

The quality of the forecast has been shown to be strongly related to the location to which it refers. However, this typically, does not correspond to the location of the PV system. A more detailed analysis would therefore be needed to test this particular

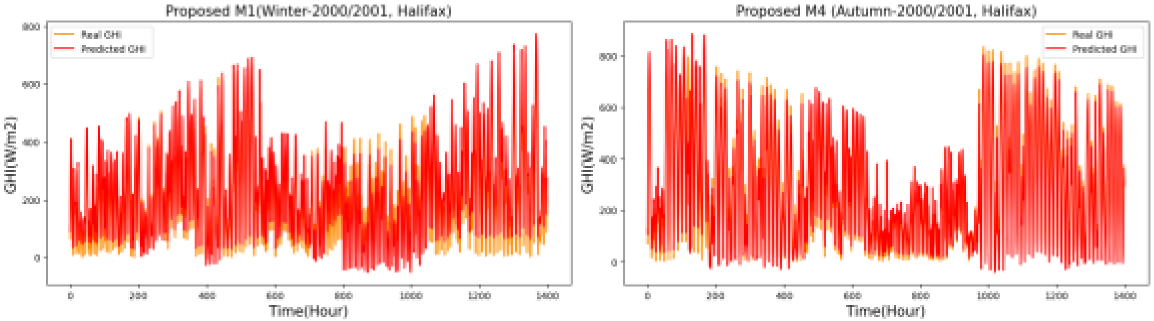


Figure 4.25: Predicted Hourly GHI for Autumn and Winter Seasons for The Proposed SCFT

Table 4.10: Köppen Climate Classification

City	Latitude and longitude coordinates	Köppen climate
	coordinates	classification
Halifax, Canada	44.6488° N, 63.5752° W	Warm humid continental climate “Dfb”
Tripoli, Libya	32.8872° N, 13.1913° E	Mid-Latitude Steppe and Desert Climate “Bsh”
Rome, Italy	41.9028° N, 12.4964° E	Mediterran Climate ”Csa”
Helsinki, Finland	60.1699° N, 24.9384° E	Warm-summer humid continental climate “Dfb”
Shanghai, China	31.2304° N, 121.4737° E	Humid Subtropical Climate) ”Cfa”
San Francisco, USA	37.7749° N, 122.4194° W	Warm-summer Mediterranean climate ”Csb”

hypothesis. In this context, the proposed method is evaluated for datasets collected from different climate regions according to the K¨oppen climate classification method [82], as described in Table 4.10. The RMSE and MAE values, along with their normalized values nRMSE and nMAE are listed in Table 4.11. The results show that the SCFT performance is consistently stable and that RMSE ranges between 13.48 W/m^2 and 17.052 W/m^2 for forecasting the highly fluctuating GHI in different climate condition locations. The study confirms the ability of the proposed technique for much more precise predictions in desert and Mediterranean climate regions, such as Tripoli, Rome, and San Francisco.

Comparative Analysis of the models

Other models in the literature are used to compare the forecasting of hourly solar irradiance. Table 4.12 presents the results of testing SCFT’s hourly GHI estimation in comparison to the models proposed in [13] [17][51][52][53], and [54]. As can be seen in the Table 4.12, the SCFT model outperforms all the others. The SCFT

Table 4.11: Performance Comparison of the Proposed SCFT for Regions with Different Climate Conditions

Location	Model	RMSE	nRMSE	MAE	nMAE	R^2
		(w/m^2)	(%)	(w/m^2)	(%)	(%)
Halifax,Canada	M1	40.2203	22.25863	38.2189	21.1535	90.4002
	M2	32.4523	17.9532	29.9823	16.593	95.1773
	M3	30.5309	16.8966	26.0309	14.4184	95.1698
	M4	40.4634	22.3929	35.9423	19.8972	94.0454
	SCFT	16.8212	9.3053	12.5307	6.9379	96.1238
Tripoli, Libya	M1	36.2119	20.0332	35.3134	19.5481	91.5437
	M2	30.1247	16.6643	25.8669	14.3105	90.9433
	M3	24.9804	13.8242	22.0924	12.2381	90.4383
	M4	33.5386	18.5536	28.6063	15.8374	91.0292
	SCFT	13.4811	7.4678	10.0452	5.5685	92.9373
Rome, Italy	M1	38.0223	21.0448	37.9352	20.9934	92.3464
	M2	30.3225	16.7896	26.0346	14.4170	92.9443
	M3	25.1343	13.9048	22.9736	12.7105	93.3838
	M4	35.7655	19.7985	30.1973	16.7136	93.2137
	SCFT	15.0998	8.3546	12.6163	6.9853	93.9844
Helsinki, Finland	M1	40.2621	22.2897	38.9509	21.5606	90.9537
	M2	33.0376	18.2753	30.4134	16.8323	92.0967
	M3	30.9735	17.1307	27.9108	15.4508	90.0437
	M4	44.2343	24.4763	39.7543	22.0038	91.0955
	SCFT	17.052	10.2425	16.4464	9.1053	92.7801
Shanghai, China	M1	39.6343	21.9353	36.6932	20.3132	93.8759
	M2	31.2069	17.2632	27.9032	15.4453	93.9835
	M3	28.1164	15.5575	24.7496	13.6927	94.0248
	M4	39.3295	21.7674	34.2333	18.9585	93.9735
	SCFT	16.0336	8.8735	11.8832	6.5848	94.8342
San Francisco, USA	M1	37.9085	20.9784	34.9574	19.3406	95.0945
	M2	32.0424	17.7339	28.6406	15.8504	94.0867
	M3	27.9647	15.4746	24.0323	13.3046	94.0756
	M4	36.3283	20.1075	31.1947	17.2684	95.0945
	SCFT	14.9903	8.29506	11.4307	6.3336	96.9834

Table 4.12: Performance Evaluation of Proposed SCFT vs Other Approaches

Ref.	Journal Publisher	Location	Contribution	Model	RMSE (w/m^2)
Ref[52]	IEEE Transactions on Industrial Informatics	Finland	Choquet integral	LSTM	17.95
Ref[53]	IEEE Access	Pretoria,South Africa	Hybrid technique	QRA	34.85
Ref[54]	IEEE Access	Hawaii	Clustered clearness index	LSTM	22.13
Ref[13]	Elsevier Energy	Santiago,Cape Verde	Feature selection	LSTM	76.245
Ref[51]	Elsevier Renewable Energy	Lamb,Texas,USA	Hybrid technique	CNN-LSTM	48.13
Ref[17]	IEEE Access	Denver,Colorado,USA	Hybrid technique	WT-ENN	25.83
Proposed SCFT	IEEE Transactions on Industrial Informatics	Libya,North Africa	Seasonal clustering	LSTM	13.48

model’s superior performance is likely due to the effectiveness of LSTM models in solar irradiance forecasting, and in the clustering strategy, which reduces uncertainty levels.

In addition to the proposed SCFT, the other model (M5) is designed by implementing the concept of applying the choquet integral to the four proposed seasonal LSTM models. The idea of the choquet integral-based LSTM forecasting models was introduced in [52] where the forecasting accuracy was improved as a result of the prediction aggregation of individual models through the fuzzy measure. The proposed SCFT is also validated by the Diebold-Mariano (DM) statistics test [83] and Granger-Newbold (GN) statistics test [84] [85] for predictive accuracy. In both tests the forecast accuracy of the two forecast methods are compared. The statistics of the DM test are formulated as follows:

- H_0 : E (d_t) = 0⍰t SCFT and M5 have the same accuracy.
- H_1 : E (d_t) = = 0⍰t SCFT and M5 have different accuracy.

where d_t is the loss differential of the SCFT and M5.The formulation of the GN test is as follows:

- H_0 : $r_{xz} > 0$ SCFT has a larger RMSE.
- H_1 : $r_{xz} < 0$ SCFT has a smaller RMSE.

where Let r_{xz} denotes the correlation coeficient between SCFT and M5. As a way of validation, the data of model [52] was entered into the proposed model, SCFT, and the results showed that the performance of SCFT outperformed the performance of model [52]. As shown in Table 4.13, the P-values for DM, which indicates that H_0 is significantly rejected and the forecasting accuracy of the two models is different.

Table 4.13: Statistics Tests for Comparing Forecasting Accuracy[1].

