BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

ARTICLE

Machine Learning Approach for Short- and Long-Term Global Solar Irradiance Prediction

Oliver O. Apeh * , ***Nnamdi I. Nwulu***

Centre for Cyber-Physical Food, Energy & Water Systems (C-C-P-F-E-W-S), University of Johannesburg, P.O. Box 524, Auckland Park, Johannesburg 2006, South Africa

ABSTRACT

Solar radiation data forecasting algorithms are important, especially in developing countries, as vast solar power plants cannot measure reliable and constant solar irradiance. The challenges of solar irradiance prediction may be resolved by machine learning using weather datasets. This study emphasises the daily and monthly global solar radiation data predictions of three locations, Pretoria, Bloemfontein, and Vuwani, at different provinces in South Africa with various solar radiation distributions. The study evaluated five different machine learning models. Forecasting models were established to evaluate global solar radiation, focusing on input data. The selected forecast models are centered on their ability to perform with time series data. These models use five years of data from meteorological parameters, such as global horizontal irradiance (GHI), relative humidity, wind speed and ambient temperature between 1 January 2018 and 31 December 2022. The datasets from these meteorological parameters are utilised for training and testing the employed algorithms, which are examined using five statistical metrics. Moreover, the inconsistency of the solar irradiance time series was equally assessed using the clearness index. The results from this study demonstrate that the R^2 value recording 0.866 datasets in Bloemfontein of random forest algorithm presents the highest performance during the training processes for all models studied, while the random tree in Vuwani showed the lowest performance of R^2 of 0.210 with other algorithms in testing processes. Additionally, the maximum solar radiation was found in December for both Pretoria and Bloemfontein, recorded as 5.347 and 5.844 kWh/m^2/day, respectively, while it was 4.692 kWh/m^2/day at Vuwani in January. Similarly, the average

*CORRESPONDING AUTHOR:

Oliver O. Apeh, Centre for Cyber-Physical Food, Energy & Water Systems (C-C-P-F-E-W-S), University of Johannesburg, P.O. Box 524, Auckland Park, Johannesburg 2006, South Africa; Email: olivera@uj.ac.za

ARTICLE INFO

Received: 15 August 2024 | Revised: 6 September 2024 | Accepted: 19 September 2024 | Published Online: 24 December 2024
DOI: https://doi.org/10.30564/jees.v7i1.7060

CITATION

Apeh, O.O., Nwulu, N.I., 2024. Machine Learning Approach for Short- and Long-Term Global Solar Irradiance Prediction. Journal of Environmental & Earth Sciences. 7(1): 321–342. DOI: https://doi.org/10.30564/jees.v7i1.7060

clearness index of 0.605, 0.657 and 0.533 are obtained at Pretoria, Bloemfontein, and Vuwani, respectively. Among the three sites under study, the solar radiation and clearness index are higher in Bloemfontein. Therefore, the proposed algorithms could be used conveniently for short- and long-term solar power plants in South Africa.

Keywords: Machine Learning; Solar Radiation; Short and Long-Term; Forecasting; Statistical Metrics

1. Introduction

The world is gradually transforming into renewable energy integrations to generate electricity and reconsider its energy network[1]. For this reason, the application and development of renewable energy have become essential to the discussion on energy security sustenance and creating a conservative and workable electricity network. However, solar energy applications, one of the sources of renewable energy, have become popular in current research fields not only in developing countries like South Africa but also in the global community due to its prominent features such as cleanliness, availability, and environmentally friendly that make it more beneficial compared to other renewable energy sources[2]. Therefore, exploring the solar photovoltaic (PV) system is essential to decrease the effect of greenhouse gas discharges in South Africa, generating the highest amount of greenhouse gas emissions in Africa[3]. However, solar PV energy generation relies on several parameters, including humidity, module temperature, wind pressure, ambient temperature and solar radiation. The quantity of power generation may be altered due to the natural disparities of these parameters in the climate[4]. The sudden variation in solar power produced interrupts power networks' consistency, stability, and forecasting. To encounter the fluctuations in prices, demand variations, and weather instability in solar radiation, the involvement in a productive forecasting model for solar energy has become critical, which is most beneficial for enhancing and regulating demand in supply[5].

Despite being the seventh largest coal provider and the fifth largest coal distributor, South Africa is still experiencing terrible energy crises. The current grid connection is heavily loaded and under stress, while efforts to meet everyday electricity demands and a substantial percentage of the people in the country lack the means to provide an electricity supply[6]. Among the several challenges facing the South African government, energy security and maintenance remain the top priority. The energy disaster started at the end of 2007, when an urgent need for help was declared, and the execution of the load-shedding agenda began in 2008 to prevent the nation from going into a total blackout[7]. Meanwhile, this strategy was the only solution that the national power station (ESKOM) could use to solve the existing electricity demand. Since several countries have embraced solar PV to tackle power shortages, South Africa has yet to incorporate the new technology; hence, the issue of load shedding schedule remains a recurring event in the country.

The evolution of South African's electricity sector has been affected by historical activities, socio-political changes, and efforts to address the country's growing energy requirements. **Figure 1** presents a full insight into this transformation, stressing significant indicators and progress that have played crucial roles in shaping the energy sector.

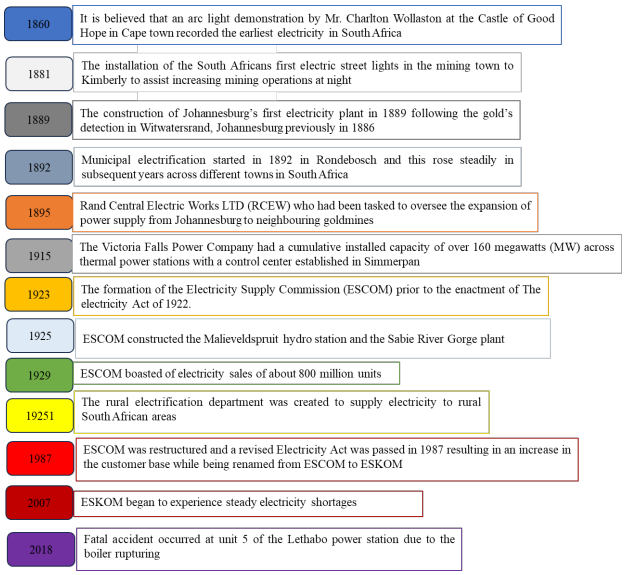

Figure 1. Many decades of electricity transformation in South Africa.

Even though the ESKOM requires more development, it is remarkable that ESKOM is one of Africa's most dependable and effective power providers, boasts a total installed capacity of approximately 10.5 GW and is better than Nigeria's power generation sub-sector of 16.384 MW, notwithstanding

that Nigeria has a larger population.

However, the government is currently initializing several efforts to improve the restrictions faced by people. These are not limited to the socioeconomic costs emanating from the interruption of the power source and failure of the state grid. As solutions to excite the nation through constant power supply are on the way, the national utility can encounter bigger demand, resulting in higher electricity charges for the entire nation. Upgrading these challenges in the energy sector is important for the financial development of a developing country, including South Africa. Therefore, reasonable prediction and proper solar radiation study might help decrease danger and allow resource management to be the most profitable method. Therefore, if solar radiation is predicted correctly, it is feasible to examine the financial description of the solar PV system at a specific position and significantly decrease prices by improving the installation capacity. Solar radiation assessment could be done through time horizon forecasting or centered on the input data type. Solar radiation prediction techniques focussing on data input are grouped into machine learning, statistical, hybrid, and persistence [8, 9].

Previously, mathematical methods were used to forecast electricity production from solar PV power plants. These approaches can be characterised as either the Statistical model or the persistence method. Regrettably, this method mostly generates low-precision forecasting and becomes inappropriate with non-linear data. Due to these restrictions, machine learning, including support vector machine, SVM, artificial neural network – ANN, metaheuristic, and extreme learning machine – ELM methods, have increased significantly [10–12]. Machine learning can conveniently solve problems that are difficult to handle using explicit processes. Machine learning-based algorithms can build a network between inputs and outputs even when illustrations are impossible, making it appropriate for design recognition, data mining, forecasting and classification [13].

Different research works are being published to forecast solar radiation, and solar prediction could be regarded as a time series challenge. Forecasting solar radiation with the Autoregressive Moving Average (ARMA) and the autoregressive integrated moving average (ARIMA) has been in existence since the 1970s [12, 14]. Examination of the numerical weather prediction algorithm (NWP) and ARIMA in short-time horizons shows growing prospects with greater accuracy [15, 16]. The statistical features and famous Box-Jenkins method demonstrated by the ARIMA techniques make it a preferred model in the system-building procedure. However, their key constraint is the pre-expected linear method of the algorithm [17]. Lately, ANN has been widely examined and applied to predict time series due to knowledge of the structures shown in the data, deducing the invisible aspect of nonlinearity and noisy data [18]. Moreover, statistical metrics, including MAE, R^2, RMSE, RAE, and RRSE, are currently used to show large-scale PV plants using forecasting methods. Different artificial intelligence (AI) algorithms, such as RF, DS, LR, DT, and MLR, are frequently used to predict solar radiation data. Numerous results from earlier investigations opine that AI models produce more reliable results than empirical algorithms in predicting solar radiation [19, 20].

However, quite a few studies based on long- and short-term solar radiation forecasts comprise data from more than one month. In addition to energy proposals and security procedures, it supports stakeholders in the design of electricity generation, transmission and supply [21]. Mellit et al. predicted monthly solar radiation using a library of Markov transition matrices and ANN [22]. The result generated a series of daily clearness indexes. In another development, Apeh et al. studied solar radiation based on monthly, seasonal, and yearly data under South African weather conditions [23]. The result illustrated an average percentage frequency of clearness index of 31.28% of clear sky days, 57% of partially cloudy days and 11.72% of cloudy days. Furthermore, on Abu Musa Island, situated in southern Iran, the assessments of hourly solar radiation were conducted using different models, including fuzzy inference system (FIS), SVR, ANFIS, Multilayer Feedforward Neural Network (MLFFNN), and Radial basis function networks (RBFNN). Their results showed that SVR outperformed ANFIS, MLFFN, FIS and MLFFNN with a correlation coefficient of 0.9999. Independent climatic parameters such as RH, local time and TEMP were used [24]. A genetic model focused on the SVM algorithm was predicted for the short-term forecast of the PV power plant. The suggested algorithm produced better predictions than the standard SVM algorithm built on RMSE and MAPE metrics [25].

