Forecasting Minute Averaged Solar Irradiance Using Machine Learning for Solar Collector Applications

By

Ronewa Collen Nemalili

14004811

A Research Project Report submitted to the Faculty of Science, Engineering and Agriculture, in partial fulfilment of the requirements for the degree of Master of Science in e-Science

in the

Department of Physics

University of Venda

Supervisor

Dr. L. Jhamba

Co-supervisors

Dr. J.K. Kirui and Dr. C. Sigauke

11 January 2023

Declaration

I, **Ronewa Collen Nemalili**, declare that this report is my own, unaided work. It is being Submitted for the degree of Master of Science in e-Science at the University of Venda. It has not been submitted for any degree or examination at any other university.

NRC

11 January 2023

Abstract

Challenges in utilising fossil fuels for generating energy call for the use of renewable energy. This study focuses on modelling and forecasting solar energy and optimum tilt angle of solar energy acceptance using historical time series data collected from one of the South African radiometric stations, USAid Venda station in Limpopo province. In the study we carried out a comparative analysis of Random Forest and Bayesian linear regression in short-term forecasting of global horizontal irradiance (GHI). To compare the predictive accuracy of the models, k-Nearest Neighbors (KNN) and Long short-term memory (LSTM) are used as benchmark models. The top two models with the best performances were then used in hourly forecasting of optimum tilt angles for harvesting solar energy. The performance measures such as MAE, MSE, and RMSE were used and the results showed RF to have better performance in forecasting GHI than other models, followed by the LSTM and the third best model was the KNN whereas the BLR was the least performing model. RF and LSTM were then used in modelling and forecasting the tilt angles of optimal solar energy acceptance and as thus, the LSTM outperformed the RF by a small margin.

Acknowledgements

I would like to thank all the people who contributed to my research project, starting with my supervisor, Dr. L. Jhamba and my co-supervisors, Dr. J.K. Kirui and Dr. C. Sigauke. Furthermore, I render hearty gratitude to my sponsor, National e-Science Postgraduate Teaching and Training Platform (NEPTTP) for sponsoring my studies.

Contents

List of Tables

List of Abbreviations

PV	**P**hoto**V**oltaic
CNN	**C**onvolutional **N**eural **N**etwork
LSTM	**L**ong **S**hort-**T**erm **M**emory
SVM	**S**upport **V**ector **M**achines
GHI	**G**lobal **H**orizontal **I**rradiance
SAURAN	**S**outhern **A**frican **U**niversities **R**adiometric **N**etwork
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **A**verage
BPNN	**B**ack **P**ropagation **N**eural **N**etwork
ANN	**A**rtificial **N**eural **N**etwork
KNN	**K**-**N**earest **N**eighbors
BLR	**B**ayesian **L**inear **R**egression
RF	**R**andom **F**orest
SFP	**S**tatistical **F**eature **P**arameters
GA	**G**enetic **A**lgorithm
FFNN	**F**eed **F**orward **N**eural **N**etwork
SVR	**S**upport **V**ector **R**egression
NWP	**N**umerical **W**eather **P**rediction
PCR	**P**rincipal **C**omponent **R**egression
LR	**L**inear **R**egression
GB	**G**radient **B**oosting

Chapter 1

Introduction

1.1 Background

The constant increase in global energy consumption has led to worsening in the shortage of primary energy resources such as fossil fuels, coal, and natural gas globally with South Africa experiencing electricity load shedding since 2005 due to energy generation and maintenance difficulties to keep up with the economic growth [1]. These, together with an increase in fossil fuels prices have compelled improved developments in such renewable energy aspects as forecasting the performance of solar energy production for the improvement of the operational performance of a solar system and for scheduling the maintenance of the solar power [2] [3] [4]. The growth of solar renewable energy has been increasing rapidly, with the total installed PVs reaching over 67.4 gigawatts and 627 gigawatts globally at the end of 2011 and 2019 respectively. Studies show that it will keep on increasing, while fossil fuels decrease [2].

It has been estimated that the earth receives solar irradiance of around 1000 W/m^2 and this kind of irradiance is estimated to generate around 85,000TW overall [5]. However, according to the Renewable Global Status Report, the solar energy contributed every day was only less than one percent of the world's energy demand between 2006 and 2007 and increased to over 27% on average by 2019 and 2020 respectively [6]. This level of performance was based on the type of semiconductor utilized to build solar panels since they were only capable of converting 15% to 40% of sunlight to electricity depending on the type of semiconductor used, and also there were not enough solar systems due to their expensiveness. To increase the

PV performances without increasing the cost, some studies focused on estimating the solar panel's tilt angles of optimal irradiation and solar collector's acceptance angle to maximize the amount of solar radiations [5].

Plenty of studies focused on forecasting the behaviour of solar irradiance using different techniques [7],[8],[9],[10],[11]. Since the behaviour of solar irradiance depends on the independent weather conditions which are always changing, it is not easy to bring balance in the power prediction of renewable energy. Since solar energy production depends on sunlight, which is also independent, some studies have focused on predicting the optimal angle of acceptance for solar panels and non-tracking solar concentrators. An accurate estimate of solar irradiance production and optimal angle of acceptance can increase the stability of solar power performance. Operators can take more economic operating decisions for the power systems. In this study, we focus on modelling and forecasting global horizontal solar irradiance and optimum tilt angle of solar irradiance acceptance using machine learning models making use of datasets from Southern African Universities Radiometric Network (SAURAN) [12].

1.2 Statement of the Problem

Many studies in the analysis of solar irradiance using machine learning techniques/models involve the use of intricate implementation models which in most cases have datasets from multiple stations at once [13],[14],[9]. Simple models like random forest (RF), and Bayesian linear regression (BLR) have lately shown good raw data approximations [15], but still need an improvement to reach the level of intricate models with better performances. This project uses two simple models, RF and BLR to predict minute-averaged global horizontal solar irradiance at USAid Venda station situated at Vuwani Science Research Centre in Vhembe District, Limpopo and benchmark them against the reliably proven KNN and LSTM forecast models. And since there are just a handful of studies in the improvements of solar irradiance, we also focus on increasing the scattering of solar irradiance by forecasting hourly averaged tilt angle of acceptance for non-tracking solar collectors making use of the same dataset averaged to hourly and two methods which will show superior performances in forecasting solar irradiance.

1.3 Subproblems

- How does the RF, BLR, and KNN perform compared to the LSTM model in forecasting solar irradiance when used with a dataset from one station at Vuwani Science Research Centre station?

- Can the resulting superior methods improve the scattering of solar irradiance by estimating the optimal tilt angle of acceptance for non-tracking solar collectors with the same dataset from Vuwani station?

1.4 Research Aims and Objectives

1.4.1 Aims

- To compare the predictive performance of random forest, Bayesian linear regression, LSTM, and KNN in forecasting minute-averaged global horizontal irradiance using a dataset from Vuwani science research Centre.

- To study the performance of superior models in forecasting the optimal hourly averaged tilt angle of acceptance and utilize the ideal way to obtain profiles of the tilt angle of acceptance for solar collectors.

1.4.2 Objectives

The objectives of this project are to:

- prepare the dataset for implementation of solar irradiance forecasting machine learning models,

- apply RF, Bayesian linear regression, KNN and LSTM models in forecasting minute-averaged global horizontal solar irradiance,

- choose methods with superior performances and apply them in forecasting hourly averaged tilt angle of acceptance for solar irradiance,
- evaluate the performances of the models and compare them against each other for solar irradiance predictions,
- obtain profiles of hourly averaged tilt angles of acceptance for solar collectors.

1.5 Overview

In Chapter 2, related work is reviewed and gaps to be filled by this research are posited. Chapter 3, explains the nature and source of data used including models and methods followed by analysis. Chapter 4 consists of the results and discussions. Chapter 5 is the conclusion and possible future work to be carried out.

Chapter 2

Related Work and Review

2.1 Literature

A great deal of solar irradiance forecasting models has been developed in the past. These models can be grouped into three main classes which are physical, statistical, and hybrid models [16]. Physical models describe the physical relationship between solar power generated, weather conditions, and solar radiation. They also make use of satellite data to keep track of the cloud's movement speed and direction. Hybrid models are often a combination of physical and statistical models or several statistical models. Statistical models focus on using time series datasets without the use of satellite datasets involved [17]. An abundant number of projects in forecasting solar energy have been carried out using statistical methods [17].

2.1.1 Related work

A great deal of projects making use of all those methods has been constantly carried out, A related study by Zhen Zhang [8] aimed to predict the second-level solar irradiance using deep learning based on CNN with the use of wavelet transformation (WT) and bootstrap sampling methods in the time-frequency domain. The method used extracts important information in the image dataset at the frequency domain to form the next second solar irradiance. The study proved to have better accuracy on a one-second scale with more performance in a stable irradiance transformation process. A hybrid model by Chang [18] was proposed where a radial basis function neural network (RBFNN) with decoupling method was presented to forecast day-ahead PV power generation. The results were compared with autoregressive integrated moving average (ARIMA), RBFNN, and back-propagation neural network (BPNN), hence the proposed method provided higher forecasting accuracy compared to ARIMA, BPNN, and RBFNN. Another project carried out by Matteo [14] used hybrid methods to forecast solar power using clear-sky models and artificial neural networks making use of a day-ahead weather forecasting with the use of a decision tree to select the best performing model by considering their accuracy. For the proposed novel hybrid model to produce the best results, the decision tree would select the appropriate model and reflect the intuition behind the use of that model.