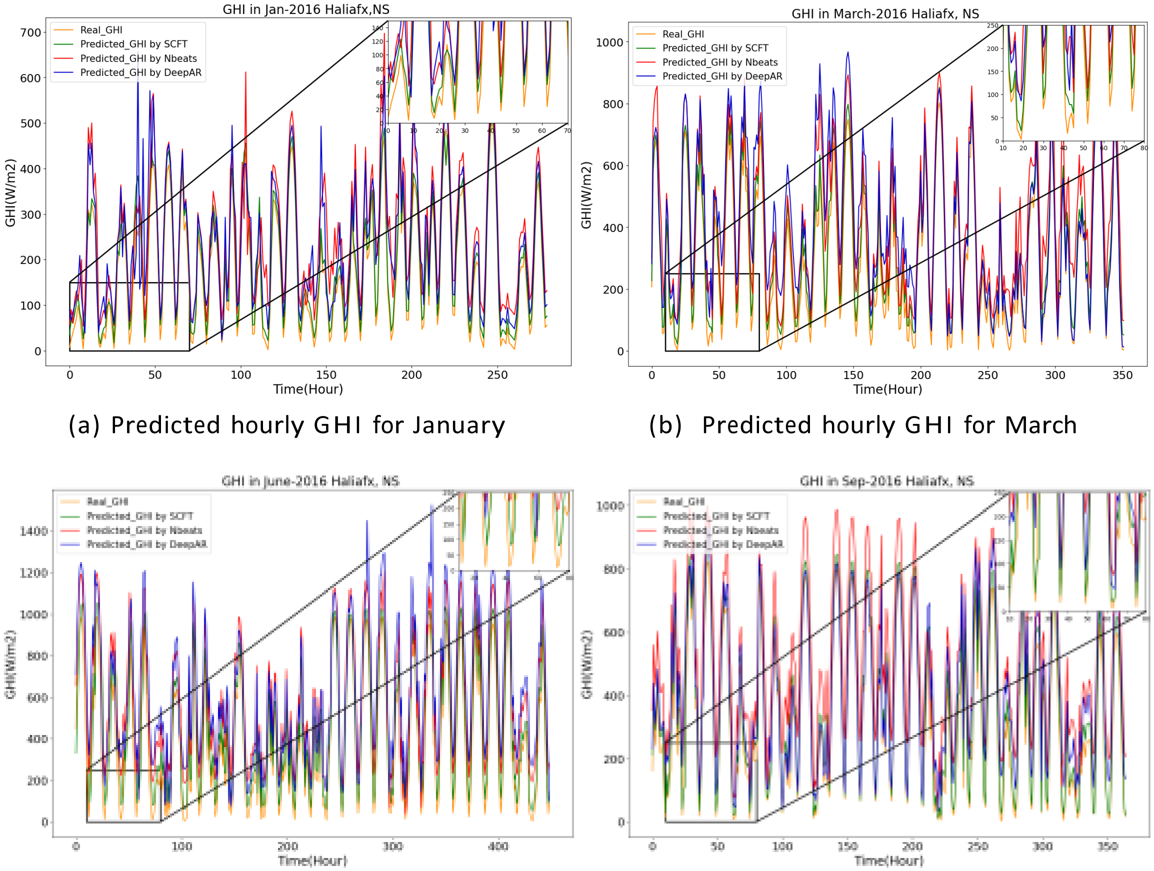
	Diebold-Mariano Test[2]		Granger-Newbold Test[3]	
	two.Sided test($\alpha = 0.0125$)		one.Sided test ($\alpha = 0.025$)	
Sample size	DM Statistic	P-Value	GN Statistic	P-Value
1200	2.596	0.00943	2.0032	0.02257
2400	2.6065	0.00917	2.0069	0.02238
3600	2.6534	0.00796	2.0702	0.01921
4800	2.769	0.00562	2.0781	0.01885
6000	2.9951	0.00274	2.1004	0.01784
7200	3.0299	0.00244	2.1093	0.01745
8400	3.289	0.001	2.1689	0.01504

[1] The DM and GN tests are designed for comparing predictive accuracy of two forecasting models (SCFT,M5) accounting for asymptotic data distribution. [2]The DM test uses two-tailed criteria for setting the alpha value, while [3]the GN test is a one-tailed statistical test.

Likewise, the p-values of the GN test indicates that we reject H_0 and accept the alternative hypothesis, and the proposed SCFT outperformed M5.

Comparison of SCFT with Baseline models

The proposed SCFT is compared with the state-of-the-art deep learning-based forecasting techniques such as Nbeats and DeepAR that are available in PyTorch Forecasting Documentation. Compared to Nbeats, LSTM-based proposed model (SCFT) adopts multivariate while the former adopts univariate. Further, for information processing, Nbeats employed deep stack of fully connected layers, whereas SCFT aggregated the data using mining clustering technique upon meteorological conditions that impact the radiance. With respect to DeepAR, it works on multivariate probabilistic forecasting with autoregressive based on recurrent networks- LSTM/GRU models. SCFT is the same as DeepAR, though with the clustering layer added. As shown in Figure 4.26 the predicted hourly GHI for four seasons for 2016, Halifax,NS. The results showed that the proposed SCFT is superior compared to the other models with average reduction of RMSE ranges between 6.14% - 9.43% .



(a) Predicted hourly GHI for January

(b) Predicted hourly GHI for March

(c) Predicted hourly GHI for June

(d) Predicted hourly GHI for September

Figure 4.26: Comparison of SCFT with Baseline Models for Different Months in 2016, for Halifax, NS

4.3.5 Summary and Recommendations

Most of the related analyses clearly show that model performance is highly dependent on training sets, and that seasonality strongly affects the accuracy of predictions due to high levels of autumn- and winter-related weather phenomena and climate uncertainty. In turn, the proposed SCFT introduced an eficient handling strategy for reliable forecasting by considering clustering the high variability of historical meteorological observations based on sky conditions. Three weather types (sunny, cloudy, and rainy) were chosen for clustering. The threshold cut-off criteria of the correlated standardized data points were chosen according to meteorological characteristics that fundamentally affect atmospheric conditions. To examine the stability of the proposed approach, it applied to forecast GHI for regions with different climate conditions. The proposed SCFT forecasting approach showed improvements in the learning tasks when training the LSTM model by enhancing the model's ability to identify patterns within a dataset. This formed the basis for dataset clusters and making predictions of the hourly GHI with suficient accuracy even for regions with highly fluctuating climates. The experimental results provided compelling evidence that the SCFT is superior to previously proposed approaches. Future smart infrastructure studies should take advantage of historical meteorological, atmospheric, and climatological data to forecast PV power generation for operating a secure and reliable smart system. However, as bias errors in GHI ,DHI, and DNI measurements could lead to corresponding errors with regard to the produced power. Solar irradiance data could be enhanced through a close investigation of the source of bias errors in data and how bias errors may be affected according to variables such as air mass, aerosols, altitude, zenith angle, and Linke turbidity factor.

Chapter 5

Conclusions and Future Work

This chapter presents in brief the study's main contributions. As well as it highlights some limitations uncovered during the design, implementation, and analysis of the models. These limitations could serve as suggestions for future research directions.

5.1 Conclusion

The recent exponential growth rate of PV systems has resulted in an increased need for greater PV power generation integration in the world's grid systems. Solar energy is characterized by its uncontrollable and intermittent nature, which makes it challenging to control and manage its produced power. This leads to several issues such as imbalances between the amount of produced and consumed energy, and regulating power to ensure a consistent supply range. Given these challenges, recent research has shifted on developing more precise predictive models for the produced power of PV systems. These models rely on an analysis of local weather and solar radiation data to more accurately predict the power generated by PV systems. Short-time horizon GHI forecasting is a crucial component in predicting PV power generation. The present study developed a robust model that can be used for an hour-ahead of GHI forecasting. The proposed approach was built in three stages, as follows: 1) the LSTM model performance was tested in relation to input feature (univariate or multivariate) influence; 2) feature selection was optimized by analyzing the impact of redundancy and relevancy measures; and 3) problems related to seasonality and its adverse affects on the model's performance were investigated.

In stage one, LSTM model behaviour was investigated under specific meteorological and geographical conditions in relation to historical meteorological data and solar irradiance. These exogenous and endogenous variables were then used for input features selection in an hour-ahead of solar irradiance prediction model, comparing it to models based on UTSF and MTSF. The performance of the LSTM forecasting model was reduced during the switch from univariate to multivariate analyses. However, forecasting accuracy was then shown to be enhanced by multivariate input in relation to explanatory and response variables.

The second stage of the thesis' model development presented a comprehensive correlation analysis aimed at optimizing feature selection. The purpose here was to permit the LSTM model to find and capture a broad range of associations while learning new nonlinearity behaviors for meteorological and solar irradiance variables. Based on these explorations, the WRFE technique was proposed as an optimizing feature selection schema, using a Least-Redundant/Highest-Relevant framework. The WRFE technique measures variance reduction in RFR and data perturbation in LSTM, thus replacing the traditional step of feature screening with the proposed optimizing feature selection method.

The thesis' third stage of model development investigated seasonal pattern effects in weather data. To offset seasonality's impact on the model's accuracy, a deep stack of the clustering connected layer with hybrid LSTM models was proposed. The results showed that SCFT offers highly accurate performance, providing not only robust reliability but also generalizability of hourly GHI prediction.

5.2 Future Research

Some points were revealed during the course of this investigation, which could be used for future work directions. Five of these points in relation to future research are listed below.

First: In this thesis, meteorological and solar irradiance data was used from a satellite-based data source with 60-second temporal coverage and 4-km spatial resolution. These relatively limited temporal and spatial ranges could be addressed with higher temporal and spatial resolution satellite-based data or with other data sources, such as ground-based measurements, NWP-based data, and Sky-Imagers data. Further, these could be integrated or applied individually.