A novel prediction technique for global solar radiation forecasting, established on SVM, has been suggested by Meenal and Selvakumar [26]. Similarly, global solar ir-

radiance has been developed using an SVM by Jiang and Dong[27], while Fan et al. modeled a long-term solar power forecasting approach that integrates a hybrid method with least square support vector regression[28]. Das suggested a forecasting approach for solar power generation that uses atmospheric data and supports vector regression with historical solar power[29]. For probabilistic solar power forecasting, several linear regression techniques have proved to have a strong performance[30]. A comparative assessment between support vector regression and multiple linear regression has been presented for short-term solar power forecasting.

1.1. Problem Formulation

The fundamental challenge motivating in writing this paper is the demand to develop an accurate solar irradiance forecast, precisely in areas with rich solar resources, including South Africa—an outstanding example in Africa. Notwithstanding the substantial solar energy prospect in this area, efficiency application encounters large challenges due to the difficulties required in accurate solar irradiance prediction. This research emphasizes the demand for an innovative forecasting model trained in acquiring and integrating solar irradiance fluctuations within the complex climatic changes and environmental conditions exclusive to each province. The present forecasting models, though promising, fall short of presenting the accuracy necessary for efficient energy planning, continuous grid integration, and ideal system performance. This research gap highlights the need for a modern and appropriate forecasting model for South Africa's distinctive climate. This paper compares the application of various ML models to tackle the large challenge of advancing the dependability and accuracy of solar irradiance forecasting in this demanding environment. The essential challenge of this research lies in developing a forecasting solution that capably steers the intricate area of solar irradiance variations while tackling the location-detailed difficulties posed by the country's characteristic meteorological and environmental conditions.

1.2. Contributions of the Study

The main contributions of the present research work can be summarised as follows.

1) The article combines different types of data, such as geographic information and historical weather data, to enhance the correctness of solar irradiance predictions.
2) The research compares the application of various ML models, including MLR, LR, RF, DS, and RT, to determine the most efficient techniques for solar irradiance forecasting.
3) By using advanced ML procedures, the article establishes significant decreases in forecast errors compared to traditional approaches, improving the consistency of solar energy prediction.
4) The results of the research have practical implications for the solar energy industry, for example, better planning and management of solar power plants, enhanced integration of solar energy into the grid, and improved decision-making for solar energy decentralization and storage.

1.3. Research Gap

The study gaps that exist in the literature and prospective future research directions are highlighted as follows:

1) The study could advance to different geographical regions with several climatic conditions and assess the effect of numerous temporal resolutions on model accuracy.
2) The utilisation of transfer learning, where models trained in one location or with one dataset are adapted to another, is a possible area for further study.
3) The study found a gap between machine learning models and traditional physical models, including numerical weather prediction models for solar irradiance prediction.

1.4. Papper Organisation

This study is organised in the following ways: Section 2 describes the forecasting horizons such as very short-term, short-term, medium-term and long-term forecasting. Section 3 compares and evaluates prediction models. Moreover, Section 4 presents the method used, including the study site, data preprocessing, PV system, introduction to machine learning, statistical description and metrics. Results relating to machine learning are presented in Section 5. This section illustrates the results regarding solar radiation in short and long-term forecasting, as well as the comparison between machine learning and traditional forecasting approaches. The

discussions of the results are described in section 6. Finally, in Section 7, the conclusions, limitations and future works are presented.

2. Forecasting Horizons

The forecast horizon is the future period by which prediction is expected to be performed. Similarly, it can also be defined as the time duration between the real period and the period in which the forecasting is performed. It is classified into three groups: long-term, medium-term, and short-term. However, Kumari and Toshniwal have presented another type of forecasting horizon called "very short-term forecasting"[31]. However, no category of forecasting horizons has been universally announced. So far, it is essential to recognise the imminent need for power generation and utilisation by an electrical energy supplier. Solar irradiance prediction, in terms of forecasting horizons, can be applied to various uses for the effective and resourceful performance of photovoltaic power systems. Generally, **Figure 2** shows the various forecasting horizons of solar radiation data in terms of their operations' needs.

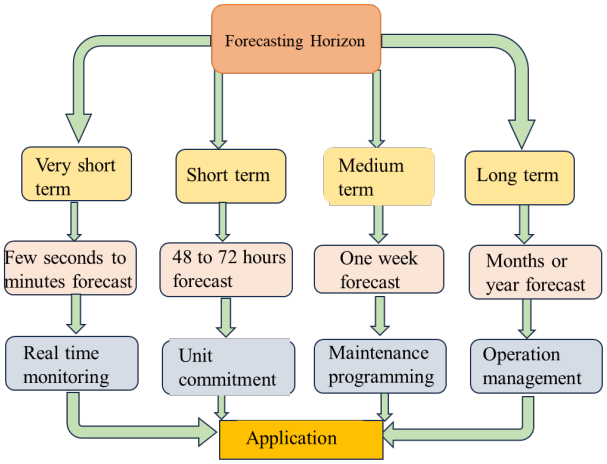

Figure 2. Forecasting horizon and their applications.

2.1. Very Short-Term Forecasting

The forecasting horizon of this type is characterized by a predicted time duration between 5 minutes and 6 hours [14]. An example of very short-term forecasting was observed by applying time series irradiance when Yang et al. studied solar irradiance prediction data logged every second[32]. Some scholars have measured a time scale of a few seconds to a few minutes or even up to a few hours under this group [33]. The forecasting category in this group is highly applicable to estimating electricity prices, requests, instantaneous notification, monitoring of PV plants and peak load matching [34].

2.2. Short-Term Forecasting

The short-term forecasting is very pertinent in the commercial supervision of electricity. It helps in electrical energy demand and supply, distributing load report opinions, unit efficiency, bulk energy storage and business development in the electricity market[35]. Typically, the short-term horizon measures from 30 minutes to 72 hours [36]. Nevertheless, few researchers deliberated the range of short-term forecasting from 1 h to numerous hours, days, or even up to 7 days. An example is observed when Jiang et al. studied a solar irradiance prediction model five days in advance, estimating to maintain consistency and effective harmonizing between demand and supply when joining the power system to the whole solar power[37].

2.3. Medium-Term Forecasting

The historical data measured in this forecasting class differs from a few days, weeks, and months in advance[38]. This type of classification is essential to build and maintain a timetable of solar power systems consisting of transformers and other equipment, including a method that experiences the least loss [39].

2.4. Long-Term Forecasting

Firstly, scientists see long-term forecasting as a few months to years ahead [40]. This prediction classification is appropriate for proposing long-term projects to implement solar power plants effectively. The long-term prediction system assists in international supervision, for example, site selection to institute a solar PV system, processing, distribution, and supply of solar energy. However, the prediction of long-term horizons is less accurate as it does not forecast meteorological variations for a prolonged period. So far, many scholars have studied long-term prediction algorithms to develop strategies and evaluate site selection [23, 41]. Thus, various prediction algorithms are often applied in solar radiation forecasting. Many of their results in terms of performance are near each other, with the research locations

having similar weather differences. Moreover, the variations of the results found, and the precision of the forecast differ in terms of the input parameters used and the capacity of data applied.

3. Comparison and Evaluation of Prediction Models

Machine learning and DL prediction methods are broadly applied in energy systems. These methods are used in several fields such as electrical load forecasting, building energy consumption forecasting as well as power and load demand forecasting[42–44]. Generally, buildings account for a substantial share of world energy waste and consumption. Hence, decreasing energy usage in buildings is a crucial technique to moderate climate change effects[45]. Consequently, most research emphasizes predicting energy demand and consumption in buildings. With respect to the forecast time horizon, building energy forecasting can be classified into short-term (up to one week ahead), medium-term (from one week to one year ahead), and long-term forecasts (more than one year ahead)[46]. Forecasting energy demand in buildings is essential at several levels, from specific households to the national level. The optimization of equipment performance contributes to balancing demand and supply through real-time renewable energy sources, including in virtually zero-energy buildings, and supports installation planning and price decreases in energy systems[47]. Exact information regarding residents' electricity usage is important to expand load prediction accuracy and guarantee the dependable operation of energy management, power systems, and planning[48].

Several recent research emphasises progress in this area. Amasyali et al. worked on building energy consumption forecasts using diverse ML and DL models. The findings showed that while ML algorithms mostly work well, each has weaknesses and strengths, demanding model selection based on definite uses. Gaps were acknowledged in long-term prediction, residential building energy consumption, and lighting energy consumption prediction, demanding more consideration[49]. Deb et al. studied nine-time series (TS) prediction methods for building energy consumption, such as NN, Fuzzy, Hybrid Model (HM), ARIMA, Case-Based Reasoning (CBR), Support Vector Machine (SVM), Gray, Moving Average (MA), Exponential Smoothing (ES), and ANN. They noted that joining TS methods, for example ANN and ARIMA, with optimization approaches like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) produces better results[50]. Walker et al. conducted research on ML algorithms, such as ANN, Boosted-Tree (BT), SVM and RF, to forecast hourly electricity demand using data from 47 commercial buildings for a period of two years. The RF model proved the best performance in terms of accuracy and prediction error[51].

In another development, Grimaldo et al. integrated the k-Nearest Neighbor (kNN) with visual analytics to forecast and assess energy demand and supply. This method produced correct results and allowed the assessment of diverse prediction possibilities and consumption patterns[52]. Similarly, various models were compared to analyse global solar irradiation in the Sudanese zone of Chad[53]. Hagh et al. introduced a hybrid model (HM) integrating SVM with quicker clustering k-medoids and ANN to forecast home application peak demand and power consumption. This algorithm accomplished a required accuracy of 99.2% using smart meter data[54]. Hafeez et al. presented an advanced HM for short-term electrical load forecast, combining a DL model called Factored Conditional Restricted Boltzmann Machine (FCRBM), Modified Mutual Information (MMI), and Genetic Wind-Driven Optimization (GWDO). This algorithm performed better than others, including LSTM, AFC-based ANN, and MI-based ANN, in terms of accuracy, average runtime, and convergence rate[55]. Khan et al. developed the Cuckoo Search Neural Network (CSNN) by integrating ANN and Cuckoo Search (CS) to advance convergence time, accuracy, and compatibility for the Organization of Petroleum Exporting Countries (OPEC) power consumption prediction. This algorithm demonstrated compatibility and superior efficiency compared to algorithms including Genetic Algorithm Neural Network (GANN), Accelerated Particle Swarm Optimization Neural Network (APSONN), and Artificial Bee Colony Neural Network (ABCNN)[56]. Kazemzadeh et al. conducted an HM for long-term forecast of total energy demand and peak electrical load using SVR, ANN and ARIMA. Their results showed that the HM performed better than the other models analysed (HM > PSO-SVR > ANN > ARIMA)[57].

Several interesting statistical metrics have been applied

to predict solar radiation, but few have shown high performance. Certainly, many of the studies undertaken to predict global solar radiation used models emanating from groups and classes. Thus, the entire statistical evaluation mostly presented results that were similar to those of all other categories of models[58]. Frequently, it becomes very difficult to evaluate algorithms successfully because of the limited number of metrics. For example, the research work by[59] in 2011 evaluated daily global solar radiation data. The authors established the R^2 values as their highest performance in the study. **Table 1** compares the statistical metrics used in the literature to describe and select the optimal forecasting algorithm.