Some of the work completed on forecasting solar energy were performed by [3] where they proposed an artificial neural network (ANN) model using statistical feature parameters (ANN-SFP) to improve the forecast accuracy of Short-term solar irradiance forecasting (STSIF). The SFP used was first reconstructed and then the sufficient information was effectively extracted from relatively few inputs and the model complexity was reduced. The results found indicated the forecasting accuracy improved under variable weather conditions. A paper by Cyril Voyant [19] describes an overview of the forecasting methods for solar irradiance using machine learning approaches, It was shown that a lot of papers pay attention to techniques like neural networks and support vector regression due to their superior performances and other techniques like regression tree, random forest, gradient boosting and many others are being used but their performance ranking is complicated due to the diversity of the data set, timestep, forecasting horizon, set-up and performance indicators. It was then indicated that, overall, the best classical models are ANN and ARIMA models and are equivalent in terms of their quality of prediction in certain variability conditions but describe ANN as the most reliable due to its flexibility as a universal non-linear approximation. Another study [9] on solar energy was focused on using a new approach to online forecasting of power production from PV systems where Linear regression, autoregressive model, and autoregressive with exogenous inputs (ARX model) were compared in forecasting hourly values of solar power for horizons of up to 36 hours using 15-min observations dataset of solar power from 21 PV systems located on rooftops in a small village in Denmark. The models were trained and evaluated, and the results indicated that ARX outperformed the LR and AR models in forecasting solar power output, the study made use of NWP datasets for LR and ARX while the AR trained using the historical datasets of the output power. A study by Hugo T.C Pedro [10] evaluates and compares models such as the Persistent model, ARIMA, KNNs, ANNs, and ANNs optimized by Genetic Algorithms (GA). The study focused on using one NWP dataset accumulated in Merced, California to train the models for the 1 and 2 h-ahead hourly averaged power output. Results showed the ANNs technique being the best performing model compared to the rest of the used models, and it was concluded that the performance can be improved with the use of GA optimization of the ANN parameters, and the performance of all models used depends strongly on seasonal characteristics of solar variability. Tendani Mutavhatsindi [11] focused on forecasting hourly solar irradiance using LSTM, Feed-forward neural network (FFNN), and PCR model as a benchmark. The study was done using a statistical dataset obtained from the SAURAN network, University of Pretoria radiometric station. The results found showed FFNN to be the best performing model compared to other models used. The benchmark model was outperformed by all models used. In addition to the results found, the study was then accelerated by combining the used models to improve the accuracy performance, these were done by using convex combination and quantile regression averaging (QRA). Hence, the QRA was found to perform better compared with all the stand-alone models used. It was concluded that to obtain more insightful models with better performances, more statistical tests and detailed evaluation metrics in forecasting should be done to understand and get the datasets prepared correctly. A study carried out by Ratshilengo [20] focused on short-term forecasting of high-frequency global horizontal irradiance data from Vuwani science research centre radiometric station. In the study, they compared performances of the genetic algorithm and recurrent neural network models with the K-nearest neighbour model as a benchmark model in forecasting global solar irradiance. Hence, the empirical results from the study showed that the genetic algorithm model had the best performing results compared to the other two models including the benchmark model. A study [21] by Jamil focused on estimating both solar radiation and optimum tilt angles in Aligarh city using dataset observed at Heat Transfer and Solar Energy Laboratory, Department of Mechanical Engineering, Aligarh Muslim University, Aligarh to estimate monthly average solar radiation and monthly, seasonal, and annual optimum tilt angles were also estimated.

Not many studies focused on finding the optimal tilt angle of acceptance for solar collectors to maximize the amount of solar irradiance constantly received [22]. In the study [5] by Kim they proposed a solar panel tilt angle optimisation model using machine learning algorithms. Their objective was to find a tilt angle that maximizes the solar irradiance on the PV systems. They made use of linear regression (LR), random forest, SVM, gradient boosting (GB), and least absolute shrinkage and selection operator (LASSO) with the solar power generation dataset from 22 PV modules involving various factors such as weather, dust level, and aerosol level. It was concluded that the GB was found to be the best performing model to use in forecasting monthly/yearly panel tilt by predicting the best tilt angles with respect to the amount of solar power generated. Another study was carried out by Manoj Kumar [23] where they focused on the optimization of the tilt angle of solar panels using artificial intelligence and solar tracking method. An analysis on the optimization of tilt angle was done and the tilt angles were optimized monthly in Kolkata, India on the latitude: 22.56670 N and longitude: 88.36670E using genetic algorithm and thus, it was proved that a significant power gain could be obtained by finding optimal tilt angle of acceptance. The study [24] by Ramaneti proposed a solution to efficient solar power by tracking the sun's relative position to earth and finding the tilt angle of the solar panel, making use of the deep learning method. The proposed method was able to predict the tilt and orientation angle of the sun, resulting in the increase of solar power by 10.6%. Chih-Chiang [25] used forecasting models such as multilayer perceptron (MLP), random forests (RF), k-Nearest Neighbours (kNN), and linear regression (LR) with the Satellite Remote-Sensing Dataset to estimate surface solar radiation on an hourly basis and the solar irradiance received by the solar panels at different tilt angles in Tainan, southern Taiwan. The study showed good performances in MLP than in RF and KNN while the LR had the least performance.

2.1.2 Conclusion from the Related Work

Most studies with outstanding results made use of the intricate techniques/models and others focused on the hybrid models. Some studies like [19] described models which are not complex and which have inconsistent results based on the types of datasets utilized. It was deduced that a lot of studies focused on using the datasets from different sources at once. In this study, models such as RF and BLR, which are considered to have inconsistent results are trained on the dataset from one source and their results are compared with the benchmark models, KNN and LSTM, which are also trained on the same dataset. Not enough studies have been carried out in forecasting the tilt angle of acceptance for solar irradiance. More studies using machine learning models in modelling and forecasting tilt angles of optimal solar irradiance acceptance should be carried out.

2.2 Models

2.2.1 Random Forest

A Random Forest (RF) is a machine learning supervised technique/model that utilizes ensemble learning to solve both regression and classification problems. It was trademarked by Leo Breiman and Adele Cutler in 2006 as an extension from Tim Kam Ho who created it in 1995 [26]. It operates by generating multiple decision trees during training time through taking a sample training data set randomly with replacement and ensemble them using an algorithm like bagging and boosting to add more diversity and reducing the correlation among ensembled decision trees. The predictions vary depending on the type of the problem being solved, for regression the average of all decision trees ensembled is considered as the prediction, whereas for classification the majority vote of the frequent categorical class is considered as the predicted class. The algorithm for RF can be given as follows,

- Given a training dataset $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$,

- random sample with replacement of the training set is selected A times and fits trees.

- Selected sample number X_a, Y_a with n training examples train a regression tree f_a on X_a, Y_a, where $a = 1, ..., A$ and the process happens for A times,

- After training, predictions for unseen samples $x^{'}$ is then made by averaging the predictions from all the individual regression trees on $x^{'}$ using the formula:

$$\hat{f} = \frac{1}{A}\sum_{a=1}^{B} f_a(x^{'}) \, , \tag{2.1}$$

2.2.2 Bayesian Linear Regression

Bayesian Linear Regression (BLR) is an approach to Linear Regression (LR) in which the statistical analysis is undertaken within the context of Bayesian inference, LR formulates its estimating function based on point estimates using training dataset X and y to find the best $\hat{\beta}^T$ to build a robust function $\hat{y}$ that can be used for estimating output value of a new data points.

$$\hat{y} = \hat{\beta}^T X \, , \tag{2.2}$$

Unlike the LR which focuses on point estimate, BLR focuses on using the Bayes theorem to find out the values of regression coefficients. The model makes use of posterior probability distributions where y as the response is assumed to be drawn from a probability distribution of the credited interval. The BLR formulation is given by

$$y \rightarrow \left(\beta^T X, \sigma^2 I\right) \, , \tag{2.3}$$

where y is output generated from the normal distribution characterized by the mean, $\beta^T X$, and variance, $\sigma^2 I$ multiplied by the identity matrix. BLR aims to determine the marginal posterior distribution for all the model's parameters including β

and intercept of the function. As shown below, Bayes' theorem finds the probability of an event occurring, given the probability of an already occurred event otherwise, it makes use of non-informative prior if there is no event that has already happened.

$$Posterior = \frac{Likelihood * Prior}{Normalization}, \tag{2.4}$$

From Equation 2.4 the posterior probability distribution of the model's parameters given the inputs and outputs is equal to the likelihood of the data, $P(y|\beta, X)$, multiplied by the prior, $P(\beta|X)$ probability of the parameters and divided by a normalization constant, $P(y|X)$.

$$P(\beta|y, X) = \frac{P(y|\beta, X) * P(\beta|X)}{P(y|X)}, \tag{2.5}$$

2.2.3 K-Nearest Neighbours

K-Nearest Neighbour (KNN) is a supervised machine learning model that can be used for both regression and classification. Like any other supervised models, it makes use of the dataset with labels training measurements X, y to find the link between them by discovering a function $h : X \rightarrow y$ which is then used for the new dataset to predict the target output. KNN is non-parametric and both classifier and regressor but in this project, we focus on using regressor since we are working with a time-series dataset. It follows the concept of finding the nearest neighbours where the closest points to the query are being found. This is done by calculating the distance from a query example to the labelled examples using a distance formula like Minkowski, Hamming, and Euclidean distance.