Second: The proposed model's eficiency is tied to the observed data's accessibility. So, for example, as some of the parameters are unavailable, cloud motion detection for a short time horizon (e.g., 10 minutes) could be used in GHI forecast applications. Potentially relevant data include cloud cover and wind vectors.

Third: This study fully investigated the effects of seasonality and geographic dependence, which significantly improved the accuracy of the solar irradiance prediction model. Therefore, future work could study on-site data in a local environment in relation to fluctuations caused by solar position. These data could be analysed from the perspective of stochastic or deterministic processes for cloudy or clear sky modeling, respectively.

Fourth: The analyses and validations of the proposed work concerned historical data in relation to meteorological and solar irradiance data tied to specific geographical locations. However, it did not conduct on-site measurements for PV systems and meteorological masts, mainly due to lack of collaboration and restrictions caused by research funding. Future work could consider real-time implementation of data from measurement stations in neighbouring locations. Such an approach could provide a more sophisticated strategy for developing a dynamic energy management platform to compare and incorporate real-world data with the designed forecasting data.

Fifth: The potential uses of recurrent neural networks such as that demonstrated in the LSTM model are considered promising research directions in the application of solar power forecasting. Research based around state-of-the-art deep learning-based forecasting techniques could be conducted to further develop the proposed LSTM model.

Bibliography

[1] C.A. Hernandez-Aramburo, T.C. Green, and N. Mugniot. Fuel consumption minimization of a microgrid. IEEE Transactions on Industry Applications, 41(3):673–681, 2005.

[2] Najiya Omar, Hamed Aly, and Timothy Little. Grid-connected photovoltaic system: System overview and sizing principles. International Journal of Electrical and Computer Engineering, 14(12):428 – 434, 2020.

[3] Leonid A. Kosyachenko. Solar Cells. IntechOpen, Rijeka, Nov 2011.

[4] Mohamed Chegaar, Amer Hamzaoui, Aboubacar Namoda, Pierre Petit, Michel Aillerie, and Axel Herguth. Effect of illumination intensity on solar cells parameters. Energy Procedia, 36:722–729, 2013.

[5] Priyanka Singh and Nuggehalli M Ravindra. Temperature dependence of solar cell performance—an analysis. Solar energy materials and solar cells, 101:36–45, 2012.

[6] David Elizondo, Gerrit Hoogenboom, and RW McClendon. Development of a neural network model to predict daily solar radiation. Agricultural and Forest Meteorology, 71(1-2):115–132, 1994.

[7] Jiaming Li, John K Ward, Jingnan Tong, Lyle Collins, and Glenn Platt. Machine learning for solar irradiance forecasting of photovoltaic system. Renewable energy, 90:542–553, 2016.

[8] Vladimir Kostylev, Alexandre Pavlovski, et al. Solar power forecasting performance–towards industry standards. In 1st international workshop on the integration of solar power into power systems, Aarhus, Denmark. Energynautics GmbH Mühlstraße Langen, Germany, 2011.

[9] Ying-Yi Hong, John Joel F. Martinez, and Arnel C. Fajardo. Day-ahead so-lar irradiation forecasting utilizing gramian angular field and convolutional long short-term memory. IEEE Access, 8:18741–18753, 2020.

[10] André Gensler, Janosch Henze, Bernhard Sick, and Nils Raabe. Deep learning for solar power forecasting — an approach using autoencoder and lstm neural networks. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 002858–002865, 2016.

[11] Bhaskar Pratim Mukhoty, Vikas Maurya, and Sandeep Kumar Shukla. Sequence to sequence deep learning models for solar irradiation forecasting. In 2019 IEEE Milan PowerTech, pages 1–6, 2019.

[12] Chun-Hung Liu, Jyh-Cherng Gu, and Ming-Ta Yang. A simplified lstm neural networks for one day-ahead solar power forecasting. IEEE Access, 9:17174–17195, 2021.

[13] Xiangyun Qing and Yugang Niu. Hourly day-ahead solar irradiance prediction using weather forecasts by lstm. Energy, 148:461–468, 2018.

[14] Jiani Heng, Jianzhou Wang, Liye Xiao, and Haiyan Lu. Research and application of a combined model based on frequent pattern growth algorithm and multi-objective optimization for solar radiation forecasting. Applied Energy, 208:845–866, 2017.

[15] Munir Husein and Il-Yop Chung. Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach. Energies, 12(10):1856, 2019.

[16] Qizal Ashfaq, Abasin Ulasyar, Haris Sheh Zad, Abraiz Khattak, and Kashif Imran. Hour-ahead global horizontal irradiance forecasting using long short term memory network. In 2020 IEEE 23rd International Multitopic Conference (INMIC), pages 1–6. IEEE, 2020.

[17] Xiaoqiao Huang, Junsheng Shi, Bixuan Gao, Yonghang Tai, Zaiqing Chen, and Jun Zhang. Forecasting hourly solar irradiance using hybrid wavelet transformation and elman model in smart grid. IEEE Access, 7:139909–139923, 2019.

[18] Najiya Omar, Hamed Aly, and Timothy 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 2021 4th International Conference on Energy, Electrical and Power Engineering, pages 1130–1135, 2021.

[19] Kasun Bandara, Christoph Bergmeir, and Hansika Hewamalage. Lstm-msnet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. IEEE transactions on neural networks and learning systems, 32(4):1586–1599, 2020.

[20] Dewang Chen, Jianhua Zhang, and Shixiong Jiang. Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and lstm neural networks. IEEE Access, 8:91181–91187, 2020.

[21] Najiya Omar, Hamed Aly, and Timothy Little. Optimized feature selection based on a least-redundant and highest-relevant framework for a solar irradiance forecasting model. IEEE Access, 10:48643–48659, 2022.

[22] Najiya Omar, Hamed Aly, and Timothy Little. Seasonal clustering forecasting technique for intelligent hourly solar irradiance systems. IEEE Transactions on Industrial Informatics, 2022.

[23] M Mohandes, S Rehman, and TO Halawani. Estimation of global solar radiation using artificial neural networks. Renewable energy, 14(1-4):179–184, 1998.

[24] SM Al-Alawi and HA Al-Hinai. An ann-based approach for predicting global radiation in locations with no direct measurement instrumentation. Renewable Energy, 14(1-4):199–204, 1998.

[25] A Sfetsos and AH Coonick. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy, 68(2):169–178, 2000.

[26] Atsu SS Dorvlo, Joseph A Jervase, and Ali Al-Lawati. Solar radiation estimation using artificial neural networks. Applied Energy, 71(4):307–319, 2002.

[27] G Mihalakakou, M Santamouris, and DN Asimakopoulos. The total solar radiation time series simulation in athens, using neural networks. Theoretical and Applied Climatology, 66(3):185–197, 2000.

[28] K Srinivas Reddy and Manish Ranjan. Solar resource estimation using artificial neural networks and comparison with other correlation models. Energy conversion and management, 44(15):2519–2530, 2003.

[29] Pravat Kumar Ray, Anindya Bharatee, Pratap Sekhar Puhan, and Sourav Sahoo. Solar irradiance forecasting using an artificial intelligence model. In 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), pages 1–5, 2022.

[30] Y Kemmoku, S Orita, S Nakagawa, and T Sakakibara. Daily insolation forecasting using a multi-stage neural network. Solar Energy, 66(3):193–199, 1999.

[31] Adel Mellit and Alessandro Massi Pavan. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected pv plant at trieste, italy. Solar energy, 84(5):807–821, 2010.

[32] Adnan Sözen, Erol Arcaklioğlu, and Mehmet Özalp. Estimation of solar potential in turkey by artificial neural networks using meteorological and geographical data. Energy Conversion and Management, 45(18-19):3033–3052, 2004.

[33] MKSRM Mohandes, A Balghonaim, M Kassas, S Rehman, and TO Halawani. Use of radial basis functions for estimating monthly mean daily solar radiation. Solar Energy, 68(2):161–168, 2000.

[34] R Meenal and A Immanuel Selvakumar. Assessment of svm, empirical and ann based solar radiation prediction models with most influencing input parameters. Renewable Energy, 121:324–343, 2018.

[35] Atika Qazi, H Fayaz, Ali Wadi, Ram Gopal Raj, NA Rahim, and Waleed Ahmed Khan. The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review. Journal of cleaner production, 104:1–12, 2015.

[36] Alfredo Nespoli, Emanuele Ogliari, Sonia Leva, Alessandro Massi Pavan, Adel Mellit, Vanni Lughi, and Alberto Dolara. Day-ahead photovoltaic forecasting: A comparison of the most effective techniques. Energies, 12(9):1621, 2019.

[37] Abdelaziz Rabehi, Mawloud Guermoui, and Djemoui Lalmi. Hybrid models for global solar radiation prediction: a case study. International Journal of Ambient Energy, 41(1):31–40, 2020.

[38] Unit Three Kartini, Hariyati, Widi Aribowo, and Ayusta Lukita Wardani. Development hybrid model deep learning neural network (dl-nn) for probabilistic forecasting solar irradiance on solar cells to improve economics value added. In 2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE), pages 151–156, 2022.