The analysis of several literature studies from **Table 1** indicates that no model presents the best results in all locations. Even though many predictions are compared with similar data types, it is commonly observed that the models present optimum results from location to location. Thus, there can be a difference among the metrics that have the same models and bring the best results for several localities. For instance, Mehdizadeh et al. obtained an optimum result in forecasting daily global solar radiation with the ANN model and estimated an appropriate RMSE value as 1.850[66]. In a different study, Antonopoulos et al. obtained an optimum result using ANN with the highest RMSE value of 3.166[67]. The differences in the results encountered by researchers may be a result of missing data, dataset size, the local climate, input variables, geographical differences and feature selection. Therefore, different research places emphasis on definite time frames or regions. However, this study progressed to several geographical locations in South Africa with various climatic conditions and analysed the effect of various temporal resolutions on model accuracy. Besides, the study found a gap between machine learning models and traditional physical models including numerical weather prediction models for solar irradiance prediction.

Briefly, this article brings a new look at the existing literature in the following areas:

I) Integrating MLR, LR, RF, DS, and RT models shows that DL models can improve forecasting accuracy for solar irradiance.
II) The inclusion of predictors resulting from clear sky index time series enhances prediction reliability.
III) Analysis of the models using a wide-ranging set of metrics, such as MAE, R^2, RMSE, RAE, and RRSE, and forecast skill, providing a robust assessment framework.

4. Methodology

4.1. Description of the Study Site and Database

South Africa is divided into nine different administrative provinces bounded by latitudes 25° and 30° south and longitudes 17° and 32° east with huge and clear skies that are endowed for solar energy application, and this energy has a variety of prospect in every province[68]. The analysis depicts that every year, the hourly solar radiation in South Africa exceeds 2500 h and an average between 4.5 and 6.5 kWh/m^2/day. The availability of solar radiation data in the country is pretty wide compared to other African countries and 2.5 times higher than in Europe[69]. **Figure 3** presents the solar radiation distribution in South Africa, representing a considerable amount of solar radiation in the country, making it an excellent contender for establishing a solar power plant.

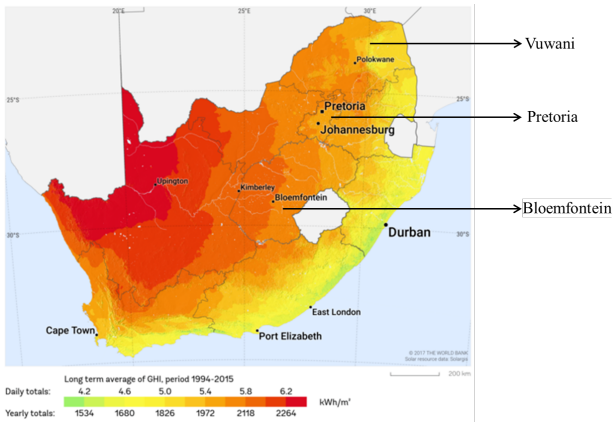

Figure 3. The annual geographical solar radiation distribution in South Africa.

The GHI map in **Figure 3** illustrates the amount of solar radiation obtained per unit area in various regions in South Africa. The usefulness of this information cannot be overemphasised mostly for designers and stakeholders of solar energy technologies.

4.1.1. Data Pre-Processing

The experiments executed in this study used Python 3.6 through third-party libraries, including Pandas and NumPy library (Sklearn). Five ML algorithms were selected to build

Table 1. Comparison of different studies in the literature on solar radiation with the present research.

Prediction Model	Optimal Model	Evaluation Metrics					Forecasting Horizon/Location	References
		RRSE	RMSE	R^2	MAE	RAE		
ANFIS, Empirical MSTree	ANFIS	Absent	0.573	0.910	Absent	Absent	Daily/China	[60]
DT, LM, DL, GB, SVM, RF	SVM	Absent	0.708	0.800	Absent	Absent	Daily/Europe	[58]
SVM, Empirical	SVM	Absent	0.495	Absent	Absent	Absent	Daily/China	[61]
ENN, LRNN, LLNN, FFNN	FFNN	Absent	0.026	Absent	Absent	Absent	Monthly/Nigeria	[62]
MLP, LSTM, GRU, CNN	CNN	Absent	0.129	0.967	Absent	Absent	Monthly/Iran	[5]
ANN, SVR, GRNN, RF	ANN	Absent	0.226	0.998	Absent	Absent	Seasonal/South Africa	[63]
LSTM, RNN, SVR	LSTM	Absent	0.032	Absent	Absent	Absent	10 years/South Africa	[64]
GSR	NAR-ANN	Absent	0.330	0.960	Absent	Absent	2 years/Nigeria	[65]

the models. The preliminary variable locations of each algorithm were chosen according to the algorithm's characteristics. The selection varieties of the variables were then set in terms of the parameter adjustment approaches for different ML algorithms. The study utilised Sklearn's GridSearchCV technique to choose parameters for each of the 5 ML algorithms, eventually saving the optimal model. However, the data sets retrieved from the solar stations were established per hour and daily in a horizontal plane with the sum of 9440 samples per day. Meteorological data were first examined to determine the presence of missing values before undergoing training processes. To take care of the missing values of the functions, the established scientific libraries in Python were used for data interpolation. Also, it was discovered that exploratory data analysis from the per-hourly values of GHI showed that a number of data equal to zero were represented early in the morning (from 5 am in summer and 7 am in winter) and at night (from 6 pm in winter and 7:30 pm in summer) are bound to be zero. However, this affects the fitting of our models on the data as 23790 GHI values are zero out of 47670, which is the total number of records. These zero values were removed since the model is affected when the values are fitted. After selecting data for a fixed time interval, the total records become 27568. The GHI values have been normalized to lie in [0, 1] using Equation (1).

$$\hat{GHI}_t = \frac{GHI_t - GHI_{\min}}{GHI_{\max} - GHI_{\min}} \quad (1)$$

Where GHI_t represents GHI at time-step t, $GHI_{\min}$, stands for the minimum value of the population, $GHI_{\max}$ is the maximum value of the population, and $\hat{GHI}_t$ is the normalized value of GHI at time-step t.

However, to test the model performances (RF, DS, LR, DT, and MLR model) at various climate stations, daily measurements of meteorological parameters during 2018–2022 at three locations across South Africa were selected. The monthly variations of ambient temperature (Ta), relative humidity (Rh) and wind speed (Ws) at each location are presented in **Figure 4**. Obviously, most of the meteorological parameters (Ta and Ws) are higher in the summer months and lower in the winter months. For instance, the monthly mean of Ta at Pretoria is 22.25 and 13.01 °C in December and July, respectively, with an annual value of 19.21 °C. Also, the highest value of the monthly mean of Ta at Bloemfontein is 23.78 °C in February, while this value decreases to 9.64 °C in July, recording an annual value of 17.91 °C. At Vuwani, the highest and lowest values of Ta are recorded as 26.49 and 16.65 °C in December and July, respectively, with an annual average of 22.24 °C. Similarly, there are obvious wind speed differences recorded at each station. At Pretoria, the annual Ws value is 2.04 m/s, with the highest and lowest values recorded as 2.60 and 1.24 m/s in December and June, respectively. A similar phenomenon of lower Ws in the winter and higher in summer has also been observed at Bloemfontein and Vuwani, which have annual values of 2.08 and 2.02 m/s, respectively.

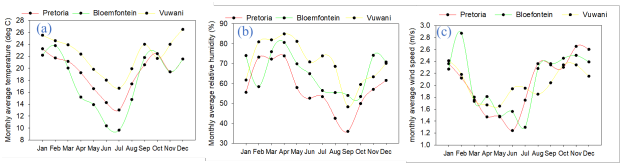

Figure 4. Monthly variations of average metrological parameters for each station. (**a**) average temperature. (**b**) average relative humidity. (**c**) average wind speed.

Meanwhile, the annual average Rh at Pretoria, Bloemfontein and Vuwani are 57.17, 65.68 and 70.44% respectively. The values of Rh are generally low in September within the

three locations studied. It is found to be lowest in Pretoria with values of 35.99%, while the maximum values are found in Vuwani and recorded as 84.91% throughout the year.

4.1.2. Distribution of Dataset

In this research, five years of historical meteorological and solar radiation datasets were utilised to investigate the efficiency of the algorithms at the study sites. The set of data retrieved was Rh, Ta, Ws and GHI as the expected output. The data was retrieved from 1 January 2018 to 31 December 2022 from the Southern African Universities Radiometric Network (SAURAN). The details of the research locations are shown in **Table 2**.

Data quality control is crucial considering the study time and the characteristic errors in the instrument-based observations. The missing and abnormal values in the weather data were excluded from the final dataset, and the solar radiation data quality control provisions proposed by [70] were utilised. On the other hand, data pre-processing involves three stages, filtering, scaling and partitioning the data. Data filtering takes care of the existence of incorrect measurements or outliers, which can lead to errors or cause uncertainties in the established models. Similarly, scaling is required so that a standardized range of disparities is obtained in the features being employed. Data partitioning is executed in two stages. The initial step is data splitting into training and testing subsets. Thereafter, the data training set is further split into other training and validation subsets to utilise the training subset and fit into the parameters of the model and use the validation subset to optimize the hyperparameters as well as finetuning the models with balanced data. The original dataset comprises the 4 input variables, including global horizontal irradiance (GHI) an output (PV generation) for the five years of data collection, as presented in **Figure 5**.

However, the data are trained, and the best input selection is the main objective of training models in various prediction horizons. The training set contains 21000 datasets while the testing data set contains 5330 datasets. Several arrangements of weather data sets have been tested to categorise the training procedure. The algorithms centered on ML propose metrological performance with datasets of a ratio of 70% for training containing 20870 records, validation set containing 20% of the total data recording 4897 while 10% represents test set with a total data of 2464 records.

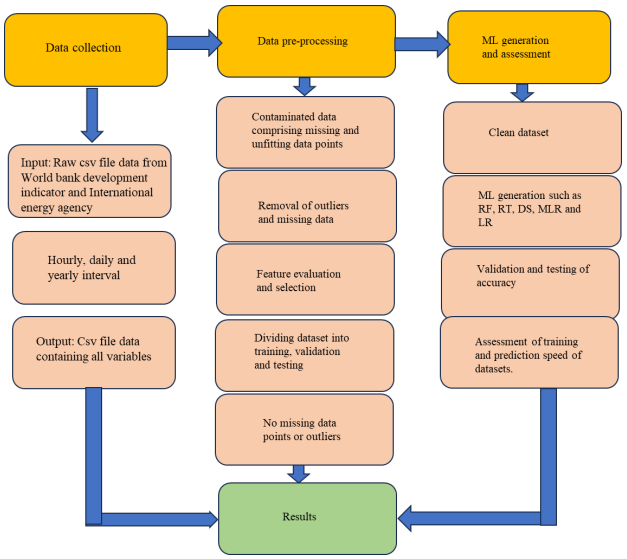

Figure 5. The research methodology from data collection to forecasting power from PV system.