In this project, the Euclidean distance in equation 2.6 was used to calculate the distances between all observations and the query point where 2.7 represent the simplified Euclidean distance

$$d(X, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}, \tag{2.6}$$

$$= \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\,, \tag{2.7}$$

The calculated distances were sorted by increasing order. Then the heuristically optimal number k of nearest neighbours was found and finally the most frequent label is voted in the case of classification problem whereas the labels are averaged for a regression problem.

2.2.4 Long Short Term Memory

Long Short Term Memory (LSTM) is a complex artificial recurrent neural network (RNN) architecture used in the field of deep learning and usually deals with a sequential dataset like time series. It was introduced by Hochreiter and Schmidhuber (1997) and subsequently refined and popularized by many researchers [27]. Instead of a Recurrent neural network (RNN), Each layer in the LSTM can be described in four steps which are Forget gate, Input gate, Update gate, and the Cell output gate,

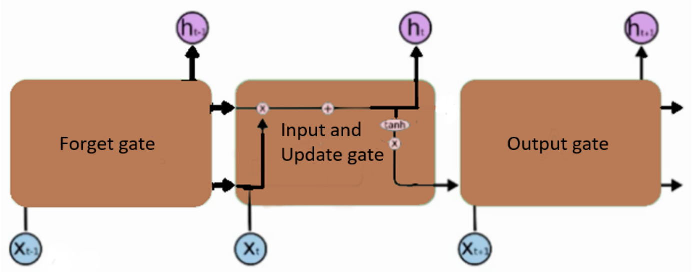

FIGURE 2.1: LSTM architecture

1. **Forget gate**: The model decides what information should be thrown away or kept for training, where equation 2.8 represent the forget gate with W_t and b_f not dependent on time.

$$f_t = g(W_t[x_t, h_{t-1} - b_f])\,, \tag{2.8}$$

2. **Input gate**: The input gate decides what information is relevant to add from the current step using the sigmoid activation function and produces the output as a hidden state represented in equation 2.10 while input gate represented in 2.9, and the weight and bias are not time-dependent.

$$i_t = g(W_i[x_t, h_{t-1} - b_i]) \ , \tag{2.9}$$

$$O_t = g(W_O[x_t, h_{t-1} - b_O]) \ , \tag{2.10}$$

3. **Update gate(Cell state)**: Takes the output from the input gate and does a point-wise addition (with the hyperbolic tangent as an activation function as shown in eq 2.11) which updates the cell state to new values that the neural network finds relevant.

$$C_t = f_t C_{t-1} + i_t tanh(g(W_C[x_t, h_{t-1} - b_C])) \ , \tag{2.11}$$

4. **The cell output state**: As shown in eq 2.12, It uses the sigmoid function to decide which information should be going to the next hidden state with the use of output state and cell state.

$$h_t = O_t tanh(C_t) \ , \tag{2.12}$$

Where the above equations from 2.8 - 2.12 are being computed every time step and weights and bias are not time dependent for all equations. The equations variables have the following meanings:

g is the sigmoid function,
$tanh$ is the hyperbolic tangent,
x_t is the input vector,
h_t is the output vector,
C_t is a cell state vector,
W represents weights and b are biases,
f_t , i_t , and O_t are gates of the block.

Chapter 3

Research Methodology

3.1 Data

3.1.1 Data Collection

FIGURE 3.1: USAid Venda station.

The study made use of a minute-averaged solar irradiance dataset from one station called USAid Venda station https://sauran.ac.za/ situated in South Africa in the province of Limpopo, on the latitude -23.13100052 and longitude 30.42399979 and the elevation 628 m [12]. Figure 3.1 shows the USAid Venda station where the paranometer used to record the datasets is situated. Due to storage of the computer, dataset was accumulated for three years from 2015-01-01 to 2017-12-31 instead of five years and this had a little effect on the models performance because the data was still large enough for analysis and models predictive accuracy increased by a very small margin. The data featured 21 variables (namely: Time Stamp, GHI

Avg(W/m^2), DIF Avg(W/m^2), DNI Avg(W/m^2), Calculated DNI Avg(W/m^2), Temperature Avg(Deg C), Relative Humidity(%), Total Rain(mm), Wind Speed Avg(m/s), Wind Vector Magnitude Avg(m/s), Wind Direction Avg(Deg), Wind Direction StdDev(Deg), Wind Speed Max(m/s), Barometric Pressure Avg(mbar), 12V Battery Avg(volts), 12V Battery min(volts), 24V Battery Avg(volts), 24V Battery min(volts), Logger Temperature Avg(Deg C), Calculated Azimuth angle(Deg), and Calculated Tilt angle(Deg)) before selecting those with the most significance in forecasting optimal global horizontal irradiance (GHI) or tilt angle.

3.2 GHI Models

The modelling and exploration of dataset were performed using python programming language in jupyter notebook installed with libraries such as keras, tensorflow, and scikit-learn. A special environment was created for PyMC3 in order to perform the bayesian inference. In the modelling part, data was first explored with the aim to study its characteristics. and as thus, denoising for the target feature was performed to reduce noise using lfilter method. The independent features were scaled to between 0 and 1 using min-max scalar to make it easy for the models to understand the problem of the study. And the feature selection was performed using LASSO regression method to help increase the performances of the models by removing features with no significance. Since there was no solar irradiance at night hours, the dataset recorded between 18H00 and 04H59 was then removed in order to increase the accuracy of the models by training on more informative dataset.

3.2.1 Random Forest Application

RF was modelled and trained on the dataset prepared, the model was first trained several times to find the best split ratio using scikit-learn, and then with the split ratio of 60% in training and 40% in testing. Secondly, the grid search was performed to find the best hyperparameters. The suitable hyperparameters were obtained and the model was trained with the number of trees set to 400. Finally the model was tested on the testing data, and the prediction were done and the performance measures were calculated and recorded in Chapter 4, Table 4.2.

3.2.2 Bayesian Linear Regression modelling

After training models with different split ratio data to find the best split ratio of data, the BLR was trained with split ratio of 70% data in training and 30% of the data was used in testing. The model was trained using PyMC3 package [28], a package which makes BLR models easier to work with. The formula for BLR was first generated from the dataset features. Then the prior and likelihood of the model was introduced, from the dataset characteristics, the prior used was a normal distribution while the likelihood was traced through sampling 4 chains and 4 jobs by default using Markov chain Monte Carlo method. The sampling were performed for 1000 tunes and 2000 draws iterations. The model was then tested on the new dataset in order to calculate its performance errors.

3.2.3 K-Nearest Neighbour Application

The models were built with the split ratio of the dataset set to 65% of training data and 35% of the testing data. The grid search for performed and the model was trained on the 65% of the dataset and the suitable k value for optimal performance was found through grid search method. After training, its performance error was then calculated on the testing data using the performance measures stated in Section 3.4 and the results were recorded in Table 4.2.

3.2.4 Long Short Term Memory Application

Prepared data were reshaped from 2 dimensional to 3 dimensional by first converting them into numpy array and then make use of reshape function . The best split ratio was found to be 60% in training and 40% in testing. The LSTM was built to use two layers starting from 64 units to 32 units and one dropout of 20%. The model was assigned to 1200 epochs and early stopping set to 50 patience, and the batch size was set to 200. After training, testing dataset was then used to calculate the performance measures making use of the testing dataset.

3.3 Tilt angles Models

The minute-averaged dataset accumulated was first converted to hourly-averaged. Resulted data was then used in modelling for hourly tilt angle of acceptance for solar irradiance. Feature selection was performed using LASSO regression method, and the obtained features were scaled to between 0 and 1 using min-max scalar. Top two of the best models in forecasting GHI, which are the RF and LSTM, were used.

3.3.1 Random Forest Application

The data was first split into the perfect ratio for best performance through training the model several times with different split ratio, and the grid search method was performed to find the hyperparameters with good effect in improving the performance of the model. The model was trained with number of trees set to 430 as explained more in Section 4.2 of chapter 4. The resulting model was then tested using test dataset and the performance measures in in Section 3.4 of chapter 3, and the results obtained were recorded in Table 4.4.

3.3.2 Long Short Term Memory Application

The LSTM also made use of the same dataset converted to hourly and prepared as explained above. The independent variables were first reshaped from 2d to 3d with a timestep of 1, and split the data into 60% training and 40% testing. The model was built with two layers starting with 64 units, and then followed by a layer with 32 units without dropout in between but at the end of the layers. After training several times while tuning hyperparameters differently, the training was assigned to 500 epochs, and set to be patience for 40 epochs. Obtained model was then tested using the testing data making use of the same performance measures stated in 3.4, and the results are recorded in Chapter 4, Table 4.4.

3.4 Performance Measures

The performance of models was measured using errors such as Mean Absolute Error (MAE), Mean Square Error (MSE), and the Root Mean Square Error (RMSE).

Mean Absolute Error:	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} \lvert Y_i - y_i \rvert$
Mean squared error:	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} (Y_i - y_i)^2$
Root Mean Square Error:	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (Y_i - y_i)}$

Where Y_i = Predicted value

y_i = Actual value

n = Total number of data points.

Chapter 4

Results and Discussion

4.1 Results and Discussion for GHI

Figure 4.1 below shows the variables selected. Variables with longer bars are more important than variables with short bars whereas those with no bars at all are not important, hence variables such as rain total, wind speed average, and wind vector magnitude average were discarded because they lacked significant effect in obtaining models with highly accurate results.