[39] Amit Kumar Yadav and SS Chandel. Solar radiation prediction using artificial neural network techniques: A review. Renewable and sustainable energy reviews, 33:772–781, 2014.

[40] Tamer Khatib, Azah Mohamed, Marwan Mahmoud, and K Sopian. Modeling of daily solar energy on a horizontal surface for five main sites in malaysia. International Journal of Green Energy, 8(8):795–819, 2011.

[41] Hatice Citakoglu. Comparison of artificial intelligence techniques via empiri-cal equations for prediction of solar radiation. Computers and Electronics in Agriculture, 118:28–37, 2015.

[42] Vahid Nourani, Gozen Elkiran, Jazuli Abdullahi, and Ala Tahsin. Multi-region modeling of daily global solar radiation with artificial intelligence ensemble. Natural Resources Research, 28(4):1217–1238, 2019.

[43] Nokuzola Mdluli, Gulshan Sharma, Kayode Akindeji, K. Narayanan, and Sachin Sharma. Development of short term solar radiation forecasting using ai techniques. In 2022 30th Southern African Universities Power Engineering Conference (SAUPEC), pages 1–6, 2022.

[44] JCSH Cao Cao and SH Cao. Study of forecasting solar irradiance using neu-ral networks with preprocessing sample data by wavelet analysis. Energy, 31(15):3435–3445, 2006.

[45] Nian Zhang and Pradeep K Behera. Solar radiation prediction based on recurrent neural networks trained by levenberg-marquardt backpropagation learning algorithm. In 2012 IEEE PES Innovative Smart Grid Technologies (ISGT), pages 1–7. IEEE, 2012.

[46] Jiacong Cao and Xingchun Lin. Application of the diagonal recurrent wavelet neural network to solar irradiation forecast assisted with fuzzy technique. Engineering Applications of Artificial Intelligence, 21(8):1255–1263, 2008.

[47] Shuanghua Cao. Total daily solar irradiance prediction using recurrent neural networks with determinants. In 2010 Asia-Pacific Power and Energy Engineering Conference, pages 1–4. IEEE, 2010.

[48] Zhihong Pang, Fuxin Niu, and Zheng O'Neill. Solar radiation prediction us-ing recurrent neural network and artificial neural network: A case study with comparisons. Renewable Energy, 156:279–289, 2020.

[49] Seyed Mohammad Jafar Jalali, Sajad Ahmadian, Abdollah Kavousi-Fard, Abbas Khosravi, and Saeid Nahavandi. Automated deep cnn-lstm architecture design for solar irradiance forecasting. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(1):54–65, 2022.

[50] Bhaskar Pratim Mukhoty, Vikas Maurya, and Sandeep Kumar Shukla. Sequence to sequence deep learning models for solar irradiation forecasting. In 2019 IEEE Milan PowerTech, pages 1–6. IEEE, 2019.

[51] Haixiang Zang, Ling Liu, Li Sun, Lilin Cheng, Zhinong Wei, and Guoqiang Sun. Short-term global horizontal irradiance forecasting based on a hybrid cnn-lstm model with spatiotemporal correlations. Renewable Energy, 160:26–41, 2020.

[52] Mohamed Abdel-Nasser, Karar Mahmoud, and Matti Lehtonen. Reliable solar irradiance forecasting approach based on choquet integral and deep lstms. IEEE Transactions on Industrial Informatics, 17(3):1873–1881, 2020.

[53] Tendani Mutavhatsindi, Caston Sigauke, and Rendani Mbuvha. Forecasting hourly global horizontal solar irradiance in south africa using machine learning models. IEEE Access, 8:198872–198885, 2020.

[54] Yunjun Yu, Junfei Cao, and Jianyong Zhu. An lstm short-term solar irradiance forecasting under complicated weather conditions. IEEE Access, 7:145651–145666, 2019.

[55] Cheng Pan and Jie Tan. Day-ahead hourly forecasting of solar generation based on cluster analysis and ensemble model. IEEE Access, 7:112921–112930, 2019.

[56] JL Bosch, G Lopez, and FJ Batlles. Daily solar irradiation estimation over a mountainous area using artificial neural networks. Renewable Energy, 33(7):1622–1628, 2008.

[57] 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. Renewable energy, 50:168–176, 2013.

[58] Saurabh Bhardwaj, Vikrant Sharma, Smriti Srivastava, OS Sastry, B Bandyopadhyay, SS Chandel, and JRP Gupta. Estimation of solar radiation using a combination of hidden markov model and generalized fuzzy model. Solar Energy, 93:43–54, 2013.

[59] Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata, and André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. Briefings in bioinformatics, 15(6):906–918, 2014.

[60] Wassily Hoeffding. A Non-Parametric Test of Independence. The Annals of Mathematical Statistics, 19(4):546 – 557, 1948.

[61] United States. National Renewable Energy Laboratory. Nrel national solar radiation database, 2020.

[62] Hae-Young Kim. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative dentistry & endodontics, 38(1):52–54, 2013.

[63] Sas ondemand for academics. https://support.sas.com/ondemand/manuals/SASStudio.pdf,Access data on March, 2021.

[64] Jiani Heng, Jianzhou Wang, Liye Xiao, and Haiyan Lu. Research and application of a combined model based on frequent pattern growth algorithm and multi-objective optimization for solar radiation forecasting. Applied Energy, 208:845–866, 2017.

[65] Anamika Yadav and Niranjan Kumar. Solar resource estimation based on correlation matrix response for indian geographical cities. International Journal of Renewable Energy Research (IJRER), 6(2):695–701, 2016.

[66] Michael J Patetta. Predictive modeling using logistic regression: Course notes. https:Predictive Modeling Using Logistic Regression : course notes (Book, 2001) [WorldCat.org].

[67] Andrew 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. Educational and psychological measurement, 53(1):87–92, 1993.

[68] Christophe Croux and Catherine Dehon. Influence functions of the spearman and kendall correlation measures. Statistical methods & applications, 19(4):497–515, 2010.

[69] Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. In Advances in neural information processing systems, pages 800–808, 2012.

[70] David Christensen. Fast algorithms for the calculation of kendall's τ. Computational Statistics, 20(1):51–62, 2005.

[71] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA, 2006.

[72] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. IEEE Transactions on neural networks, 20(2):189–201, 2009.

[73] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is dificult. IEEE transactions on neural networks, 5(2):157–166, 1994.

[74] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[75] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In Third IEEE international conference on data mining, pages 99–106. IEEE, 2003.

[76] Biao Jia, Zherong Pan, Zhe Hu, Jia Pan, and Dinesh Manocha. Cloth manipulation using random-forest-based imitation learning. IEEE Robotics and Automation Letters, 4(2):2086–2093, 2019.

[77] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In CVPR 2011, pages 1297–1304. IEEE, 2011.

[78] Innocent Kamwa, SR Samantaray, and Geza Joós. On the accuracy versus transparency trade-off of data-mining models for fast-response pmu-based catastrophe predictors. IEEE Transactions on Smart Grid, 3(1):152–161, 2011.

[79] M Maroof Khan, M Jamil Ahmad, and Basharat Jamil. Development of models for the estimation of global solar radiation over selected stations in india. In Energy, Transportation and Global Warming, pages 149–160. Springer, 2016.

[80] Juntao Wang and Xiaolong Su. An improved k-means clustering algorithm. In 2011 IEEE 3rd international conference on communication software and networks, pages 44–46. IEEE, 2011.

[81] Kipp zonen: Solar radiation measurement.the benefits of accurately measuring solar irradiance. https://pages.kippzonen.com/1912$_K$Z$_C$NT$_c$ − met$_W$hitepaper − benefits − Measuring − Solar − Irradiance − USA − Feb − March2019$_1$9 − Landing − Page.html.

[82] Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. World map of the köppen-geiger climate classification updated. 2006.

[83] Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. Journal of Business and Economic Statistics, 13(3):253–263, 1995.

[84] Ali Azadeh, Morteza Saberi, Anahita Gitiforouz, and Zahra Saberi. A hybrid simulation-adaptive network based fuzzy inference system for improvement of electricity consumption estimation. Expert Systems with Applications, 36(8):11108–11117, 2009.

[85] M Ghiassi, H Saidane, and D K Zimbra. A dynamic artificial neural network model for forecasting time series events. International Journal of Forecasting, 21(2):341–362, 2005.