4.2. Machine Learning Prediction Techniques

Machine learning (ML) has become an essential technique in advancing the accuracy of solar radiation predictions, making it crucial for both renewable energy integration and other fields depending on solar energy. This system is often utilised by artificial intelligence (AI) and remains famous for gradually discovering different application areas and efficiency [71]. ML offers the systems the capacity to comprehend the issue and proceed to evaluate the unknown outputs. Certainly, the output of any ML is highly subjected to its variety of features and training success. In this research work, five various ML algorithms were applied. They are Decision Stump (DS), Random Forest (RF), Random Tree (RT), Multivariable Linear Regression (MLR) and Linear Regression (LR). As stated earlier, the dataset was arbitrarily split as the shuffled selection method, and 70% of the entire data was utilised in the training stage of the model, 20% was used for validation, and the remaining 10% was used in the testing stage. Generally, a similar dataset for each location was utilised for the training, validation and testing data, and the similar symbols were predicted to generate a higher assessment among the ML algorithms.

4.2.1. Decision Stump

A decision stump is an ML technique that comprises a one-level decision tree. In this case, the stated decision tree has one internal node, the root directly linked to the

Table 2. Geographical descriptions of the research locations.

Site	Province	Longitude	Latitude	Elevation (m)	Topography
Pretoria	Gauteng	28.22	−25.75	1410	The roof of the university building
Bloemfontein	Free state	26.21	−29.12	1397	The roof of the Engineering building
Vuwani	Limpopo	30.42	−23.13	6280	Vuwani Science Research Center

terminal nodes (its leaves). The prediction of a decision stump focuses on the value of a single input variable and is sometimes referred to as a 1 rule by many researchers[72]. It is fast to train and make predictions, which can be useful when computational resources are inadequate.

4.2.2. Random Forest

Random forest (RF) is a group of classifiers that conceptualise an ensemble of non-identical and independent decision trees with the knowledge focusing on randomization. Equation (2) is used for the description of the random forest:

$$\left\{ \; h\left(x, \theta_k\right), k = 1, ..., L \; \right\} \tag{2}$$

Where θ_k represents a variable with a mutually independent random vector with input data x[73]. Every member of the decision tree employs a random vector as a variable that randomly chooses both the feature samples and the subsequent subset sample data set for training purposes. When building the random forest model, k represents the number of decision trees in the random forest, and n is the number of corresponding decision trees that each sample uses to train the dataset. Therefore, every member tree is trained on various subsections of samples (because of bagging), along with several subgroups of features (because of random feature selection). The selection of random characteristics in all member trees allows the dissociation of the forecasts of the various trees[74]. One criterion for selecting RF is to handle outliers and noisy data well due to the averaging effect of multiple trees.

The hyperparameter selections are essential for optimizing the performance of an RF model to improve prediction accuracy. Compared to advanced DL models, RF has moderately limited hyperparameters, making the process of selecting the best settings easier[75]. The main hyperparameters for an RF regressor model comprise: 1) n-estimators: This regulates the number of trees in the forest. According to Díaz-Uriarte and Alvarez de Andrés, this number should be set adequately high to obtain strong performance[76]. 2) min samples leaf: This stipulates the minimum number of samples needed to exist in a leaf node. 3) max depth: This sets the maximum depth of each tree, making this value too high can bring about overfitting the model during data training[77]. Nevertheless, Gressling opines that finding an ideal set of hyperparameters during the validation process relies deeply on the dataset's features[78].

4.2.3. Random Tree

RT splits a data set into batches and applies a constant to every group. A single-tree algorithm is prone to inconsistency and presents poor forecast precision. Therefore, RT bagging as a decision tree model can produce correct results[79]. RT has better adaptability and rapid training competence. When dealing with smaller datasets or situations where overfitting is a concern, RT might be a good choice.

4.2.4. Multivariable Linear Regression

When several explanatory variables are applied, the regression model is considered a Multilinear regression (MLR) model. As a result, more slopes need to be calculated, and the model should be analysed using a cross-validation method to reduce the risk of overfitting[80]. MLR model is a machine learning technique established focusing on the general formula depicted in Equation (3).

$$Y_i = b_1 + b_2 A_1 + b_3 A_2 + b_4 A_3 \ldots, + b_{n+1} A_n \tag{3}$$

While A_i represents meteorological parameters, b_1, b_2, b_3, b_4, and b_{n+1} represent regression coefficients, and Y_i represents the dependent variable.

MLR is mainly an algorithm that studies the relationship between dependent and multiple independent parameters. Additionally, it is widely used in assessing solar radiation research[33]. Training and making predictions with MLR are computationally efficient, making it appropriate for huge datasets.

4.2.5. Linear Regression (Time Series) Algorithms

Linear regression (LR) forecasting refers to an algorithm that forecasts the future results of any system with

historical information. Meanwhile, this prediction method requires the description of the data either by a non-linear or linear autoregressive process acquired. Therefore, the time series algorithm equation is presented in Equation (4).

$$y(m+n) = f\,[y(\,m), y(m-1), \\ y(m-2), \ldots, y(m-h-1)] \tag{4}$$

Here, the function *f* represents the current and historical values of *y*. However, different possible prediction methods involve forecasting the next values of the time series using two techniques. The first method is the independent value prediction (preparing the direct model to predict *y*(*m*+*n*). At the same time, the other step involves an iterative approach and reiterating one-method-upfront forecasting until the preferred possibility. To ensure the consistency of predicted future values in regression models, various assumptions must be considered. The most essential assumption is that the correlation between the residuals and the explanatory variables must be zero.

4.3. Statistical Description and Analysis

The standard deviation (SD) calculates the variation or dispersion of a set of values. The mean describes the average values of the parameters measured, while the minimum and maximum measure the range of parameters. **Table 3** presents the pertinent descriptive statistical factors used in the study.

This paper comprises four independent variables, including relative humidity and ambient temperature, along with the dependent variable being GHI. Solar irradiance measurements have daily and monthly variations demonstrating the hourly global solar radiation values and extraterrestrial solar radiation at noon. On the other hand, the clearness index, K_t, is usually utilised to characterize and estimate solar irradiance, as this index tolerates the monthly, seasonal, and yearly changes experimented in solar irradiance. However, K_t is described as the ratio of the solar global horizontal irradiance, *H*, obtained several times, to the extraterrestrial global horizontal irradiance for the same time and is expressed according to Equation (5):

$$K_t = \frac{H}{H_o} \tag{5}$$

Where $H_{o,t}$ represents extraterrestrial global horizontal irradiance at a time, t, and H is the global horizontal irradiance obtained simultaneously, t. On considering the hourly period, hourly extraterrestrial global radiation on a horizontal surface H_o is found by applying the well-recognised Equation (6).

$$H_{o,h} = I_{sc}E_o(\sin\delta + \sin\phi + \cos\delta\cos\phi\cos\omega_h) \tag{6}$$

Similarly, considering the monthly average daily extraterrestrial radiation, Equation (7) is considered[23]:

$$H_o = \frac{24x3600I_{sc}}{\Pi}{}_{sc}\left[1 + 0.333\cos\left(\frac{360n}{365}\right)\right] \\ \left[\cos\phi\cos\delta\sin\omega + \frac{\Pi\omega}{180}\sin\phi\sin\delta\right] \tag{7}$$

Where n = day number, δ = the declination angle calculated using Equation (8), I_{sc} = solar constant (1367 W/m^2), ϕ = latitude of the location, ω = hour angle and is calculated using Equation (9).

$$\delta = 23.45\sin\left(\frac{360}{365}(248+n)\right) \tag{8}$$

$$\omega = \cos^{-1}(\ -\tan\phi\tan\delta\ ) \tag{9}$$

However, the K_t model is not specifically limited to *H*, as presented in Equation (5). Different studies have focused on correlating the diffuse transmittance index or diffuse coefficient with the GHI parameter to improve the precision of DHI evaluation as presented in Equation (10)[23, 81].

$$\left(K_d = \frac{H_d}{H_o}\right) \approx f\left(K_t = \frac{H}{H_o}\right) \tag{10}$$

4.4. Statistical Performance Assessment

The daily and monthly global solar irradiation forecast, analysed with different machine learning algorithms, is compared with the data measured by statistical factors. For example, root relative square error (RRSE), mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2) and root absolute error (RAE).

The RMSE indicates the variation between the predicted and measured values presented by a model. Indeed, the RMSE represents model correctness by equating the difference between the actual and forecast data. The RMSE is calculated using Equation (11) and always has a positive value.

$$RMSE = \sqrt{\sum_{i=1}^{n}(O_i - P_i)^2} \tag{11}$$

In Equation (11), O_i represents the i-th observed or calculated value by the engaged methods, P_i represents the i-th predicted value, and the number of all observations is represented by *n*. As the RMSE becomes smaller, the model deviation is reduced and vice versa.

Table 3. Statistical description analysis of the used data from three locations in South Africa.

Pretoria					
Variables	**Mean**	**Std Dev**	**Std. Error**	**Max**	**Min**
GHI	391.0400	98.1800	28.3420	511.6210	249.7690
T_av	19.2080	3.2670	0.9430	23.2500	13.0100
Rh_av	57.1650	11.8190	3.4120	73.8600	35.9900
WS_av	2.0380	0.4750	0.1370	2.6500	1.2400
Bloemfontein					
GHI	248.6660	70.7480	20.4230	350.6110	162.9930
T_av	17.9130	4.9250	1.4220	23.7800	9.6400
Rh_av	65.6770	9.730	2.8090	80.6000	53.4600
WS_av	2.0830	0.4890	0.1410	2.8700	1.2960
Vuwani					
GHI	210.6380	45.6100	13.1660	281.5400	145.9640
T_av	22.2420	3.0780	0.8880	26.4900	16.6500
Rh_av	70.4350	10.9370	3.1570	84.9100	48.2900
WS_av	2.0230	0.2650	0.0765	2.4100	1.6500

In another development, the MAE describes an average of all errors given in Equation (12).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \tag{12}$$

It can also be seen as a measure of the errors between the measured and true values. The coefficient of determination R^2 describes the strength of a linear correlation between the forecast and the measured values, as defined by Equation (13).