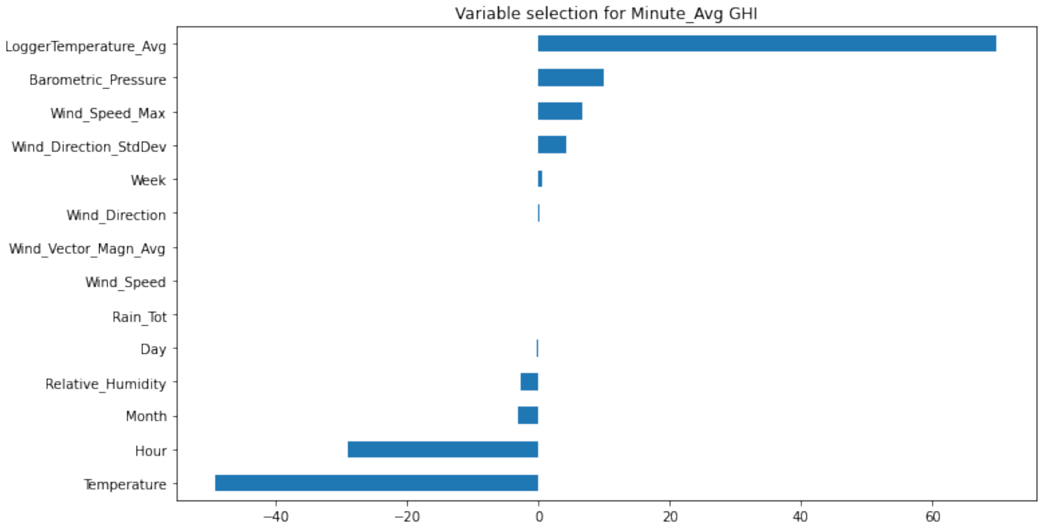

FIGURE 4.1: Variable selected using LASSO regressor method.

Variables such as temperature, hour, relative humidity, month, day, wind direction, week, and others are selected for the training process. The Figure 4.2 below shows the variables before scaling and after scaling using min-max scalar.

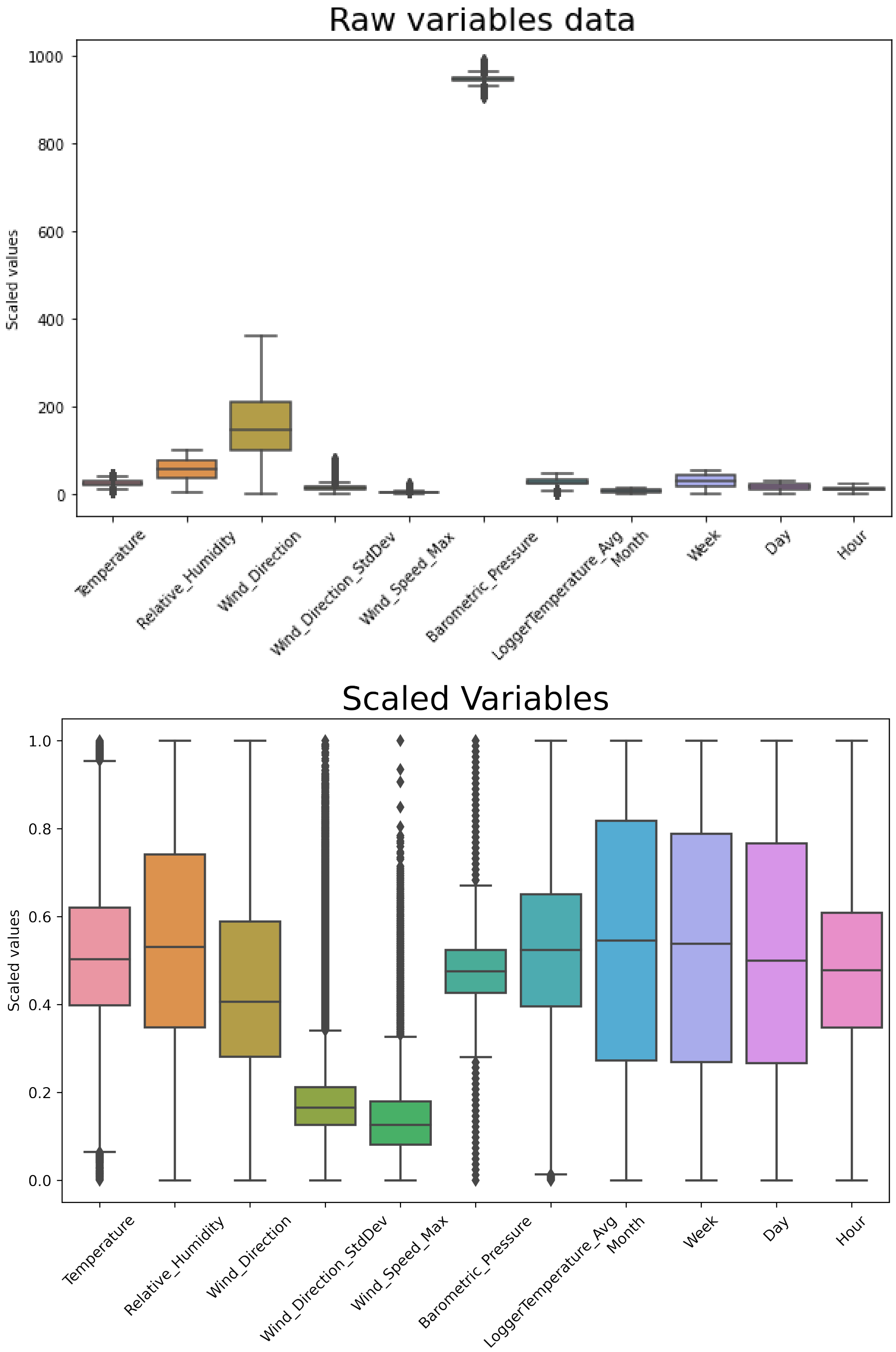

FIGURE 4.2: Comparison of the variables before scaling and after scaling dataset.

Visualization of global horizontal irradiance is depicted in Figure 4.3 which shows seasonal changes from 2015 January to 2017 December with the peak performances of GHI happening in summer times. Figure 4.3 further shows the performance of GHI for the term of three years visualized in hourly. The GHI density distribution is a descending unimodal with the mean of 405.228 W/s^2 and median of 333.1364 W/s^2 as shown also in Table 4.1 which gives the statistics of GHI measured in W/S^2. The minimum energy measured is 0.0032 W/s^2 whereas the maximum energy recorded is 1521.7970 W/s^2 with most energy measured between 110.1275 W/s^2 and 671.4653 W/s^2 as shown in the box plot in Figure 4.4.

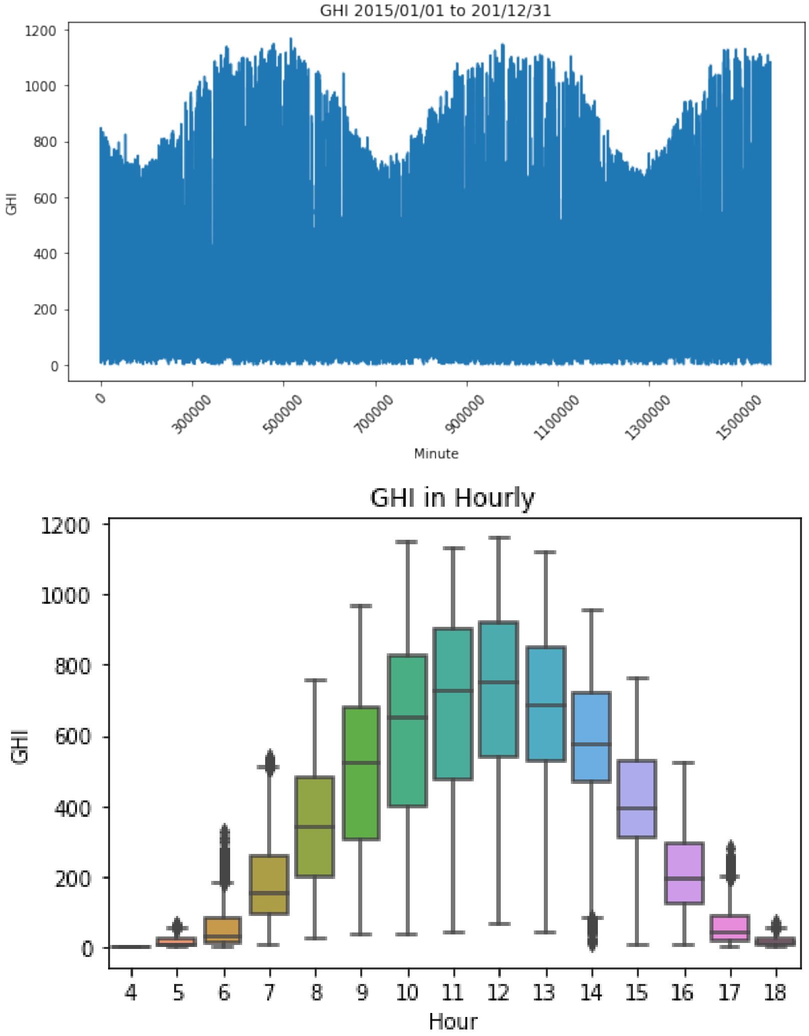

FIGURE 4.3: Visualization of GHI measured every minute from 2015 January 1st to 2017 December 31, plotted with noise reduced using lfilter method.

TABLE 4.1: GHI stats.