Appendix A

Completed Results of the Relevance Measure Validation

Table A.1: Nonparametric Relevance Measure for a 95% Confidence Interval (Year 2005)

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Correlation	Spearman p	Hoeffding Correlation	Hoeffding p	Kendall Correlation	Kendall p
DNI	1	3	2	0.78264	<.0001	0.2484	<.0001	0.59665	<.0001
CGHI	2	1	1	0.78156	<.0001	0.28438	<.0001	0.6105	<.0001
SZA	3	2	3	-0.75127	<.0001	0.24868	<.0001	-0.57432	<.0001
DHI	4	4	4	0.63027	<.0001	0.20379	<.0001	0.49028	<.0001
RH	5	5	5	-0.57344	<.0001	0.11813	<.0001	-0.40611	<.0001
T	6	6	6	0.24819	<.0001	0.01899	<.0001	0.16696	<.0001
WS	7	7	7	-0.22951	<.0001	0.01528	<.0001	-0.15538	<.0001
WD	8	8	9	0.15437	<.0001	0.00807	<.0001	0.0993	<.0001
P	9	9	8	0.14673	<.0001	0.00516	<.0001	0.11102	<.0001
PW	10	10	10	-0.11698	<.0001	0.00467	<.0001	-0.07674	<.0001
DP	11	11	11	0.06308	<.0001	0.00255	<.0001	0.04096	<.0001
SA	12	12	12	0.0235	0.1228	0.00188	<.0001	0.01524	0.1545

Table A.2: Nonparametric Relevance Measure for a 95% Confidence Interval (Year 2010)

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Correlation	Spearman p	Hoeffding Correlation	Hoeffding p	Kendall Correlation	Kendall p
DNI	1	2	1	0.80019	<.0001	0.26532	<.0001	0.61622	<.0001
CGHI	2	1	2	0.78788	<.0001	0.28805	<.0001	0.61494	<.0001
SZA	3	3	3	-0.77051	<.0001	0.26162	<.0001	-0.59026	<.0001
DHI	4	4	4	0.65147	<.0001	0.22198	<.0001	0.5118	<.0001
RH	5	5	5	-0.61544	<.0001	0.13295	<.0001	-0.43432	<.0001
T	6	6	6	0.34091	<.0001	0.03459	<.0001	0.23246	<.0001
WS	7	7	8	-0.22162	<.0001	0.01465	<.0001	-0.15066	<.0001
P	8	8	7	0.2073	<.0001	0.01095	<.0001	0.156	<.0001
DP	9	10	9	0.13563	<.0001	0.00662	<.0001	0.09025	<.0001
WD	10	11	11	0.1249	<.0001	0.00499	<.0001	0.08246	<.0001
SA	11	9	10	0.12163	<.0001	0.00837	<.0001	0.08324	<.0001
PW	12	12	12	-0.02887	0.0581	0.00158	<.0001	-0.01928	0.0579

Table A.3: Nonparametric Relevance Measure for a 95% Confidence Interval (Year 2015)

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Correlation	Spearman p	Hoeffding Correlation	Hoeffding p	Kendall Correlation	Kendall p
CGHI	1	1	1	0.81734	<.0001	0.32748	<.0001	0.64818	<.0001
SZA	2	2	2	-0.7958	<.0001	0.29287	<.0001	-0.61711	<.0001
DNI	3	3	3	0.78947	<.0001	0.25033	<.0001	0.59911	<.0001
DHI	4	4	4	0.63279	<.0001	0.19413	<.0001	0.48626	<.0001
RH	5	5	5	-0.53387	<.0001	0.09485	<.0001	-0.37293	<.0001
T	6	6	6	0.24284	<.0001	0.01852	<.0001	0.16295	<.0001
WS	7	7	7	-0.23827	<.0001	0.01578	<.0001	-0.16171	<.0001
WD	8	8	8	0.19159	<.0001	0.01276	<.0001	0.12608	<.0001
SA	9	9	9	0.11532	<.0001	0.00693	<.0001	0.08029	<.0001
PW	10	10	10	-0.10657	<.0001	0.00394	<.0001	-0.07026	<.0001
P	11	12	11	0.09359	<.0001	0.00221	<.0001	0.07021	<.0001
DP	12	11	12	0.08071	<.0001	0.00283	<.0001	0.05274	<.0001

Table A.4: Nonparametric Relevance Measure for a 95% Confidence Interval (Year 2018)

Variable	Spearman Rank	Hoeffding Rank	Kendall Rank	Spearman Correlation	Spearman p	Hoeffding Correlation	Hoeffding p	Kendall Correlation	Kendall p
DNI	1	3	3	0.81317	<.0001	0.27617	<.0001	0.61672	<.0001
CGHI	2	1	1	0.80578	<.0001	0.3079	<.0001	0.63816	<.0001
SZA	3	2	2	-0.78969	<.0001	0.28928	<.0001	-0.61871	<.0001
DHI	4	4	4	0.71814	<.0001	0.27418	<.0001	0.56463	<.0001
RH	5	5	5	-0.6185	<.0001	0.13871	<.0001	-0.4357	<.0001
T	6	6	6	0.31537	<.0001	0.02915	<.0001	0.21003	<.0001
WS	7	7	7	-0.23815	<.0001	0.01779	<.0001	-0.16183	<.0001
PW	8	8	8	0.18549	<.0001	0.01144	<.0001	0.12543	<.0001
DP	9	9	9	0.11214	<.0001	0.00439	<.0001	0.07256	<.0001
PW	10	10	11	0.08631	<.0001	0.00433	<.0001	0.05507	<.0001
SA	11	11	10	0.078	<.0001	0.00264	<.0001	0.05782	<.0001
PW	12	12	12	-0.07353	<.0001	0.00261	<.0001	-0.04864	<.0001

Appendix B

Clustering Deep Stack Output (Clustered Datasets)

Table B.1: The 13 Standardized Features and Assigned Clusters (Data Points 1-30)

	Feature0	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12	Cluster
0	0.660644	-0.953530	-0.418768	1.485284	1.474560	-1.881645	-0.417786	0.180305	1.586512	-1.223525	-0.061421	1.645248	-0.494096	4
1	0.926257	-0.687330	-0.154342	0.166677	0.152925	-0.031432	-0.511022	0.103713	-0.166889	-1.607259	0.506483	0.011200	-1.514745	4
2	-1.116920	-0.977192	-1.122606	-0.847071	-0.087372	0.791071	-0.504116	1.788739	-0.435541	-0.092920	0.717681	-0.240193	-3.556042	2
3	-0.728716	1.084381	-0.107679	-0.619345	0.032777	0.611096	-0.473037	-0.738799	-0.275529	1.200283	-0.165960	0.136896	-0.494096	2
4	2.969433	-0.563103	0.903359	1.588128	0.873817	-1.670833	-0.479944	-1.045168	0.165769	-0.530854	-1.554641	1.519552	0.526553	4
5	0.977336	-0.938741	-0.290444	0.824144	0.273074	-0.681810	-0.490303	-0.126063	0.048707	-1.476096	-0.808737	0.639680	1.547202	4
6	1.202086	-0.941699	-0.201006	1.301634	1.114114	-1.637753	-0.455771	-0.815391	1.609250	0.270217	0.502245	1.016768	-0.494096	4
7	-1.116920	-0.430003	-1.079831	-1.537595	1.114114	1.461072	-0.448865	-0.355839	0.598645	0.287561	1.214951	0.765376	-0.494096	3
8	-1.096488	-0.912121	-1.110940	-1.552287	1.474560	1.453784	-0.435052	-1.045168	1.574721	1.213291	1.214951	1.142464	-0.494096	3
9	-0.463103	1.646360	1.222225	0.636821	0.273074	-0.442965	-0.486850	-0.738799	-0.391749	0.480512	-2.012355	1.142464	0.526553	1
10	-1.249726	-0.977192	-1.173158	-1.526576	-1.288858	1.453784	2.085787	0.486674	0.015020	-1.283145	-0.243659	-1.371457	0.526553	0
11	-1.014761	-0.784937	-1.060388	-1.456789	1.234262	1.357348	-0.462678	-0.968576	0.477372	-0.447386	1.214951	0.891072	-0.494096	3
12	-0.728716	0.498741	-0.558758	-1.045413	0.513371	0.960954	-0.476490	0.410082	-0.176995	-2.263075	-0.213992	0.639680	0.526553	3
13	-1.004545	-0.977192	-1.079831	-1.335580	0.032777	1.195875	-0.483397	1.252594	1.029836	-0.124356	1.214951	-0.365889	0.526553	3
14	-1.280374	-0.977192	-1.184824	-1.585344	0.993965	1.422386	-0.483397	-0.202655	1.271540	0.108702	1.214951	0.639680	-0.494096	3
15	0.732155	-0.956488	-0.399325	1.073908	0.032777	-1.135393	-0.497210	0.639858	1.205008	-1.135722	0.285396	0.011200	0.526553	4
16	1.978493	0.297611	1.420544	1.305307	1.234262	-1.398347	-0.466131	0.333490	0.697179	0.135802	-0.165960	1.393856	0.526553	4
17	0.527837	-0.888459	-0.442099	-0.002281	0.032777	0.043137	-0.441958	-0.738799	-0.135729	1.311935	0.873078	-0.240193	-0.494096	3
18	-0.871739	-0.977192	-1.029279	-1.353945	-0.327669	1.176813	-0.500663	2.554659	0.609593	-1.411056	0.830697	-0.617281	0.526553	3
19	-0.228138	1.699600	1.074458	0.497247	-1.409007	-0.130671	2.085787	1.329186	-1.169073	0.431732	-1.397126	-1.120065	-1.514745	0
20	-1.229295	-0.977192	-1.165381	-1.592690	-1.649304	1.589466	2.085787	-0.202655	-1.145492	1.275079	-1.517911	-1.371457	-0.494096	0
21	-0.595910	0.912830	-0.029907	-0.523847	-0.207521	0.471489	-0.521382	-0.585615	-0.486072	-1.480432	-0.787547	0.011200	0.526553	2
22	-0.166843	-0.977192	-0.760965	0.287886	1.234262	-0.294948	-0.459224	-0.738799	0.603698	0.797038	0.839173	1.016768	-0.494096	3
23	1.426835	0.359725	1.078346	1.081254	-0.327669	-0.833752	2.085787	-0.509023	-0.004350	-0.733560	1.125951	-0.617281	0.526553	4
24	-0.125979	-0.764232	-0.659861	-0.843398	1.354411	0.655950	-0.448865	-1.198352	1.470292	-0.117852	1.214951	1.016768	-0.494096	3
25	-0.228138	1.749882	2.093273	1.459573	0.273074	-1.336673	-0.469584	-0.355839	-0.718511	0.100030	-2.184704	1.268160	1.547202	1
26	-0.350729	-0.977192	-0.830960	0.379711	-0.327669	-0.474923	-0.486850	-0.049471	0.704758	-1.422980	1.051078	-0.617281	0.526553	3
27	0.129418	-0.820430	-0.574312	-0.347543	-0.928412	0.482702	-0.514476	0.180305	-1.082329	1.037684	-0.784722	-0.868673	1.547202	2
28	-0.197490	1.522133	1.560534	0.956372	0.513371	-0.879727	-0.466131	-0.891984	-0.298268	-0.253351	-1.166856	1.016768	0.526553	1
29	-0.289433	1.406779	1.218336	0.633148	0.993965	-0.602756	-0.479944	-0.968576	0.612962	0.148810	-0.707730	1.268160	-0.494096	1