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{n=1}^{n}(O_i - O_{ave})^2}\right] \tag{13}$$

Usually, the closer the R^2 value is to 1, the better the model's fitness. In addition, R^2 is an instrument that discovers and examines the ability of a statistical model to describe and predict future results.

Similarly, the RAE is defined as a quantity that compares a predictive model's performance with a simple model's performance. This statistical model confirms whether a model works better than predicting only the average, which is given by Equation (14).

$$RAE = \frac{\sum_{i=1}^{n} \mid O_i - P_i \mid}{\sum_{i}^{n} \mid O_i - O_{ave} \mid} \tag{14}$$

The RRSE describes the square root of the sum of squared errors normalized by the sum of squared errors of a simple algorithm as represented by Equation (15).

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O_{ave})^2}} \tag{15}$$

Where O_{ave} is the average of the observed values.

5. Results and Discussion

5.1. Prediction with Machine Learning

The analysis in this paper presents the predictability of daily global solar radiation falling on the horizontal surface through various machine learning models in three provinces in South Africa. Several statistical models that are regularly used in the literature were used to authenticate the success of the algorithms. **Table 4** presents the training and testing scores in the different provinces under study. The dataset of the RF algorithm in Bloemfontein presents the highest performance of R^2 with a value of 0.8659 in training processes for all models studied, while Random Tree in Vuwani showed the lowest performance of R^2 of 0.2100 with other algorithms in testing processes. In other words, it can be said that all models in terms of R^2 exhibit good performance in predicting the daily global solar radiation. In contrast, the random forest performances in all three provinces are very close, where the algorithm could forecast global solar radiation values. In addition, it realised median error values in relation to the DS and RT algorithms.

It is established that the measured performance variables measured during the training stage are better than those achieved during the testing stage. In this case, the algorithm is over-fitted[82]. The slight change between the training and testing values shows that the algorithm takes neither under-

Table 4. Training and testing scores of data sets in different provinces.

Dataset						
Training	**Model**	**R^2**	**MAE**	**RMSE**	**RAE (%)**	**RRSE (%)**
Pretoria	**Random Tree**	0.7589	50.4163	63.9259	83.9146	86.6379
	Random Forest	0.8474	46.6435	57.6265	77.6351	78.1005
	Decision Stump	0.6011	44.2971	58.9462	73.7296	79.889
Testing						
	Random Tree	0.4536	34.1923	47.1923	56.6669	63.2978
	Random Forest	0.8606	46.8928	57.9526	77.7155	77.7302
	Decision Stump	0.5971	45.73.19	59.775	75.7915	80.1746
Training						
Bloemfontein	**Random Tree**	0.4941	57.9095	73.9708	81.4037	86.9073
	Random Forest	0.8659	52.985	64.7248	74.4812	76.0444
	Decision Stump	0.6003	55.5215	68.0403	78.0468	79.9396
Testing						
	Random Tree	0.5115	59.0075	71.6768	86.0792	86.6761
	Random Forest	0.8371	51.2585	62.4675	74.7751	75.5397
	Decision Stump	0.5115	59.0075	71.6768	86.0792	86.6761
Training						
Vuwani	**Random Tree**	0.3135	63.0092	78.7388	93.0238	94.9921
	Random Forest	0.8383	52.0188	64.0407	76.7982	77.2600
	Decision Stump	0.6002	53.1017	66.2741	78.3969	79.9544
Testing						
	Random Tree	0.2100	69.2305	83.0989	100.0000	100.0000
	Random Forest	0.8502	52.4642	63.4804	75.7818	76.3914
	Decision Stump	0.6023	52.0105	66.3464	75.1266	79.8403

fit nor overfit. The closer the R^2 value is to one, the better the datasets will fit with the regression line. As observed in **Table 3**, the RMSE for all algorithms in the different locations fluctuates between 47.1923 and 83.0989 for both the training and the testing process. Therefore, by comparing all the algorithms and following [82], the results showed that forecasting solar radiation using RF is more efficient than the DS and RT algorithms.

5.2. The Performance Analysis of Solar Radiation Based on Short-Term Weather Uncertainty

The hourly historical solar irradiance datasets from Pretoria, Bloemfontein and Vuwani were each divided into 80% for the training set and 20% for the testing set with data standardisation performances. Furthermore, the data set is also analysed in various conditions to illustrate several climate categories and uncertainty in solar irradiance, such as a clear day, a partially cloudy day, and a cloudy day. The analysis is necessary since the accuracy of the prediction irradiance is influenced by fluctuations during various weather categories. Therefore, the prediction algorithms were then trained for the real-time data with two different weather days. The days selected are the 1st of January and the 14th of May, all in 2018. Each day is chosen from a different station and analysed as a clear or partially cloudy and cloudy day, as presented in **Figure 7**. On a clear day, there are no fluctuations. **Figure 7a** the 1st of January is classified as a clear day in Pretoria with a maximum measured and predicted solar irradiance of 1124.20 and 945.45 W/m^2, respectively, while the same clear day is observed in **Figure 7b** in Bloemfontein with maximum measured and predicted solar irradiance of 1181.4 and 1165.3 W/m^2 respectively. **Figure 7c** shows that the same day in Vuwani is partially cloudy with variations with maximum measured and predicted solar irradiances of 1100.0 and 1024.4 W/m^2, respectively.

Similarly, the 14th of July is observed as a cloudy day in both Pretoria and Vuwani, as observed in **Figure 6a,c**, while **Figure 6b** in Bloemfontein is seen as a clear day. The maximum irradiances for the measured and predicted values are 1124.20 and 945.45 W/m^2, respectively, while in the case of a partially cloudy day, there are variations with measured and predicted values of irradiance recorded as 523.95 and 467.34 W/m^2, respectively. On a cloudy day, multiple severe irradiance fluctuations are noticed. A high prediction preci-

sion in clear-sky conditions and a low one in partly cloudy conditions can be observed.

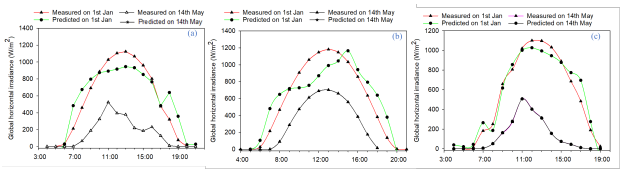

Figure 6. Forecasting of solar irradiance predicted and measured in different weather conditions at (**a**) Pretoria, (**b**) Bloemfontein (**c**) Vuwani on the 1st of January and the 14th of May 2018.

Since the type of weather on the 14th of May varies meaningfully from that of the 1st of January, in **Figure 6a,b**, the prediction approaches experienced terrible forecasting performance for higher varied days. Moreover, it has been noticed that the prediction of the performance of conventional standard algorithms decreased on cloudy days. This could be due to bad weather situations. Even a slight change produces a significant difference in irradiance. The quick variation of the cloud layers on cloudy days produces enormous challenges with irradiance forecasts. The results are tested and validated for each type of day using RMSE evaluations, as presented in **Figure 7**.

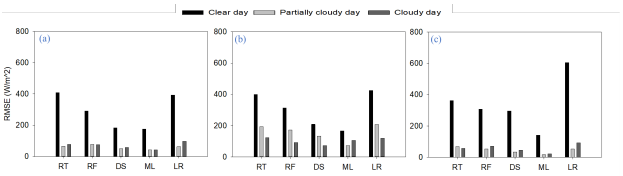

Figure 7. Prediction performance of the RMSE metric against different algorithms in (**a**) Pretoria, (**b**) Bloemfontein (**c**) Vuwani.

This proposed method shows that on a clear day, the RMSE values in Pretoria, Bloemfontein and Vuwani are 409.089, 398.372 and 361.604 W/m^2, respectively, for the RT model. On the contrary, during the cloudy day, the values of RMSE in Pretoria, Bloemfontein and Vuwani are 76.43, 92.89 and 71.28 W/m^2, respectively, for the RF model. Moreover, the study offers a different method of applying different models to predict solar irradiance. Compared to other statistical metrics, the RMSE is very high in forecasting solar irradiance on a clear day with the least value on a cloudy day; hence, **Figure 8** shows the normalised RMSE of **Figure 7**. It is obvious from the results established in **Figures 6–8** that with the increment in the prediction horizon, the forecast precision of algorithms decreases.

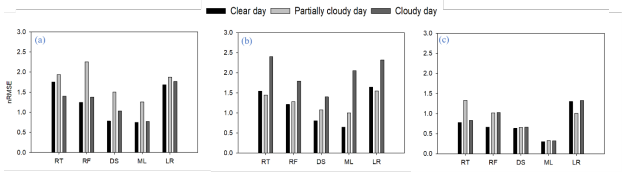

Figure 8. Prediction performance of the nRMSE metric against different algorithms in (**a**) Pretoria, (**b**) Bloemfontein (**c**) Vuwani.

Table 5 presents several numerical values of the metrics designed for the locations studied and the study models. As depicted in **Table 4**, R^2 values differ between 0.2 and 1.0, subject to the type of day, model, and location. However, it can be easily inferred from the results shown in **Table 5** that several algorithms, in terms of R^2 on a clear day in all locations, perform well in predicting daily global solar radiation.

Considering the clear day in these locations, the RF algorithm in terms of R^2 in Pretoria shows a performance of 0.942. This means that 94.2% of the data fit the regression model and are considered a strong correlation. Also, in Bloemfontein, both RF and MLR show an equal value of R^2 0.948 each as the highest value among all the algorithms. Therefore, this result shows that 94.8% of the data fit each regression model. The town of Vuwani has MLR as the highest model with an R^2 value of 0.964, therefore regarded as the best performance on a clear day among the three locations studied. Generally, the average RF performance is presented as the best performance among all the algorithms studied. The result agrees with many research works [35, 83] that examined the use of RF to forecast solar radiation values.

5.3. Long-Term Monthly and Yearly Analysis of Solar Radiation, Clearness Index, and Diffuse Fraction

The solar radiation attained on the Earth's surface changes over time with respect to the Sun's location and other weather parameters. Nevertheless, this radiation is unlimited. Solar energy flow remains an essential aspect of the energy source on Earth's surface. If the flux attained on the Earth's surface in a certain place differs significantly, particularly seasonally, the flux emitted by the Sun remains moderately unchanged. **Figure 9** shows that solar radiation fluctuates from month to month due to the rotation of the Earth in addition to South African weather conditions. In each location,

Table 5. Machine learning technique to characterise different daily weather conditions in different locations.