Number of GHI data entries	Mean	Standard dev	Min	Q1	Median	Q3	Max
732865	405.228	325.914	0.0032	110.1275	333.1364	671.4653	1521.7970

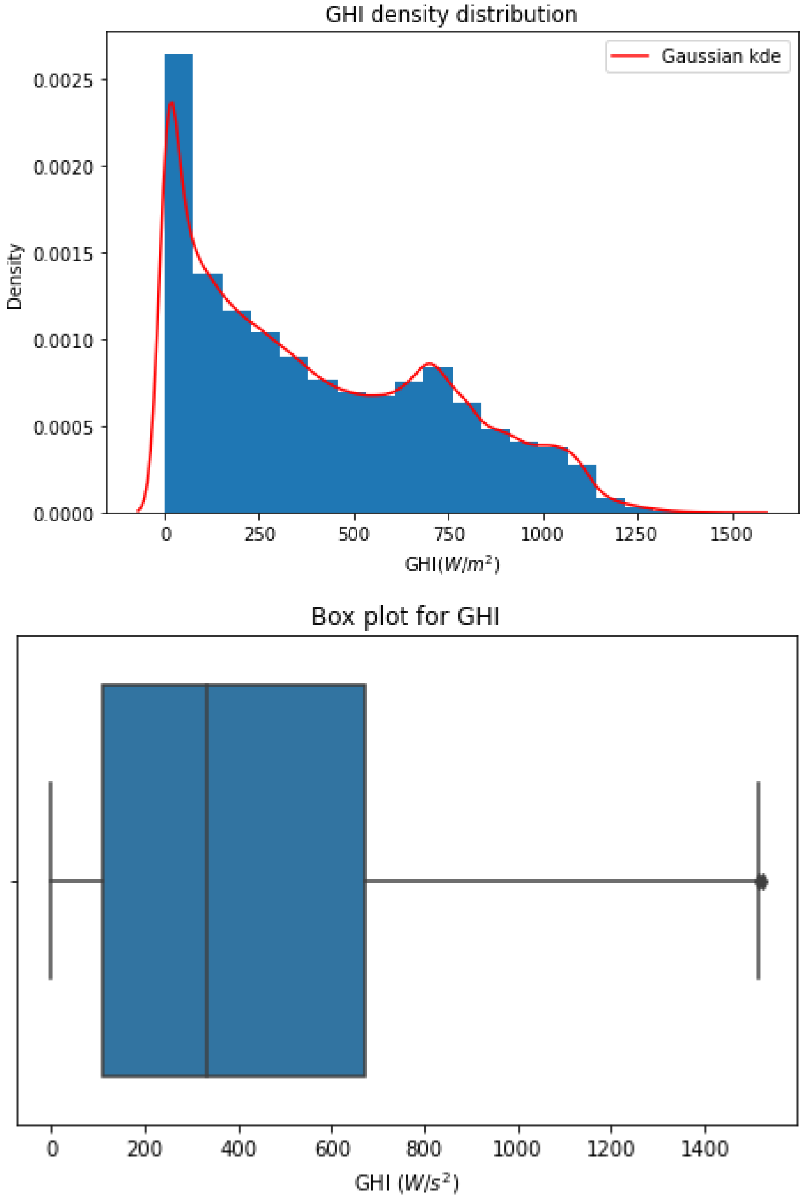

FIGURE 4.4: Histogram with a gaussian kernel density estimation fitted, and boxplot for GHI on the right.

Models were trained on 11 independent variables. The best results for random forest were obtained with a number of trees set to 400. The BLR was modelled focusing on the Markov Chain Monte Carlo (MCMC) and variational inference. Optimal results for KNN were obtained with a hyperparameter k value set to 3. Overall, the RF was found to be the best performing model compared to the LSTM, KNN, and BLR with 16.74 MAE, 899.86 MSE, and 30.0 RMSE. This is due to the simplicity of its implementation with very few hyperparameters to tune and easy to find the best with grid search method, and also the use of decision trees to train

RF makes it a robust and the best performing model. BLR is the least performing with 115,81 MAE, 147.30 RMSE, and 21697.29 MSE. The runner-up to the RF is the LSTM, and KNN is the third-best model while BLR is the least.

TABLE 4.2: Error measures results for RF, BLR, KNN, and LSTM models evaluated on testing data.

Testing	RF	BLR	KNN	LSTM
MAE	15.23	115.81	23.5	22.78
MSE	899.86	21697.29	1747.59	1656.49
RMSE	30.0	147.30	41.8	40.7

All the models showed good results including the BLR regardless of being last. BLR has specialities that the other three models do not have, it was modelled to give the posterior distribution of all possible values whilst other models only give point estimates.

The BLR trace plot in figures 4.5 and 4.6 follow Gaussian distribution and Generalized Linear Models (GLM). They reflect a marginal posterior distribution for all variables and the intercept of the BLR obtained, on the right side of Figure 4.6 are shown plots of the progression of Markov Chain Monte Carlo (MCMC). It may be noted that chains for 1000 tune and 2 000 draw iterations were sampled as shown in the distributions.

The resulting posterior distributions are characterized by the mean and variance, where the peak of each distribution is the mean of that parameter as shown in figure 4.5, and it is considered to be the maximum a posteriori estimate (MAP) which is the most likely parameter to be used in estimation before considering other values in the credible interval due to uncertainties.

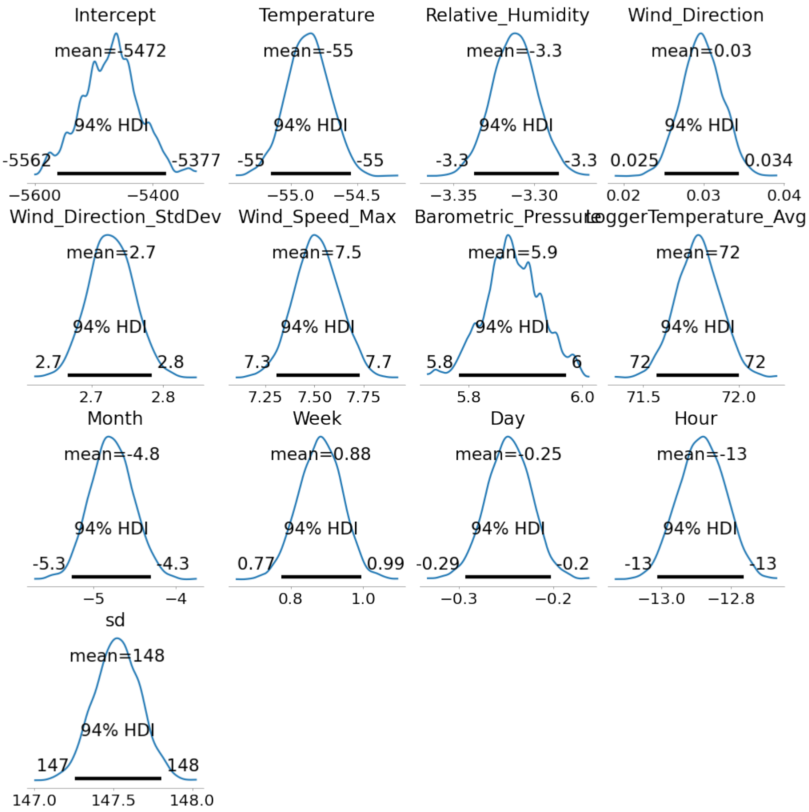

FIGURE 4.5: Posterior distribution plots with HPD and credible interval specified.

Figure 4.5 show a detailed look at the posterior distributions for all variable parameters together with the intercepts of the equation to be obtained and standard deviation. The distributions give a credible interval of all possible parameters to be used for estimations. The highest posterior density (HPD) for the credible interval is 94%, which means HPD is 94% confident that the parameters are in that interval.

Figure 4.7 below compares the probability distributions of the observed GHI and predicted GHI for RF, KNN, and LSTM. Models indicate to follow the observed density plots on the testing dataset. The BLR results vary depending on the parameters used since they are given in the whole posterior distribution showing all possible estimates.