Table B.2: The 13 Standardized Features and Assigned Clusters (Data Points 121-155)

	Feature0	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12	Cluster
121	-1.219079	-0.977192	-1.161492	-1.276812	1.114114	1.137566	-0.417786	1.176002	1.255538	-1.497776	1.144316	0.891072	-0.494096	3
122	1.641369	0.200005	1.027795	0.996775	0.393222	-0.884212	-0.462678	-1.504720	0.246617	-0.642505	-1.687435	1.142464	0.526553	1
123	1.304244	-0.929868	-0.150454	1.279596	-0.808264	-1.056338	-0.493756	-0.049471	-0.848205	0.870750	-0.569992	-0.742977	-0.494096	4
124	-1.270158	-0.977192	-1.180935	-1.537595	-0.688115	1.493031	2.085787	0.027121	-0.651137	0.609507	1.190935	-1.120065	-2.535394	0
125	1.723096	-0.841134	0.082863	1.147368	-0.808264	-0.815250	2.085787	1.252594	-0.720195	-2.182859	-0.173730	-0.868673	0.526553	0
126	-0.687853	-0.977192	-0.959284	-0.769938	-0.447818	0.669406	-0.528288	2.478067	1.350704	-1.396964	1.169038	-0.742977	-0.494096	3
127	0.190713	1.309173	1.478873	1.239193	-1.529155	-0.829267	2.085787	3.090804	-1.029272	1.522229	-1.500959	-1.245761	-0.494096	0
128	-0.892170	-0.977192	-1.037056	-1.269466	-0.567966	1.203164	2.085787	2.171699	0.006598	-1.656039	1.214951	-0.994369	-0.494096	0
129	0.395031	1.217481	1.031683	0.456844	-0.447818	-0.202437	2.085787	-1.734496	-0.477650	0.044747	0.969141	-0.742977	1.547202	0
130	-0.912602	0.679165	-0.698748	-1.177641	-0.808264	1.147097	-0.490303	0.639858	-0.841468	0.821970	-0.959190	-0.617281	0.526553	2
131	-0.514183	1.572415	0.658377	0.104236	-0.688115	0.112099	-0.531742	2.171699	-1.081487	-1.860913	-0.946475	-0.365889	1.547202	2
132	0.231577	1.504386	2.342144	1.694645	0.873817	-1.974156	-0.448865	-0.891984	0.324097	-0.211076	-0.122167	1.016768	0.526553	1
133	0.701507	-0.882543	-0.364327	0.093217	0.633520	-0.033114	-0.507569	-0.049471	0.301358	0.385121	1.214951	0.388288	-1.514745	3
134	-0.432456	1.338750	0.549496	0.001392	-0.688115	0.110417	-0.486850	-0.355839	-0.961056	0.872917	-2.066037	-0.114497	0.526553	1
135	-0.074900	1.637486	2.279926	1.635877	1.354411	-1.764465	-0.459224	-0.585615	0.157347	-0.238176	-0.311468	1.519552	0.526553	1
136	1.447267	-0.474370	0.242296	0.563361	-0.087372	-0.365592	-0.504116	1.252594	-0.310900	1.175352	-0.440024	0.011200	-0.494096	4
137	0.864961	-0.950572	-0.337107	1.430189	-0.327669	-1.474598	-0.486850	-1.121760	0.034390	-0.543862	0.585594	-0.491585	-0.494096	4
138	0.200929	1.465935	2.038833	1.408151	0.273074	-1.471795	-0.462678	-0.968576	0.090816	0.316829	-0.514897	0.513984	0.526553	1
139	-0.616342	1.040015	-0.022130	-0.538539	0.393222	0.533723	-0.493756	-0.355839	-0.338692	0.895681	-0.375040	0.513984	0.526553	2
140	0.967120	0.217751	0.576716	0.537650	-0.808264	-0.316253	-0.486850	1.329186	-1.024219	-1.476096	-0.301579	-0.868673	1.547202	2
141	-0.718500	0.992690	-0.166008	-0.674440	0.513371	0.639690	-0.448865	-0.891984	-0.233420	0.620347	0.632213	0.388288	-0.494096	3
142	-0.013604	-0.977192	-0.702636	0.096890	-0.327669	-0.056102	-0.528288	0.256897	-0.307531	1.250147	-0.209754	-0.240193	-0.494096	2
143	-0.360945	0.714659	0.001202	-0.516501	0.993965	0.422711	-0.479944	-0.509023	0.821820	-0.114600	0.173793	1.016768	-0.494096	3
144	-0.636773	0.084651	-0.714302	-1.192333	-0.688115	1.131959	2.085787	2.937620	-0.849890	0.807878	1.014348	-1.120065	-1.514745	0
145	1.804823	0.069862	0.969465	0.989429	0.753668	-0.894304	-0.483397	-0.968576	0.180086	-0.361751	-0.972610	1.142464	0.526553	4
146	-0.473319	1.430442	0.743926	0.185042	0.032777	-0.073483	-0.476490	-0.355839	-0.509653	1.504886	-2.027895	0.891072	1.547202	1
147	0.926257	-0.406341	0.040088	0.107909	0.273074	-0.083014	-0.528288	-1.121760	-0.297425	-0.310803	0.457745	0.136896	-0.494096	4
148	-0.136195	1.066635	0.565050	0.016084	-0.447818	0.016785	-0.459224	-1.045168	-0.562709	-2.213211	0.540388	-0.617281	0.526553	1
149	-1.208863	-0.977192	-1.157603	-1.603709	-0.207521	1.559750	-0.517929	1.558963	-0.247737	0.168322	0.259261	-0.240193	-0.494096	2
150	-0.718500	0.874379	-0.302109	-0.802995	0.633520	0.758552	-0.441958	-0.738799	-0.495336	1.464778	0.317888	0.639680	0.526553	3
151	0.200929	0.504656	0.304513	0.027103	0.993965	0.067806	-0.479944	0.410082	0.059655	-0.337903	0.792554	0.765376	0.526553	3
152	-0.800227	0.191131	-0.745411	-1.221717	0.032777	1.120185	-0.524835	-0.355839	-0.483545	-1.214853	1.214951	-0.365889	1.547202	3
153	2.734468	-0.852966	0.495055	1.749740	0.152925	-2.149646	-0.421239	-0.509023	0.687073	-1.778530	0.324245	0.011200	0.526553	4
154	0.037475	1.495513	1.999947	1.371421	0.633520	-1.458339	-0.428145	-1.121760	0.238195	-0.346575	0.072079	0.639680	0.526553	1

Table B.3: The 13 Standardized Features and Assigned Clusters (Data Points 3050-3079)