Clear Day in Pretoria					**Clear Day in Bloemfontein**				**Clear Day in Vuwani**			
Model	R^2	**MAE**	**RAE(%)**	**RRSE(%)**	R^2	**MAE**	**RAE (%)**	**RRSE(%)**	R^2	**MAE**	**RAE (%)**	**RRSE (%)**
RT	0.713	363.897	88.719	90.718	0.800	328.766	76.014	83.621	0.816	291.183	72.726	82.137
RF	0.942	268.866	65.550	64.258	0.948	281.339	65.049	65.772	0.901	283.988	70.929	70.099
DS	0.912	146.207	35.646	40.224	0.895	161.053	37.237	43.789	0.758	197.054	49.216	67.203
MLR	0.929	129.720	31.626	38.633	0.948	118.838	27.477	35.023	0.964	100.605	25.127	32.105
LR	0.824	344.764	84.054	86.647	0.838	373.476	86.350	89.243	0.905	544.767	136.062	137.466
Partially cloudy day in Pretoria					**Partially cloudy day in Bloemfontein**				**Partially cloudy day in Vuwani**			
RT	0.852	45.963	51.329	62.078	0.402	167.160	95.530	87.400	0.336	54.799	82.650	91.524
RF	0.506	64.273	71.777	72.305	0.659	127.241	72.717	77.819	0.556	45.273	68.026	70.029
DS	0.659	39.459	44.066	48.089	0.769	83.598	47.775	60.747	0.882	24.500	36.818	45.255
MLR	0.704	30.502	34.063	40.292	0.667	45.719	73.519	26.128	0.885	14.509	21.807	22.805
LR	0.792	53.232	59.447	60.039	0.123	161.524	92.309	93.912	0.759	45.131	67.813	69.400
Cloudy day in Pretoria					**Cloudy day in Bloemfontein**				**Cloudy day in Vuwani**			
RT	0.417	47.606	66.038	77.935	0.160	104.087	100.000	100.000	0.331	37.824	39.771	54.971
RF	0.513	51.755	71.793	76.655	0.400	73.873	70.973	74.7134	0.412	61.121	64.268	67.715
DS	0.502	44.271	61.411	57.592	0.397	44.472	42.726	58.395	0.593	31.824	33.463	43.851
MLR	0.630	30.243	41.952	42.806	0.520	65.102	62.546	85.339	0.586	18.615	19.574	21.585
LR	0.247	56.140	77.876	98.417	0.334	99.432	95.589	96.544	0.408	77.884	81.897	87.896

it is observed that solar radiation varies between 2.433 and 5.347 kWh/m^2/day in all provinces in different months of the year.

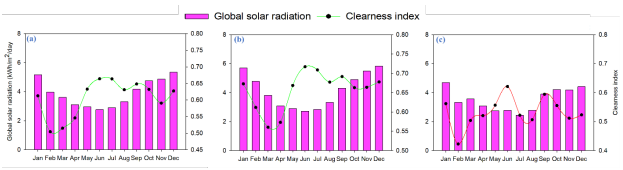

Figure 9. Comparison between solar radiation and clearness index with months of the year in different provinces (**a**) Pretoria (**b**) Bloemfontein (**c**) Vuwani.

In **Figure 9**, it can be observed that the maximum solar radiation found in December for both Pretoria and Bloemfontein is recorded as 5.347 and 5.844 kWh/m^2/day, respectively, while the maximum solar radiation maximum is found in January for Vuwani with a value of 4.692 kWh/m^2/day. Similarly, the average clearness index is 0.605, 0.657 and 0.533 in Pretoria, Bloemfontein and Vuwani, respectively, while their respective clearness index maximum is 0.663, 0.717 and 0.621, all in June. Among the three sites under study, the solar radiation and clearness index are higher in Bloemfontein. **Figure 9** suggests that the proposed method effectively predicts the monthly average global solar radiation. However, the month of June has the lowest solar radiation in Pretoria and Bloemfontein, recording 2.76 and 2.717 kWh/m^2/day, respectively, compared to that found at Vuwani in July with a value of 2.433 kWh/m^2/day. This study could conclude that the maximum amount of energy is available during summer (November to March) and the minimum in winter (June through August) in the studied locations.

Figure 10 shows a notable correlation between the diffuse fraction and the clearness index values, especially in Pretoria and Vuwani, with R^2 values of 0.8802 and 0.9034, respectively. It is noticed that there is a best fit of the scatter plot between the diffuse index and the clearness index value in Vuwani Limpopo province.

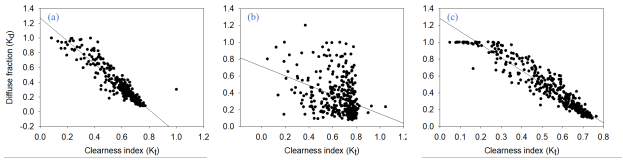

Figure 10. Scatter plot of diffuse fraction versus clearness index in different provinces (**a**) Pretoria, (**b**) Bloemfontein (**c**) Vuwani.

It is observed that there are comparative agreements between the values of the diffuse fraction and the clearness index, especially in **Figure 10a,c**. The presence of scattered plots between diffuse fraction and clearness index in **Figure 10b** could imply a high level of noise or randomness in the data, leading to uncertainty in the analysis. South Africa is dependent on coal, but efforts are ongoing to expand its ageing coal-fire. The decline of the existing fleet is evident as renewables, including solar PV, are fast growing in the country. **Figure 11a,b** demonstrate the relationship between the predicted and measured solar radiation on summer and winter days, respectively.

It can be seen that there is a good positive correlation between the model results and the measured data. The appropriate accuracy of the model can be observed from high R^2 values and a close relationship between the forecasted and measured solar radiations for both seasons, as indicated in **Figure 11a,b**. The RMSE for the chosen summer and winter

days is low, and it can be attributed to the fact that the model performs well for various seasons. While the model results indicate good forecasting ability, it is essential to note its deficiencies.

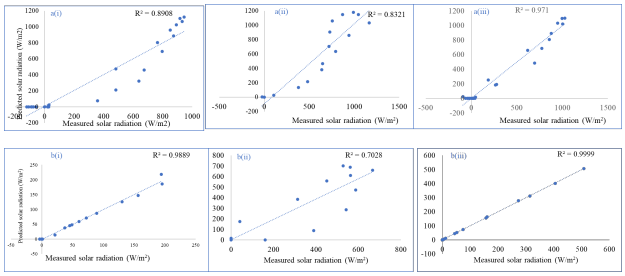

Figure 11. The relationship between predicted solar radiation and the measured solar radiation during (**a**) a summer day in **a(i)** Pretoria **a(ii)** Bloemfontein **a(iii)** Vuwani (**b**) a winter day in **b(i)** Pretoria **b(ii)** Bloemfontein **b(iii)** Vuwani.

5.4. Assessment of Machine Learning and Traditional Forecasting Approaches

Machine learning arises as a better option for solar radiation prediction because of its special ability to manage difficult patterns and various data sources. Compared to the traditional statistical approaches that depend on historical data patterns, ML performs well by expertly capturing difficult relationships within datasets, familiarising non-linear patterns, and providing more prediction accuracy. While physical models, including those focusing on satellite images and numerical weather predictions, have their advantages, they are often restricted in addressing the essential challenges of atmospheric conditions and may struggle to capture delicate connections, mainly under challenging situations. ML proposes the flexibility to integrate the strengths of statistical and physical models, generating hybrid models that excel in forecasting. An extensive summary of the advantages of ML and limitations in solar irradiance forecasting, as described in **Table 6**, emphasises the features contributing to its preference over traditional techniques.

Table 6. Comparison of machine learning and traditional forecasting approaches.

Feature	Machine Learning	Traditional Approaches	Capability	References
Managing difficult relationships	ML can capture difficult, non-linear relationships in data.	Traditional approaches may struggle with difficult relationships and non-linear patterns.	ML	
Nonlinearity	MLs are well-matched for capturing nonlinearities in data.	Traditional methods may find it difficult to model non-linear relationships.	ML	
Forecasting errors	ML models can reach little forecasting errors with suitable training data.	Traditional techniques' forecasting errors may be inadequate by their basic assumptions.	ML	
Automation and scalability	ML can be automated and scaled.	Traditional approaches may be deficient in automation and scalability.	ML	
Overfitting	MLs, if not correctly normalised, can be susceptible to overfitting.	Traditional approaches may be stronger when overfitting becomes an issue.	Tradition	
Managing missing Data	ML models may struggle with missing data, and imputation methods may be needed.	Traditional approaches may manage missing data more smartly	Tradition	[84]
Computational difficulty	Some developed MLs, mainly DL models, can be computationally costly.	Traditional approaches are mostly computationally cheap.	Tradition	
Interpretability	ML models, mostly DL, are normally regarded as "black boxes," making interpretation and understanding more difficult.	Traditional models are mostly more interpretable as they focus on identified physical principles or statistical relationships.	Tradition	[85]
Dynamic retraining	ML can be easily retrained and updated as new data becomes available, affirming that the models remain accurate over time.	Traditional approaches may be difficult to retrain and update, regularly demanding a complete improvement of the model to integrate new data.	ML	[13, 85]
High cost of maintenance	ML models may be expensive to maintain because of the demand for skilled personnel, data acquisition, and advanced computational resources.	Traditional approaches mostly have lesser initial costs since they depend on the current computational structure and simpler models.	Tradition	[86]

6. Discussion

The constant benefit of the random forest algorithm in terms of the training time emphasizes its effectiveness compared to other algorithms analysed in this work [87]. This effectiveness is remarkably important given the fluctuations in the number of GHI measured. Similarly, the reliability of RMSE parameters across numerous GHI selections indicates that the essential method of retrieving data from the weather station is sufficient for correct GHI prediction. This result undermines previous research that proposed substantial effects of GHI variety on model performance.

The findings in this research point to constant overvaluation of GHI in May under cloudy sky conditions in the three provinces, supporting the idea of the models' effort to perform successfully during months described by such weather patterns [88]. The existence of clouds due to the hindrance of sunlight is caused by factors such as the length of unclearness of the sun by clouds, the optical thickness of the clouds and secondary effects, including reflections between cloud layers or from the sides of clouds. These factors jointly influence the quantity of irradiance that attains the surface. Moreover, Weyll et al. [89] proposes that the decrease in statistical model activities can be characterized by the random performance of the atmosphere and the effect of clouds in controlling the pattern of irradiance over time. During overcast or partially cloudy conditions, the irradiance pattern leans to exhibit negligible autocorrelation within its time sequences. In this case, cloud transmission is, hence, the most flexible feature influencing surface irradiance in several geographical settings.

This study strengthens the efficiency of data-driven ML methods in medium-term GHI prediction. Significantly, it is established that the amount of solar radiation employed in the research did not meaningfully affect the model's productivity within the study areas. This indicates that the methodology shows an intensity of generalization capability, possibly appropriate to other provinces with varying datasets. Nevertheless, it is critical to recognize the impact of dataset attributes on study performances. The model described in this work has impacts on an increasing body of research on solar power forecasting for PV power plants in South Africa. While similar and complex research has been performed in other regions, just a few numbers are available for South Africa [90, 91]. Because of variations in the orientation and axis tracking capabilities of South African power plants, direct assessment with these studies is difficult. However, the model can be tested against others established in regions with similar and different climatic conditions to validate its forecasting accuracy.