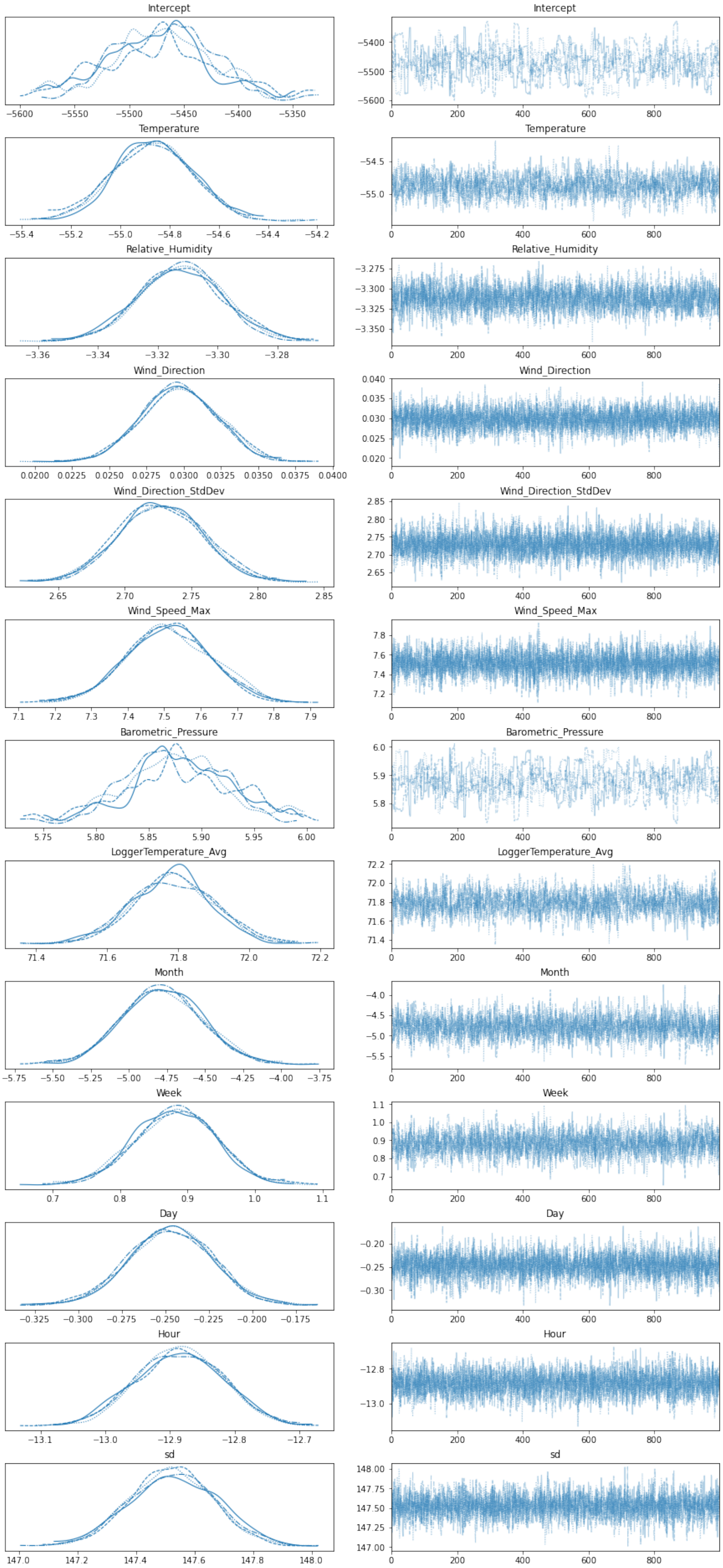

FIGURE 4.6: Posterior distributions with the sampling progression on the right.

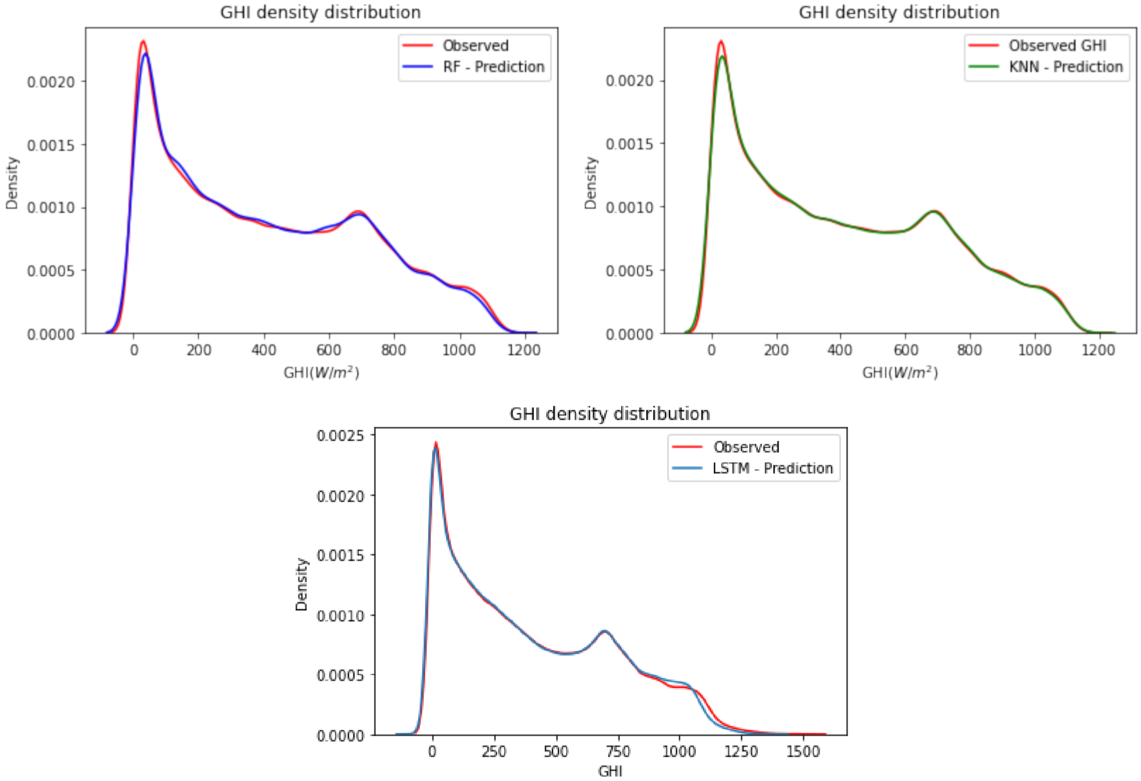

FIGURE 4.7: Probability densities for RF, KNN, LSTM evaluated on the testing dataset.

4.2 Results and Discussion for Tilt angle

The Figures 4.8, and 4.9 below show how the tilt angles were changing throughout the time for three years, with Figure 4.8 showing the raw data of tilt angles before scaling whilst Figure 4.9 is the scaled data,

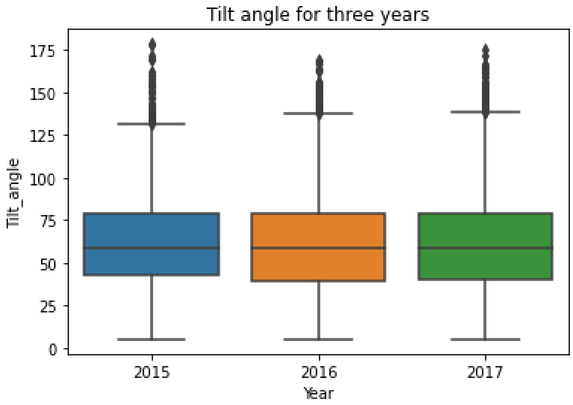

FIGURE 4.8: Box plots showing tilt angles for three years.

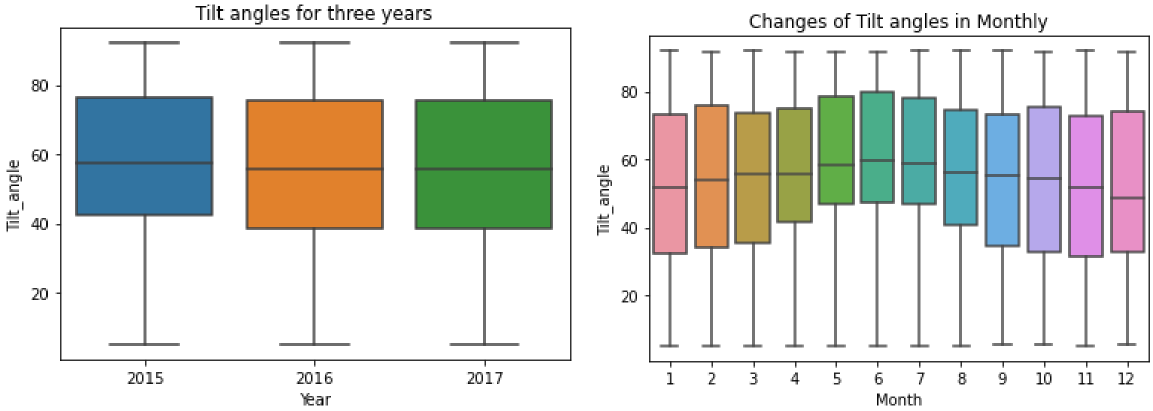

FIGURE 4.9: Box plots showing yearly and monthly tilt angles for three years.

TABLE 4.3: Hourly tilt angle's stats.

Number of data points for Tilt angles	Mean	Standard dev	Min	Q1	Median	Q3	Max
12736	55.63	22.90	5.14	39.44	56.20	75.55	92.00

Figures 4.10 and 4.11 show the change of tilt angles throughout the day in every hour and every week for the term of three years. It can be deduced that the angle in the morning is at maximum and become small as it reaches mid-day, so during mid-day where maximum GHI is scattered, the tilt angle of acceptance is at its minimum where the very minimum angle recorded is 5.014 degrees as shown in Table 4.3. The tilt angles increases from midday and reaches maximum at sunset. The tilt angle of acceptance is mostly high during winter times as shown by Figure 4.11 with a belly-shaped weekly tilt angles recording.

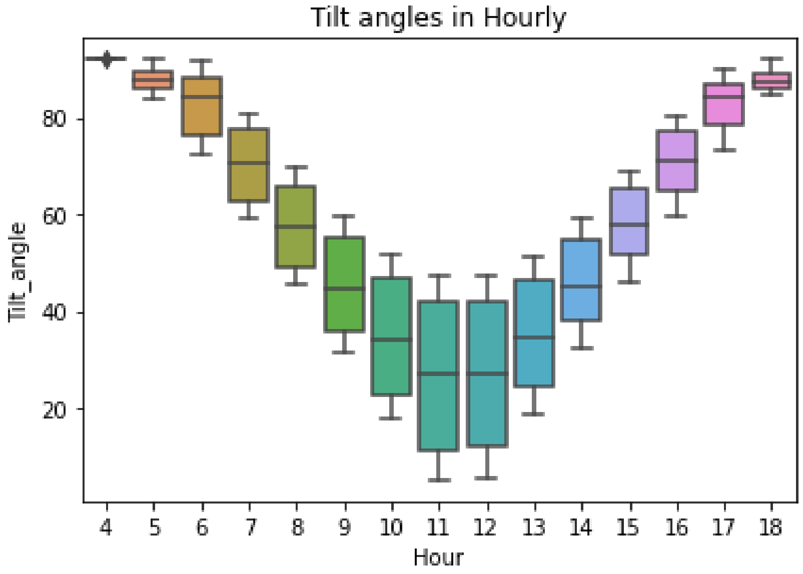

FIGURE 4.10: Box plots showing hourly tilt angles for three years.