	Feature0	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12	Cluster
3050	-0.756348	0.823505	-0.385706	-0.878447	0.643189	0.832025	-0.493225	-0.357080	-0.448645	1.353806	-0.826721	1.027837	0.518017	4
3051	-1.085088	-0.218503	-1.045881	-1.499167	0.763392	1.430293	-0.493225	-0.052993	-0.467164	1.402472	-0.038255	0.776752	0.518017	0
3052	-0.889899	0.378202	-0.755090	-1.225755	0.643189	1.136782	-0.493225	-0.889234	-0.364469	1.135347	0.384721	0.525667	0.518017	0
3053	-0.643343	1.055062	-0.032041	-0.545919	0.643189	0.538514	-0.493225	-1.041277	-0.379621	1.188339	-0.327809	0.776752	1.547859	4
3054	-0.499519	1.393492	0.694938	0.137611	0.522985	-0.046258	-0.493225	-1.117299	-0.394773	1.154814	-0.802591	0.902294	1.547859	2
3055	-0.407061	1.583488	1.327606	0.732467	0.522985	-0.587735	-0.493225	-1.269343	-0.400665	1.110473	-1.258213	1.027837	1.547859	2
3056	-0.345422	1.690360	1.799159	1.175838	0.522985	-1.041496	-0.493225	-1.345365	-0.391406	0.995836	-1.583251	1.153379	1.547859	2
3057	-0.283783	1.725984	2.070303	1.430776	0.522985	-1.338943	-0.493225	-1.345365	-0.369520	0.791436	-1.747189	1.278922	0.518017	2
3058	-0.263237	1.725984	2.117458	1.475113	0.643189	-1.401919	-0.493225	-1.345365	-0.342583	0.507007	-1.635768	1.278922	0.518017	2
3059	-0.242690	1.669579	1.940626	1.308849	0.643189	-1.210181	-0.493225	-1.269343	-0.315647	0.271245	-1.533573	1.278922	0.518017	2
3060	-0.232417	1.553801	1.547664	0.939373	0.643189	-0.822207	-0.493225	-1.041277	-0.294603	0.129571	-1.464023	1.278922	0.518017	2
3061	-0.324875	1.399429	0.985729	0.411023	0.643189	-0.315591	-0.493225	-0.813212	-0.279451	0.066845	-1.377441	1.278922	0.518017	2
3062	-0.499519	1.173809	0.309836	-0.224475	0.763392	0.253438	-0.493225	-0.585146	-0.277768	0.049541	-0.889174	1.153379	0.518017	2
3063	-0.694709	0.648352	-0.432862	-0.922784	0.883596	0.848894	-0.493225	-0.509124	-0.292919	0.033319	-0.137612	1.027837	0.518017	0
3064	-1.095362	-0.387718	-1.069459	-1.521335	1.003799	1.447161	-0.493225	-0.357080	-0.310596	0.067926	0.460657	0.902294	0.518017	0
3065	-0.766621	-0.708336	-0.935852	-1.251618	0.763392	1.148028	-0.493225	-0.129015	-0.233154	0.482133	0.935440	0.525667	0.518017	0
3066	-0.458426	0.571166	-0.181366	-0.601341	0.763392	0.549760	-0.493225	-0.281058	-0.145611	0.408592	1.098669	0.525667	0.518017	0
3067	-0.252963	0.995688	0.514176	0.063716	0.883596	-0.034450	-0.493225	-0.281058	-0.060593	0.330726	0.471303	0.776752	0.518017	2
3068	0.199055	1.028344	1.076110	0.647487	0.883596	-0.574241	-0.493225	-0.281058	0.036209	0.298282	0.312332	0.902294	0.518017	2
3069	-0.057774	1.485521	1.696989	1.079774	0.883596	-1.025753	-0.493225	-0.205036	0.141429	0.291793	-0.073030	1.027837	0.518017	2
3070	0.034685	1.515208	1.952414	1.319933	0.883596	-1.319826	-0.493225	-0.129015	0.230656	0.265837	-0.495295	1.153379	0.518017	2
3071	0.096324	1.491458	1.991710	1.356881	0.883596	-1.381677	-0.493225	-0.052993	0.294630	0.236637	-0.810398	1.278922	0.518017	2
3072	1.421560	0.565229	1.433705	1.186922	0.883596	-1.189939	-0.493225	0.023029	0.345136	0.206356	-0.761429	1.278922	0.518017	2
3073	0.969541	0.588979	1.115407	0.821141	1.003799	-0.803089	-0.493225	0.099051	0.384698	0.190134	-0.696848	1.278922	0.518017	2
3074	0.394245	0.627571	0.667430	0.300180	1.003799	-0.297598	-0.493225	0.175073	0.405742	0.178237	-0.609556	1.278922	0.518017	2
3075	0.342879	-0.251159	-0.161718	-0.331623	1.003799	0.270306	-0.493225	0.175073	0.403217	0.176074	-0.130515	1.153379	0.518017	0
3076	-0.489246	-0.203660	-0.637201	-0.996680	1.124003	0.865762	-0.493225	0.175073	0.396483	0.187971	0.416657	1.027837	0.518017	0
3077	-1.095362	-0.714273	-1.100896	-1.550893	1.124003	1.464592	-0.493225	0.251095	0.412476	0.216089	0.936150	0.902294	0.518017	0
3078	-0.787167	-0.562871	-0.916204	-1.318123	1.244206	1.158711	-0.493225	-0.129015	1.448685	0.637866	1.217897	0.902294	0.518017	0
3079	-0.109139	-0.936925	-0.715794	-0.693710	1.244206	0.560443	-0.493225	-0.357080	1.349357	0.622725	0.949634	1.027837	0.518017	0

Table B.4: The 13 Standardized Features and Assigned Clusters (Data Points 47370-47406)

	Feature0	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12	Cluster
47370	-0.242690	-0.990361	-0.794386	-0.486803	-0.198235	0.529518	-0.517206	0.023029	-0.547131	-0.190547	1.217897	-0.478672	-4.631195	1
47371	-0.653616	0.897722	-0.287466	-0.786078	-0.198235	0.784231	-0.517206	0.403139	-0.537030	-0.263007	1.217897	-0.478672	-4.631195	1
47372	-0.859079	0.387108	-0.751160	-1.222060	-0.198235	1.154213	-0.517206	0.479161	-0.547973	-0.203525	1.217897	-0.604215	-4.631195	1
47373	-1.146727	-0.990361	-1.140192	-1.443745	-1.159863	1.378001	2.059056	2.227665	-0.869526	1.003406	0.406721	-1.483012	-3.601353	2
47374	-0.499519	-0.889426	-0.865119	-0.967122	-1.159863	0.959101	2.059056	2.227665	-0.888886	0.996918	-0.099998	-1.357469	-3.601353	2
47375	-0.540612	-0.990361	-0.908345	-0.553309	-1.159863	0.641412	2.059056	2.227665	-0.903196	0.991510	-0.191548	-1.357469	-3.601353	2
47376	0.260694	-0.485685	-0.362128	-0.302065	-1.280067	0.452485	2.059056	2.227665	-0.909930	0.979614	-0.280969	-1.357469	-3.601353	2
47377	0.322333	-0.687555	-0.428932	-0.246644	-1.280067	0.412563	2.059056	2.227665	-0.907405	0.963392	-0.313615	-1.357469	-3.601353	2
47378	0.147689	-0.714273	-0.519313	-0.398129	-1.280067	0.525582	2.059056	2.227665	-0.897304	0.947169	-0.266775	-1.357469	-3.601353	2
47379	-0.222144	-0.542090	-0.629342	-0.738047	-1.159863	0.779171	2.059056	2.227665	-0.883836	0.926621	-0.164580	-1.357469	-3.601353	2
47380	-0.982357	-0.990361	-1.077318	-1.199891	-1.159863	1.148590	2.059056	2.227665	-0.872893	0.914725	-0.057417	-1.357469	-3.601353	2
47381	-1.002903	0.027897	-0.971219	-1.428966	-1.280067	1.379688	2.059056	1.847555	-0.976430	1.075866	0.217943	-1.483012	-1.541668	2
47382	-0.550885	-0.984424	-0.908345	-0.937563	-1.280067	0.959663	2.059056	1.771533	-0.990740	1.081273	-0.221355	-1.357469	-1.541668	2
47383	-0.098866	-0.230378	-0.428932	-0.516361	-1.159863	0.640850	2.059056	1.771533	-0.994948	1.080191	-0.173806	-1.357469	-1.541668	2
47384	0.055231	0.060553	-0.173507	-0.250339	-1.159863	0.450236	2.059056	1.771533	-0.995790	1.080191	-0.483231	-1.231927	-1.541668	2
47385	0.209328	-0.144286	-0.201014	-0.194917	-1.159863	0.408627	2.059056	1.695511	-0.998315	1.088843	-0.336325	-1.231927	-1.541668	2
47386	0.086050	-0.188816	-0.303184	-0.342708	-1.039660	0.521084	2.059056	1.619489	-1.002524	1.107228	-0.191548	-1.231927	-1.541668	2
47387	-0.448153	0.220862	-0.440721	-0.678930	-1.039660	0.773548	2.059056	1.543468	-1.008417	1.132103	-0.072320	-1.231927	-1.541668	2
47388	-0.694709	-0.628182	-0.892626	-1.155554	-1.039660	1.142405	2.059056	1.391424	-1.011784	1.152651	0.018520	-1.231927	-1.541668	2
47389	-1.043996	0.401952	-0.947641	-1.406798	-1.520474	1.380250	2.059056	0.935292	-1.290407	1.084517	-0.242646	-1.608554	-0.511826	2
47390	-0.797440	1.164903	-0.401425	-0.893226	-1.400270	0.959663	2.059056	1.011314	-1.288723	1.063969	-0.760010	-1.357469	-0.511826	2
47391	-0.653616	1.455834	0.050481	-0.468329	-1.400270	0.639725	2.059056	1.087336	-1.250844	1.062888	-1.004143	-1.231927	-0.511826	2
47392	-0.561158	1.547863	0.309836	-0.224475	-1.280067	0.447987	2.059056	1.163358	-1.170035	1.048829	-1.219889	-1.106385	-0.511826	2
47393	-0.509792	1.512239	0.349132	-0.187528	-1.280067	0.405254	2.059056	1.239380	-1.073232	1.032606	-0.984272	-1.106385	-0.511826	2
47394	-0.520065	1.375680	0.180158	-0.346402	-1.159863	0.516023	2.059056	1.239380	-0.999999	1.001243	-0.625879	-1.106385	-0.511826	2
47395	-0.427607	0.170394	-0.448580	-0.690015	-1.039660	0.767925	2.059056	1.163358	-0.946126	0.956903	-0.253291	-1.106385	-0.511826	2
47396	-0.828260	0.591947	-0.684357	-1.159249	-0.919456	1.136220	2.059056	1.011314	-0.898988	0.880118	0.053295	-1.106385	-0.511826	2
47397	-1.208366	-0.643025	-1.155911	-1.602620	-0.919456	1.593354	2.059056	1.087336	-0.859425	0.792518	0.248460	-1.106385	-0.511826	2
47398	-1.013176	0.069459	-0.971219	-1.428966	-0.799253	1.380812	2.059056	0.327117	-0.856058	0.904992	0.266202	-1.106385	0.518017	2
47399	-0.725528	0.888816	-0.448580	-0.937563	-0.799253	0.959101	2.059056	0.175073	-0.874577	0.911481	0.481948	-1.106385	0.518017	2
47400	-0.571431	1.242089	-0.004534	-0.520056	-0.799253	0.638038	2.059056	0.175073	-0.879627	0.902829	0.217943	-0.980842	-0.511826	2
47401	-0.489246	1.390523	0.266610	-0.265118	-0.679049	0.445176	2.059056	0.175073	-0.867001	0.867140	0.404592	-0.855300	-0.511826	2
47402	-0.468700	1.414273	0.321624	-0.213391	-0.679049	0.400755	2.059056	0.251095	-0.836697	0.846592	0.135619	-0.855300	-0.511826	2
47403	-0.520065	1.334118	0.164440	-0.361181	-0.679049	0.510400	2.059056	0.251095	-0.772723	0.841184	0.266912	-0.855300	-0.511826	2
47404	-0.633070	1.096623	-0.193155	-0.697404	-0.679049	0.761740	2.059056	0.251095	-0.674237	0.832533	0.406721	-0.855300	-0.511826	2
47405	-0.828260	0.577104	-0.684357	-1.159249	-0.679049	1.129472	2.059056	0.251095	-0.573226	0.835777	1.007829	-0.980842	0.518017	2
47406	-1.208366	-0.622244	-1.151981	-1.598925	-0.679049	1.586045	2.059056	0.251095	-0.494942	0.830370	0.898536	-0.980842	0.518017	2