Besides, this research highlights the comparative advantages of data-driven ML models over traditional models, remarkably regarding the computational proficiency and predictive precision for solar radiation predictions. By presenting a sustainable alternative to resources, the method not only offers the scientific knowledge of GHI forecast but also holds practical effects for improving the functional efficiency of solar energy management systems [92]. This supports wider sustainability aims and supports the incorporation of renewable energy sources into the energy grid.

7. Conclusion, Limitations and Future Works

This study investigated the short and long-term global solar radiation data predictions of three locations in different provinces in South Africa with various solar radiation distributions. The performance of five machine learning, including random tree, random forest, decision stump, multilinear regression and linear regression, are examined to predict daily and monthly solar radiation. It can be deduced from the results of hourly short-term forecasting that the R^2 values are 0.873 on a clear day and 0.990 on a cloudy day in Pretoria. Also, at Vuwani on a clear day, the values of R^2 are 0.8730 and 0.990 on a cloudy day. Therefore, the clearer the day, the less correlation and the better the forecast. In fact, the R^2 values in all the models in this study change from 53.7% to 98.6%, whereas the RMSE of all the algorithms in different provinces fluctuates between 47.1923 and 83.0989 for both the training and testing process depending on the locations. The algorithm used in all provinces has presented a very successful result. Considering all statistical metrics and study locations in South Africa, the best results are achieved within Bloemfontein. When the three locations and all algorithms are examined regarding RMSE, it is observed that many values found are close to zero. This means that several predicted results could be considered 'reasonable prediction' or "good prediction". Furthermore, it is pertinent to note that during the general assessment, random forest, multiliear

regression, and decision stump models present close results in the analysis based on the daily category in all locations. Hence, several parameters should be considered to choose the best of these models. The selection between machine learning and traditional approaches for solar irradiance forecast relies on several factors, such as the specific demands of the role, needed results, and available resources. Machine learning produces better accuracy, scalability, and adaptability but also accompanies higher computational difficulty and operation costs. However, traditional approaches offer easier interpretability at lower prices but may be rigid and robustly desirable for handling intricate, non-linear data designs.

Thus, it is obvious that the presented article has bridged an essential gap in the literature, contributing a strong and effective alternative to traditional models. The implications of the results go outside academia, indicating an improvement in the working efficiency of solar energy management systems, hence contributing to wider sustainability targets and renewable energy integration. Similarly, the results obtained in this research can assist South African governments in better policy implementations in terms of considerations for integrating solar photovoltaic systems as the country's core sources of electricity.

Although machine learning methods present considerable prospects for adaptive and accurate solar irradiance forecasts, they come with several limitations and complexities. These comprise reliance on huge and high-quality datasets, challenges in model interpretability, high computational prices, and the complexity of model selection, tuning, and feature engineering.

The result based on the solar radiation prediction of the proposed algorithm is compared with standard algorithms, bearing in mind weather unreliability, and the research has resulted in various weather types. It is noticed that the proposed model exceeded even with severe variations in the irradiance during cloudy and partially cloudy days. Hence, future research work will emphasise decreasing the error in solar irradiance prediction because of non-linear weather conditions.

Author Contributions

O.O.A.: Conceptualizations, methodology, writing-original draft, writing- final draft, software development. N.I.N.: Final corrections, funding, supervision.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data will be provided on request.

Acknowledgments

The authors wishes to express their gratitude to the Centre for Cyber-Physical Food, Energy & Water Systems (C.C.P-F.E.WS), University of Johannesburg, Johannesburg, South Africa.

Conflicts of Interest

The authors declare no known conflict of interest.

Abbreviation

AI	Artificial Intelligence
AIP	Adaptive Internet Protocol
ANFIS	Adaptive Neuro-Fuzzy Inference Systems
ANN	Artificial neural network
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
CNN	Convolution Neural Network
CO	Carbon (iv) Oxide
DHI	Direct Horizontal Irradiance
DL	Deep Learning
DS	Decision Stump
Ta	Amient Temperature
Rh	Relative Humidity
Ws	Wind Speed
DT	Decision Tree
EI_2	Energy Internet and Energy System Integration
ELM	Extreme Learning Machine
FFNN	Feedforward Neural Network
FIS	Fuzzy Inference System
GB	Gradient Boosting

GHI	Global Horizontal Irradiance
GRNN	General Regression Neural Network
GSR	Global solar radiation
IREC	International Renewable Energy Congress
Long-Term	Long term
LR	Linear Regression
LSTM	Long-Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLFFNN	Multilayer Feedforward Neural Network
MLP	Multilinear Perception
MLR	Multivariale Linear Regression
MPPT	Maximum Power Point Tracker
nRMSE	Normalized Root Mean Square Error
NWP	Numerical Weather Prediction Algorithm
PV	Photovoltaic
RAE	Root Absolute Error
RBFNN	Radial basis function networks
RF	Random Forest
RH	Relative Humidity
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
RRSE	Root Relative Square Error
RT	Random Tree
SAURAN	Southern African Universities Radiometric Network
SD	Standard Deviation
SVM	Supportive Vector Machine
SVR	Supportive Vector Regression

References

[1] Jebli, I., Belouadha, F.-Z., Kabbaj, M.I., et al., 2021. Deep learning based models for solar energy prediction. Advances in Science, Technology and Engineering Systems Journal. 6(1), 349–355.

[2] Apeh, O.O., Meyer, E.L., Overen, O.K., 2022. Contributions of Solar Photovoltaic Systems to Environmental and Socioeconomic Aspects of National Development—A Review. Energies. 15(16), 5963.

[3] Ritchie, H., Roser, M., Rosado, P., 2020. CO_2 and greenhouse gas emissions. Our World in Data.

[4] Okampo, E.J., Nwulu, N.I., 2020. Optimal energy mix for a reverse osmosis desalination unit considering demand response. Journal of Engineering Design and Technology. 18(5), 1287–1303.

[5] Azizi, N., Yaghoubirad, M., Farajollahi, M., et al., 2023. Deep learning based long-term global solar irradiance and temperature forecasting using time series with multi-step multivariate output. Renewable Energy. 206, 135–147.

[6] Maleki, S.A.M., Hizam, H., Gomes, C., 2017. Estimation of hourly, daily and monthly global solar radiation on inclined surfaces: models re-visited. Energies. 10(134).

[7] da Silva, V.J., da Silva, C.R., Almorox, J., et al., 2016. Temperature-based solar radiation models for use in simulated soybean potential yield. Australian Journal of Crop Science. 10(7), 926–932.

[8] Alfailakawi, M.S., Michailos, S., Ingham, D.B., et al., 2022. Multi-temporal resolution aerosols impacted techno-economic assessment of concentrated solar power in arid regions: Case study of solar power tower in Kuwait. Sustainable Energy Technologies and Assessments. 52, 102324.

[9] Akhter, M.N., Mekhilef, S., Mokhlis, H., et al., 2019. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. IET Renewable Power Generation. 13(7), 1009–1023.

[10] Antonanzas, J., Osorio, N., Escobar, R., et al., 2016. Review of photovoltaic power forecasting. Solar Energy. 136, 78–111.

[11] Ramedani, Z., Omid, M., Keyhani, A., et al., 2014. Potential of radial basis function based support vector regression for global solar radiation prediction. Renewable and Sustainable Energy Reviews. 39, 1005–1011.

[12] Das, U.K., Tey, K.S., Seyedmahmoudian, M., et al., 2018. Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews. 81, 912–928.

[13] Voyant, C., Notton, G., Kalogirou, S., et al., 2017. Machine learning methods for solar radiation forecasting: A review. Renewable Energy. 105, 569–582.

[14] Reikard, G., 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy. 83(3), 342–349.

[15] Reikard, G., Haupt, S.E., Jensen, T., 2017. Forecasting ground-level irradiance over short horizons: Time series, meteorological, and time-varying parameter models. Renewable Energy. 112, 474–485.

[16] Diagne, M., David, M., Lauret, P., et al., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renewable and Sustainable Energy Reviews. 27, 65–76.

[17] Zhang, G., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 50, 159–175.

[18] Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting. 14(1), 35–62.

[19] Bayrakçı, H.C., Demircan, C., Keçebaş, A., 2018. The development of empirical models for estimating global solar radiation on horizontal surface: A case study. Renewable and Sustainable Energy Reviews. 81, 2771–2782.

[20] Liu, Y., Zhou, Y., Chen, Y., et al., 2020. Comparison of support vector machine and copula-based nonlinear quantile regression for estimating the daily diffuse solar radiation: A case study in China. Renewable Energy. 146, 1101–1112.

[21] Sharma, A., Kakkar, A., 2018. Forecasting daily global solar irradiance generation using machine learning. Renewable and Sustainable Energy Reviews. 82, 2254–2269.

[22] Mellit, A., Benghanem, M., Arab, A.H., et al., 2005. A

simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach. Solar Energy. 79(5), 469–482.

[23] Apeh, O.O., Overen, O.K., Meyer, E.L., 2021. Monthly, seasonal and yearly assessments of global solar radiation, clearness index and diffuse fractions in Alice, South Africa. Sustainability. 13(4), 1–15.

[24] Khosravi, A., Koury, R.N.N., Machado, L., et al., 2018. Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. Journal of Cleaner Production. 176, 63–75.

[25] VanDeventer, W., Jamei, E., Thirunavukkarasu, G.S., et al., 2019. Short-term PV power forecasting using hybrid GASVM technique. Renewable Energy. 140, 367–379.

[26] Meenal, R., Selvakumar, A.I., 2018. Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. Renewable Energy. 121, 324–343.

[27] Jiang, H., Dong, Y., 2016. A nonlinear support vector machine model with hard penalty function based on glowworm swarm optimization for forecasting daily global solar radiation. Energy Conversion and Management. 126, 991–1002.

[28] Fan, J., Wang, X., Wu, L., et al., 2018. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy Conversion and Management. 164, 102–111.

[29] Das, U.K., Tey, K.S., Seyedmahmoudian, M., et al., 2017. SVR-based model to forecast PV power generation under different weather conditions. Energies. 10(7), 876.

[30] Wu, Y.-K., Huang, C.-L., Phan, Q.-T., et al., 2022. Completed review of various solar power forecasting techniques considering different viewpoints. Energies. 15(9), 3320.

[31] Kumari, P., Toshniwal, D., 2021. Deep learning models for solar irradiance forecasting: A comprehensive review. Journal of Cleaner Production. 318, 128566.