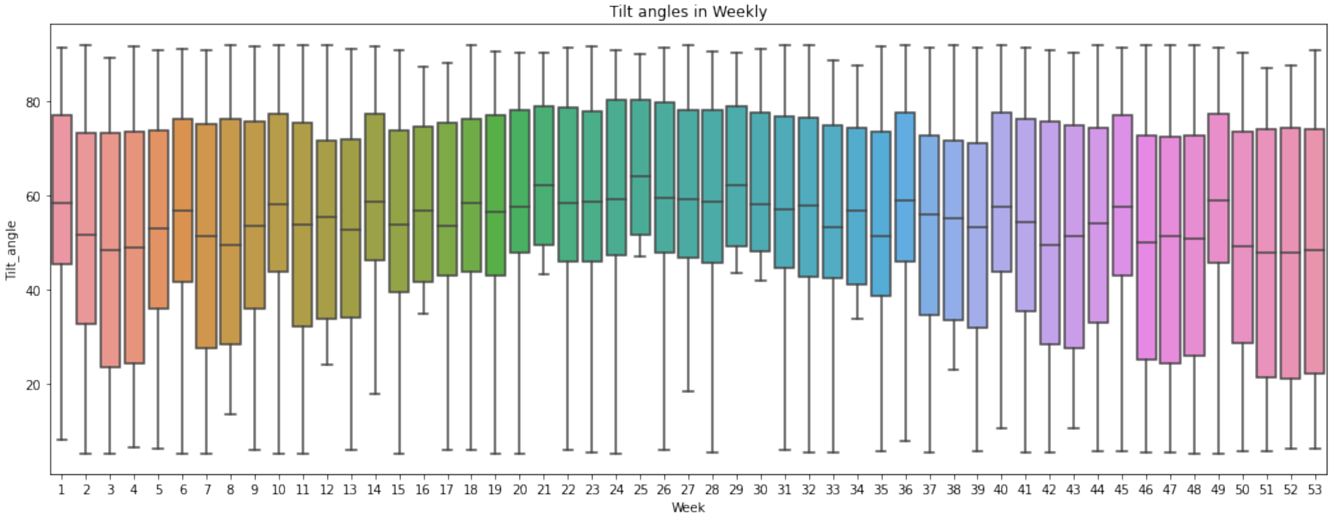

FIGURE 4.11: Box plots showing weekly tilt angles for three years.

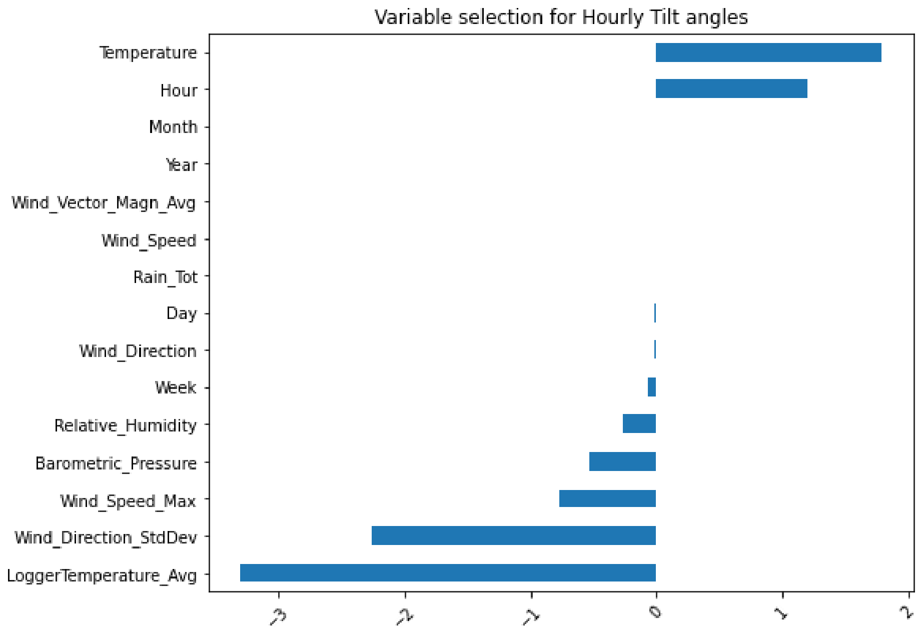

FIGURE 4.12: Variable selected for training models to forecast optimum tilt angles.

Figure 4.12 shows the results obtained after variable selections using the LASSO regressor. Variables such as rain total, wind speed, wind vector magnitude average, year, month were all removed since they are not useful in the forecasting tilt angle of acceptance. The models were trained on the dataset such as temperature, relative humidity, wind direction,wind direction standard dev, wind speed max, barometric pressure, logger temperature, week, day, hour as shown in Figure 4.12.

Figure 4.13 shows how the tilt angles change throughout the years. Since the dataset was being measured 24 hours automatically using a pyranometer. Irrelevant data recorded during night hours were then removed, in this case, the angles higher than 92 degrees as shown in Figure 4.13 were all removed and the max angle of acceptance became 92 degrees, which the pyranometer reaches at around 18H00 night. This was done with consideration to the fact that beyond that time there was no solar irradiance scattered.

The density distribution of the tilt angle indicate to be bimodal as shown in Figure 4.13. The figure also shows the changes of tilt angles throughout the time for three years recorded hourly.

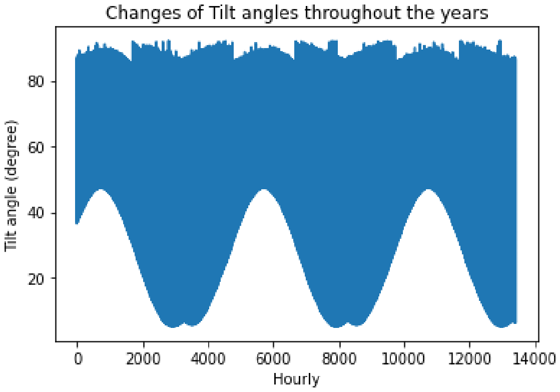

FIGURE 4.13: Tilt angles probability density and hourly changes of tilt angles for optimal GHI recorded for three years.

The optimal results for RF were obtained when trained with the number of trees set to 430, the dataset was split into 70% for training and 30% testing. Optimal results for LSTM were obtained after training the model for 231 epochs when the model was set to stop training after 40 iterations of epochs when there was no improvement. Table 4.4 below summarizes the results obtained for both models.

TABLE 4.4: Testing errors of the RF and LSTM models for predicting tilt angle of acceptance for solar energy.

Testing	RF	LSTM
MAE	1.72	1.23
MSE	8.23	2.37
RMSE	2.87	1.52

Error measures in Table 4.4 show the results obtained for the Random Forest (RF), and Long Short Term Memory (LSTM) evaluated on the testing dataset. LSTM

FIGURE 4.14: Density plots for RF and LSTM on the testing data.

proved to have better results than RF in all error measures calculated, with a mean absolute error of 1.23, more accurate than RF by 0.49 less, mean square error of 2.37 show to be more precise than MSE of the RF by 5.86 less, and the root mean square error of 1.52 which shows to be more precise by 1.37 less. Generally, the predictive results obtained for tilt angles and GHI vary with a large margin and this is due to the feature selection which lead the models to train on the different independent variables, and also the models were trained to forecast GHI on the minute averaged dataset whereas the tilt angle were trained on the very same data but converted to hourly averaged by aggregation. Figure 4.14 shows the density plots of the actual testing data and predicted data for both RF and LSTM,

Chapter 5

Conclusion

The study focused on training machine learning models such as Random Forest (RF), Bayesian Linear Regression (BLR) and compare them against the benchmark models, K- Nearest Neighbour (KNN) and Long Short-Term Memory (LSTM) in forecasting short-term global horizontal irradiance. Overall, the RF outperformed all the models including the benchmarks, KNN, and LSTM, the BLR had the worst performance. The optimal performances of the models were achieved by performing variable selection using the LASSO method and de-noising dataset using the lfilter method. The RF and the LSTM were then used to estimate hourly averaged tilt angles of acceptance for solar energy. The results obtained for RF and LSTM were different by a small margin. In conclusion, the random forest can be considered as one of the best models to be used in forecasting solar energy and angles of solar collectors for solar energy because it has surprisingly showed good performances in both forecasting GHI and tilt angles of acceptance. The LSTM has proved to be reliable by being consistent with the study [19] and being the best in forecasting tilt angle of acceptance and being the second best model in estimating GHI. Possible future work should focus on using more machine learning models to model and estimate the angles of acceptance for solar irradiance scattering improvements.

Bibliography

[1] "Eskom: SA on brink of load shedding". In: (2019). URL: https://www.news24.com/Fin24/eskom-sa-on-brink-of-further-load-shedding-20190821.