Appendix C

IEEE Copyright and Consent Forms Paper

- Copyright © 2021, IEEE. Najiya Omar, Hamed Aly, and Timothy 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 2021 4th International Conference on Energy, Electrical and Power Engineering (CEEPE), pages 1130–1135. IEEE, 2021.

- Najiya Omar, Hamed Aly, and Timothy Little. Optimized feature selection based on a least-redundant and highest-relevant framework for a solar irradiance forecasting model. IEEE Access, 10:48643–48659, 2022.

- Copyright © 1969, IEEE. Najiya Omar, Hamed Aly, and Timothy Little. Seasonal clustering forecasting technique for intelligent hourly solar irradiance systems. IEEE Transactions on Industrial Informatics, pages 1–1, 2022.

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

LSTM and RBFNN Based Univariate and Multivariate Forecasting of Day-ahead Solar Irradiance for Atlantic Region in Canada and Mediterranean Region in Libya
Najiya Omar, Hamed Aly, Timothy Little
2021 4th International Conference on Energy, Electrical and Power Engineering (CEEPE)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

You have indicated that you DO wish to have video/audio recordings made of your conference presentation under terms and conditions set forth in "Consent and Release."

CONSENT AND RELEASE

1. In the event the author makes a presentation based upon the Work at a conference hosted or sponsored in whole or in part by the IEEE, the author, in consideration for his/her participation in the conference, hereby grants the IEEE the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive, in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IEEE and live or recorded broadcast of the Presentation during or after the conference.
2. In connection with the permission granted in Section 1, the author hereby grants IEEE the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IEEE from any claim based on

right of privacy or publicity.

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Najiya Omar	07-04-2021
Signature	**Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any

previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

Creative Commons Attribution License (CCBY)

Optimized Feature Selection Based on a Least-Redundant and Highest-Relevant Framework for a Solar Irradiance Forecasting Model
NAJIYA OMAR,IEEE),HAMED ALY,TIMOTHY LITTLE
IEEE Access

By clicking the checkbox at the bottom of this page you, as the author or representative of the author, confirm that your work is licensed to IEEE under the Creative Commons Attribution 4.0(CCBY 4.0). As explained by the Creative Commons web site, this license states that IEEE is free to share, copy, distribute and transmit your work under the following conditions:

> Attribution - Users must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse the users or their use of the work).

With the understanding that:

> **Waiver** - Any of the above conditions can be waived if users get permission from the copyright holder.
>
> **Public Domain** - Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.
>
> **Other Rights** - In no way are any of the following rights affected by the license:
>
> - A user's fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
> - The author's moral rights;
> - Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

For any reuse or distribution, users must make clear to others the license terms of this work.

Upon clicking on the checkbox below, you will not only confirm that your submission is under the CCBY license but you will also be taken to IEEE's Terms of Use, which will require your signature.

[X] I confirm the submitted work is licensed to IEEE under the Creative Commons Attribution 4.0 United States (CCBY 4.0)

TERMS AND CONDITIONS OF AN AUTHOR'S USE OF THE CREATIVE COMMONS ATTRIBUTION LICENSE (CCBY)

1. Creative Commons Licensing

To grow the commons of free knowledge and free culture, all users are required to grant broad permissions to the general public to re-distribute and re-use their contributions freely. Therefore, for any text, figures, or other work in any medium you hold the copyright to, by submitting it, you agree to license it under the Creative Commons Attribution 4.0 Unported License.

2. Attribution

As an author, you agree to be attributed in any of the following fashions: a) through a hyperlink (where possible) or URL to the article or articles you contributed to, b) through a hyperlink (where possible) or URL to an alternative, stable online copy which is freely accessible, which conforms with the license, and which provides credit to the authors in a manner equivalent to the credit given on this website, or c) through a list of all authors.

3. Terms of Publication

A. By submitting your work to IEEE, you agree to comply with the IEEE Publication Services and Products Board Operations

Manual (the "Operations Manual"), including, but not limited to, the specific provisions referenced herein(except to the extent any provision of the Operations Manual requires assignment of copyright in your work to IEEE).

B. Submission to this IEEE journal does not guarantee publication. By submitting your work to this journal you, as author, recognize that your work may be rejected for any reason. All submissions shall be reviewed by the Editor in accordance with section 8.2.2 of the Operations Manual.

C. Should your paper be rejected IEEE will not exercise any of the rights granted to it under the Creative Commons Attribution 4.0 Unported License.

D. IEEE takes intellectual property protection seriously and is opposed to plagiarism in any fashion. Accordingly, you consent to having your work submitted to a plagiarism detection tool and to be bound by IEEE policies concerning plagiarism and author misconduct.

E. IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. You must ensure that your work meets the requirements as stated in section 8.2.1 of the Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at https://www.ieee.org/publications/rights/author-rights-responsibilities.html.

F. You warrant that your work, including and any accompanying materials, is original and that you are the author of the work. To the extent your work incorporates text passages, figures, data or other material from the works of others, you represent and warrant that you have obtained all third party permissions and consents to grant the rights herein and have provided copies of such permissions and consents to IEEE. As stated in section 8.2.1B12 of the Operations Manual: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it."

G. You are advised of Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

H. You agree that publication of a notice of violation as a corrective action for a confirmed case of plagiarism, as described in Section 8.2.4 of the IEEE PSPB Publications Operations Manual, does not violate any of your moral rights.

I. You agree to indemnify and hold IEEE and its parents, subsidiaries, affiliates, officers, employees, agents, partners and licensors harmless from any claim or demand, including reasonable attorneys' fees, due to or arising out of: (1) content you submit, post, transmit or otherwise make available through IEEE's publishing program; (2) your use of this IEEE journal; (3) your violation of these Terms of Use; or (4) your violation of any rights of another party.

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Najiya Omar	27-04-2022
Signature	**Date**

Questions about the submission of the form or manuscript must be sent to the publication's editor. Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

IEEE COPYRIGHT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Seasonal Clustering Forecasting Technique forIntelligent Hourly Solar Irradiance Systems
Omar, Najiya; Aly, Hamed; Timothy, Little
IEEE Transactions on Industrial Informatics

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Najiya Omar	22-05-2022
Signature	**Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966