[32] Yang, D., Ye, Z., Lim, L.H.I., et al., 2015. Very short term irradiance forecasting using the lasso. Solar Energy. 114, 314–326.

[33] Chow, C.W., Urquhart, B., Lave, M., et al., 2011. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. Solar Energy. 85(11), 2881–2893.

[34] Engerer, N.A., 2015. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. Solar Energy. 116, 215–237.

[35] Ibrahim, I.A., Khatib, T., 2017. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Conversion and Management. 138, 413–425.

[36] Kalogirou, S.A., 2001. Artificial neural networks in renewable energy systems applications: A review. Renewable and Sustainable Energy Reviews. 5(4), 373–401.

[37] Jiang, H., Dong, Y., Xiao, L., 2017. A multi-stage intelligent approach based on an ensemble of two-way interaction model for forecasting the global horizontal radiation of India. Energy Conversion and Management. 137, 142–154.

[38] Olatomiwa, L., Mekhilef, S., Shamshirband, S., et al., 2015. A support vector machine–firefly algorithm-based model for global solar radiation prediction. Solar Energy. 115, 632–644.

[39] Heng, J., Wang, J., Xiao, L., et al., 2017. 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.

[40] Mishra, A., Kaushika, N.D., Zhang, G., et al., 2008. Artificial neural network model for the estimation of direct solar radiation in the Indian zone. International Journal of Sustainable Energy. 27(3), 95–103.

[41] Olorunfemi, B.O., Nwulu, N.I., Ogbolumani, O.A., 2023. Solar panel surface dirt detection and removal based on Arduino color recognition. MethodsX. 10, 101967.

[42] Yun, G.Y., Kong, H.J., Kim, H., et al., 2012. A field survey of visual comfort and lighting energy consumption in open plan offices. Energy and Buildings. 46, 146–151.

[43] Xuan, Z., Xuehui, Z., Liequan, L., et al., 2019. Forecasting performance comparison of two hybrid machine learning models for cooling load of a large-scale commercial building. Journal of Building Engineering. 21, 64–73.

[44] Runge, J., Zmeureanu, R., Le Cam, M., 2020. Hybrid short-term forecasting of the electric demand of supply fans using machine learning. Journal of Building Engineering. 29, 101144.

[45] Ghodrati, A., Zahedi, R., Ahmadi, A., 2022. Analysis of cold thermal energy storage using phase change materials in freezers. Journal of Energy Storage. 51, 104433.

[46] Fan, C., Xiao, F., Wang, S., 2014. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy. 127, 1–10.

[47] Bot, K., Ruano, A., Ruano, M.G., 2020. Forecasting electricity demand in households using MOGA-designed artificial neural networks. IFAC-PapersOnLine. 53(2), 8225–8230.

[48] Bian, H., Zhong, Y., Sun, J., et al., 2020. Study on power consumption load forecast based on K-means clustering and FCM–BP model. Energy Reports. 6,

693–700.

[49] Amasyali, K., El-Gohary, N.M., 2018. A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews. 81, 1192–1205.

[50] Deb, C., Zhang, F., Yang, J., et al., 2017. A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews. 74, 902–924.

[51] Walker, S., Khan, W., Katic, K., et al., 2020. Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. Energy and Buildings. 209, 109705.

[52] Grimaldo, A.I., Novak, J., 2020. Combining machine learning with visual analytics for explainable forecasting of energy demand in prosumer scenarios. Procedia Computer Science. 175, 525–532.

[53] Soulouknga, M.H., Coban, H.H., Falama, R.Z., et al., 2022. Comparison of different models to estimate global solar irradiation in the Sudanese Zone of Chad. Journal of Electronics and Telecommunications. 22(2), 63–71.

[54] Haq, E.U., Lyu, X., Jia, Y., et al., 2020. Forecasting household electric appliances consumption and peak demand based on hybrid machine learning approach. Energy Reports. 6, 1099–1105.

[55] Hafeez, G., Alimgeer, K.S., Khan, I., 2020. Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid. Applied Energy. 269, 114915.

[56] Khan, A., Chiroma, H., Imran, M., et al., 2020. Forecasting electricity consumption based on machine learning to improve performance: A case study for the organization of petroleum exporting countries (OPEC). Computers and Electrical Engineering. 86, 106737.

[57] Kazemzadeh, M.-R., Amjadian, A., Amraee, T., 2020. A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting. Energy. 204, 117948.

[58] Nematchoua, M.K., Orosa, J.A., Afaifia, M., 2022. Prediction of daily global solar radiation and air temperature using six machine learning algorithms; a case of 27 European countries. Ecological Informatics. 69, 101643.

[59] Moreno, A., Gilabert, M.A., Martínez, B., 2011. Mapping daily global solar irradiation over Spain: A comparative study of selected approaches. Solar Energy. 85(9), 2072–2084.

[60] Wang, L., Kisi, O., Zounemat-Kermani, M., et al., 2017. Prediction of solar radiation in China using different adaptive neuro-fuzzy methods and M5 model tree. International Journal of Climatology. 37(3), 1141–1155.

[61] Chen, J.-L., Li, G.-S., Wu, S.-J., 2013. Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy Conversion and Management. 75, 311–318.

[62] Badrudeen, T.U., Nwulu, N.I., Gbadamosi, S.L., 2023. Neural network based approach for steady-state stability assessment of power systems. Sustainability. 15(2), 1667.

[63] Govindasamy, T.R., Chetty, N., 2021. Machine learning models to quantify the influence of PM10 aerosol concentration on global solar radiation prediction in South Africa. Clean Engineering and Technology. 2, 100042.

[64] Obiora, C.N., Ali, A., Hasan, A.N., 2020. Forecasting hourly solar irradiance using long short-term memory (LSTM) network. Proceedings of the 2020 11th International Renewable Energy Congress (IREC); Hammamet, Tunisia; 29–31 October 2020. pp. 1–6.

[65] Ozoegwu, C.G., 2019. Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number. Journal of Cleaner Production. 216, 1–13.

[66] Mehdizadeh, S., Behmanesh, J., Khalili, K., 2016. Comparison of artificial intelligence methods and empirical equations to estimate daily solar radiation. Journal of Atmospheric and Solar-Terrestrial Physics. 146, 215–227.

[67] Antonopoulos, V.Z., Papamichail, D.M., Aschonitis, V.G., et al., 2019. Solar radiation estimation methods using ANN and empirical models. Computers and Electronics in Agriculture. 160, 160–167.

[68] Adeala, A.A., Huan, Z., Enweremadu, C.C., 2015. Evaluation of global solar radiation using multiple weather parameters as predictors for South Africa provinces. Thermal Science. 19(suppl. 2), 495–509.

[69] Bugaje, I.M., 2006. Renewable energy for sustainable development in Africa: a review. Renewable and Sustainable Energy Reviews. 10(6), 603–612.

[70] Younes, S., Claywell, R., Muneer, T., 2005. Quality control of solar radiation data: Present status and proposed new approaches. Energy. 30(9), 1533–1549.

[71] Meddage, D.P.P., Ekanayake, I.U., Herath, S., et al., 2022. Predicting bulk average velocity with rigid vegetation in open channels using tree-based machine learning: a novel approach using explainable artificial intelligence. Sensors. 22(12), 4398.

[72] Al-Rousan, N., Al-Najjar, H., Alomari, O., 2021. Assessment of predicting hourly global solar radiation in Jordan based on Rules, Trees, Meta, Lazy and Function prediction methods. Sustainable Energy Technologies and Assessments. 44, 100923.

[73] Ren, Q., Cheng, H., Han, H., 2017. Research on machine learning framework based on random forest algorithm. Proceedings of the International Conference on Advances in Materials, Machinery, Electronics (AMME 2017); Wuhan, China; 25–26 February 2017. Volume 1820, no. 1.

[74] Ho, T.K., 1995. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition; Montreal, QC, Canada; 14–16 August 1995. Volume 1, pp. 278–282.

[75] Probst, P., Wright, M.N., Boulesteix, A., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9(3), e1301.

[76] Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 7, 1–13.

[77] Nadi, A., Moradi, H., 2019. Increasing the views and reducing the depth in random forest. Expert Systems with Applications. 138, 112801.

[78] Gressling, T., 2020. Automated machine learning. Data Science in Chemistry. Springer Publishing, New York, NY, USA.

[79] Nhu, V.-H., Shahabi, H., Nohani, E., et al., 2020. Daily water level prediction of Zrebar Lake (Iran): A comparison between M5P, random forest, random tree and reduced error pruning trees algorithms. ISPRS International Journal of Geo-Information. 9(8), 479.

[80] Petropoulos, F., Apiletti, D., Assimakopoulos, V., et al., 2022. Forecasting: theory and practice. International Journal of Forecasting. 38(3), 705–871.

[81] Khorasanizadeh, H., Mohammadi, K., Goudarzi, N., 2016. Prediction of horizontal diffuse solar radiation using clearness index based empirical models: A case study. International Journal of Hydrogen Energy. 41(47), 21888–21898.

[82] Gupta, R., Yadav, A.K., Jha, S.K., et al., 2023. Long term estimation of global horizontal irradiance using machine learning algorithms. Optik (Stuttgart). 283, 170873.

[83] Sun, H., Gui, D., Yan, B., et al., 2016. Assessing the potential of random forest method for estimating solar radiation using air pollution index. Energy Conversion and Management. 119, 121–129.

[84] Marquez, R., Coimbra, C.F.M., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. Solar Energy. 85(5), 746–756.

[85] Lefèvre, M., Oumbe, A., Blanc, P., et al., 2013. McClear: A new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmospheric Measurement Techniques. 6(9), 2403–2418.

[86] Reikard, G., Hansen, C., 2019. Forecasting solar irradiance at short horizons: Frequency and time domain models. Renewable Energy. 135, 1270–1290.

[87] Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

[88] Smith, C.J., Bright, J.M., Crook, R., 2017. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. Solar Energy. 144, 10–21.

[89] Weyll, A.L.C., Kitagawa, Y.K.L., Araujo, M.L.S., et al., 2024. Medium-term forecasting of global horizontal solar radiation in Brazil using machine learning-based methods. Energy. 300, 131549.

[90] Apeh, O.O., Nwulu, N.I., 2024. The water-energy-food-ecosystem nexus scenario in Africa: Perspective and policy implementations. Energy Reports. 11, 5947–5962.

[91] Apeh, O.O., Nwulu, N., 2024. The Food-Energy-Water Nexus Optimization: A Systematic Literature Review. Research on World Agricultural Economy. 5(4).

[92] Apeh, O.O., Nwulu, N.I., 2024. Unlocking Economic Growth: Harnessing Renewable Energy to Mitigate Load Shedding in Southern Africa. e-Prime-Advances in Electrical Engineering, Electronics and Energy. 100869.