[2] Ke Wang Ke Wang Zhimin Huang Zhaohua Wang Yi Li. "Environment-adjusted operational performance evaluation of solar photovoltaic power plants: A three stage efficiency analysis". In: 104 (Sept. 2017). URL: https://www.researchgate.net/publication/316007906_Environment-adjusted_operational_performance_evaluation_of_solar_photovoltaic_power_plants_A_three_stage_efficiency_analysis.

[3] Shi Su Hongshan Zhao Fei Wang Zengqiang Mi. "Short-Term Solar Irradiance Forecasting Model Based on Artificial Neural Network Using Statistical Feature Parameters". In: (2012), p. 16. URL: https://www.researchgate.net/publication/272647724_Short-Term_Solar_Irradiance_Forecasting_Model_Based_on_Artificial_Neural_Network_Using_Statistical_Feature_Parameters.

[4] Alexander Jung Timo Huuhtanen. "Predictive maintenance of photovoltaic panels via deep learning". In: (2018), p. 05. URL: https://www.researchgate.net/publication/327129244_PREDICTIVE_MAINTENANCE_OF_PHOTOVOLTAIC_PANELS_VIA_DEEP_LEARNING.

[5] Gi Yong Kim, Doo Sol Han, and Zoonky Lee. "Solar panel tilt angle optimization using machine learning model: a case study of Daegu City, South Korea". In: 13.3 (2020), p. 529. URL: https://www.mdpi.com/1996-1073/13/3/529.

[6] Eric Martinot et al. "Renewables 2007-global status report". In: (2008). URL: https://inis.iaea.org/search/search.aspx?orig_q=RN:46105564.

[7] Mehrnoosh Torabi et al. “A hybrid machine learning approach for daily prediction of solar radiation”. In: (2018), pp. 266–274. URL: https://link.springer.com/chapter/10.1007/978-3-319-99834-3_35.

[8] Zhen Zhang Yongfen Jiang Zengwei Zhu Guoan Xu and Ming Liu. “Very Short-Term Solar Irradiance Forecasting at a Sub-Minute Scale Based on WT-Cnns”. In: (2020). URL: https://www.researchgate.net/publication/346480416_Very_Short-Term_Solar_Irradiance_Forecasting_at_a_Sub-Minute_Scale_Based_on_WT-Cnns/citations.

[9] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. “Online short-term solar power forecasting”. In: 83.10 (2009), pp. 1772–1783. URL: https://www.sciencedirect.com/science/article/pii/S0038092X09001364?casa_token=B0H9e1U9Zs4AAAAA:CuyacnfRcIgP14VTDT7PsVPu-uYowy-ROU9o0RhsbOOQssFGXvPye5t_vV8SXu4d0Eu6W8bLyms.

[10] Hugo TC Pedro and Carlos FM Coimbra. “Assessment of forecasting techniques for solar power production with no exogenous inputs”. In: 86.7 (2012), pp. 2017–2028. URL: https://www.sciencedirect.com/science/article/pii/S0038092X12001429?casa_token=gWVWWDgx-nUAAAAA:QLmtMRXFjyBnFDSW8-fvS3h7HKj4sfA2trM8uyDAPVCMxU6w791NQs_H_P0baoYTs-T3A7_vjiY.

[11] Tendani Mutavhatsindi, Caston Sigauke, and Rendani Mbuvha. “Forecasting Hourly Global Horizontal Solar Irradiance in South Africa Using Machine Learning Models”. In: 8 (2020), pp. 198872–198885. URL: https://ieeexplore.ieee.org/abstract/document/9244163/.

[12] du Clou S. van Niekerk J.L. Gauche P. Leonard C. Mouzouris M.J. Meyer A.J. van der Westhuizen N. van Dyk E.E. Brooks M.J. and F Vorster. “SAURAN: A new resource for solar radiometric data in Southern Africa”. In: (2015), p. 26. URL: https://sauran.ac.za/.

[13] José F Torres et al. “Big data solar power forecasting based on deep learning and multiple data sources”. In: *Expert Systems* 36.4 (2019), e12394. URL: https://doi.org/10.1111/exsy.12394.

[14] Matteo Saviozzi. Federico Silvestro S. MassuccoS. Gabriele Mosaico. “A Hybrid Technique for Day-Ahead PV Generation Forecasting Using Clear-Sky Models or Ensemble of Artificial Neural Networks According to a Decision Tree Approach”. In: (2019). URL: https://www.researchgate.net/publication/332219407_A_Hybrid_Technique_for_Day-Ahead_PV_Generation_Forecasting_Using_Clear-Sky_Models_or_Ensemble_of_Artificial_Neural_Networks_According_to_a_Decision_Tree_Approach.

[15] Florian Barbieri, Sumedha Rajakaruna, and Arindam Ghosh. “Very short-term photovoltaic power forecasting with cloud modeling: A review”. In: 75 (2017), pp. 242–263. URL: https://www.researchgate.net/publication/332219407_A_Hybrid_Technique_for_Day-Ahead_PV_Generation_Forecasting_Using_Clear-Sky_Models_or_Ensemble_of_Artificial_Neural_Networks_According_to_a_Decision_Tree_Approach.

[16] Shi Su. Hongshan Zhao Fei Wang. Zengqiang Mi. “Solar Irradiance Forecasting Using Deep Neural Networks”. In: (2017). URL: https://www.researchgate.net/publication/320364772_Solar_Irradiance_Forecasting_Using_Deep_Neural_Networks.

[17] Wolfgang Lehner. Robert Ulbricht. Ulrike Fischer. and Hilko Donker. “First steps towards a systematical optimized strategy for solar energy supply forecasting”. In: 2327 (2008). URL: https://www.semanticscholar.org/paper/First-Steps-Towards-a-Systematical-Optimized-for-Ulbricht-Fischer/8abf007afbf6b7e6b3501ef8adc5afadb76980d5.

[18] G.W.Chang H.J.Lu. “A Hybrid Approach for Day-Ahead Forecast of PV Power Generation”. In: (2018), pp. 634–638. URL: https://reader.elsevier.com/reader/sd/pii/S2405896318334955?token=A94532896D1F0047663524F7F9E37BA9DD0A58063
originRegion=eu-west-1&originCreation=20210726102517.

[19] Cyril Voyant et al. “Machine learning methods for solar radiation forecasting: A review”. In: *Renewable Energy* 105 (2017), pp. 569–582. URL: https://scholar.googleusercontent.com/scholar.bib?q=info:e2pVZmG2SnEJ:scholar.google.com/&output=citation&scisdr=CgVExwWmEJuW6QNJ0m4:

AAGBfm0AAAAAYP1Mym6103SE_gl27xR77Qd8TBi-FNpC&scisig=AAGBfm0AAAAAYP1MyikjOFE7WbH
scisf=4&ct=citation&cd=-1&hl=en.

[20] Mamphaga Ratshilengo, Caston Sigauke, and Alphonce Bere. "Short-Term Solar Power Forecasting Using Genetic Algorithms: An Application Using South African Data". In: *Applied Sciences* 11.9 (2021), p. 4214. URL: https://www.mdpi.com/2076-3417/11/9/4214.

[21] Basharat Jamil, Abid T Siddiqui, and Naiem Akhtar. "Estimation of solar radiation and optimum tilt angles for south-facing surfaces in Humid Subtropical Climatic Region of India". In: *Engineering Science and Technology, an International Journal* 19.4 (2016), pp. 1826–1835. URL: https://www.sciencedirect.com/science/article/pii/S2215098616306085.

[22] Manoj Kumar Panjwani. "Solar Concentrator's Effect on Solar Panel Efficiency". In: *Sukkur IBA Journal of Emerging Technologies* 1.1 (2018), pp. 15–27. URL: http://journal.iba-suk.edu.pk:8089/SIBAJournals/index.php/sjet/article/view/187.

[23] Swarnavo Datta, Supantho Bhattacharya, and Priyanka Roy. "Artificial Intelligence based Solar Panel Tilt Angle Optimization and its Hardware Implementation for Efficiency Enhancement". In: *International journal of advanced research in electrical, electronics and instrumentation engineering* 5 (2016), pp. 7830–7842. URL: https://www.semanticscholar.org/paper/Artificial-Intelligence-based-Solar-Panel-Tilt-and-Datta-Bhattacharya/237ea2b83d7a63674e21e270d041b19c

[24] Ketan Ramaneti, Pranavi Kakani, and Surya Prakash. "Improving Solar Panel Efficiency by Solar Tracking and Tilt Angle Optimization with Deep Learning". In: (2021), pp. 102–106. URL: https://ieeexplore.ieee.org/abstract/document/9490485?casa_token=Y5DqUtOJiQkAAAAA:3S5RLz-IQgWVIBcg1wLPdM1F_jcumwmHrYf64QNqg17oJth2ceaE-KgxeHFGCGQMNl5kISRr5KHa.

[25] Chih-Chiang Wei. "Predictions of surface solar radiation on tilted solar panels using machine learning models: A case study of Tainan city, Taiwan". In: *Energies* 10.10 (2017), p. 1660. URL: https://www.mdpi.com/1996-1073/10/10/1660.

[26] Tin Kam Ho. “Random decision forests”. In: 1 (1995), pp. 278–282. URL: https://doi.org/10.1109/ICDAR.1995.598929.

[27] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780. URL: https://ieeexplore.ieee.org/abstract/document/679596.

[28] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. “Probabilistic programming in Python using PyMC3”. In: *PeerJ Computer Science* 2 (2016), e55. URL: https://peerj.com/articles/cs-55.pdf.