sustainability

MDPI

Article

A Novel Machine Learning Approach for Solar Radiation Estimation

Hasna Hissou [1], Said Benkirane [2], Azidine Guezzaz [2,*], Mourade Azrour [3,*] and Abderrahim Beni-Hssane [1]

[1] Faculty of Science, Science and Technology Research Structure, Chouaïb Doukkali University, El Jadida 24000, Morocco; hissou.h@ucd.ac.ma (H.H.); beni-hsaane.a@ucd.ac.ma (A.B.-H.)
[2] Technology Higher School Essaouira, Cadi Ayyad University, Essaouira 44000, Morocco; said.benkirane@uca.ma
[3] IDMS Team, Faculty of Sciences and Techniques, Moulay Ismail University of Meknès, Errachidia 25000, Morocco
* Correspondence: a.guezzaz@uca.ma (A.G.); mo.azrour@umi.ac.ma (M.A.)

Abstract: Solar irradiation (Rs) is the electromagnetic radiation energy emitted by the Sun. It plays a crucial role in sustaining life on Earth by providing light, heat, and energy. Furthermore, it serves as a key driver of Earth's climate and weather systems, influencing the distribution of heat across the planet, shaping global air and ocean currents, and determining weather patterns. Variations in Rs levels have significant implications for climate change and long-term climate trends. Moreover, Rs represents an abundant and renewable energy resource, offering a clean and sustainable alternative to fossil fuels. By harnessing solar energy, we can actively reduce greenhouse gas emissions. However, the utilization of Rs comes with its own challenges that must be addressed. One problem is its variability, which makes it difficult to predict and plan for consistent solar energy generation. Its intermittent nature also poses difficulties in meeting continuous energy demand unless appropriate energy storage or backup systems are in place. Integrating large-scale solar energy systems into existing power grids can present technical challenges. Rs levels are influenced by various factors; understanding these factors is crucial for various applications, such as renewable energy planning, climate modeling, and environmental studies. Overcoming the associated challenges requires advancements in technology and innovative solutions. Measuring and harnessing Rs for various applications can be achieved using various devices; however, the expense and scarcity of measuring equipment pose challenges in accurately assessing and monitoring Rs levels. In order to address this, alternative methods have been developed with which to estimate Rs, including artificial intelligence and machine learning (ML) models, like neural networks, kernel algorithms, tree-based models, and ensemble methods. To demonstrate the impact of feature selection methods on Rs predictions, we propose a Multivariate Time Series (MVTS) model using Recursive Feature Elimination (RFE) with a decision tree (DT), Pearson correlation (Pr), logistic regression (LR), Gradient Boosting Models (GBM), and a random forest (RF). Our article introduces a novel framework that integrates various models and incorporates overlooked factors. This framework offers a more comprehensive understanding of Recursive Feature Elimination and its integrations with different models in multivariate solar radiation forecasting. Our research delves into unexplored aspects and challenges existing theories related to solar radiation forecasting. Our results show reliable predictions based on essential criteria. The feature ranking may vary depending on the model used, with the RF Regressor algorithm selecting features such as maximum temperature, minimum temperature, precipitation, wind speed, and relative humidity for specific months. The DT algorithm may yield a slightly different set of selected features. Despite the variations, all of the models exhibit impressive performance, with the LR model demonstrating outstanding performance with low RMSE (0.003) and the highest R2 score (0.002). The other models also show promising results, with RMSE scores ranging from 0.006 to 0.007 and a consistent R2 score of 0.999.

Citation: Hissou, H.; Benkirane, S.; Guezzaz, A.; Azrour, M.; Beni-Hssane, A. A Novel Machine Learning Approach for Solar Radiation Estimation. *Sustainability* **2023**, *15*, 10609. https://doi.org/10.3390/su151310609

Academic Editor: Ljubomir Jankovic

Received: 22 May 2023
Revised: 20 June 2023
Accepted: 3 July 2023
Published: 5 July 2023

Keywords: sustainable energy; solar radiation; times series; machine learning; feature selection; forecasting

1. Introduction

Rs refers to the energy that is emitted by the Sun and travels through space to reach the Earth. It consists of electromagnetic waves, including visible light, ultraviolet (UV) rays, and infrared (IR) radiation. The Sun emits solar radiation in all directions, and a small fraction of it reaches the Earth's atmosphere. As the radiation passes through the atmosphere, it may be absorbed, scattered, or reflected via gases, particles, and clouds. Eventually, a portion of the solar radiation reaches the Earth's surface.

All life is powered by the Sun. It keeps the thermal energy and power equilibrium in the Earth's atmospheric conditions and ecological systems. Even a minor fluctuation in the Sun's radiation emission would have a notable effect on the climate of the Earth [1]. It influences Earth's climate system, influencing temperature variations, atmospheric circulation, and weather system formation, through heating of the atmosphere, oceans, and land. Its distribution across the globe contributes to regional climate differences and temperature gradients between the equator and poles. In the Earth's water cycle, solar radiation plays a crucial role, in providing the energy needed for processes like evaporation, condensation, and precipitation, which, in turn, affect mean sea level fluctuations. It also fuels atmospheric instability, which leads to the formation of severe weather phenomena such as storms, hurricanes, and extreme weather events, through the heating of the Earth's surface and the creation of pressure differences. Moreover, solar radiation stands as the most abundant and readily available energy source on Earth, serving as the primary energy source for natural processes and ecosystems, as well as representing the most environmentally friendly and sustainable form of energy [2]. It is also non-polluting, highly accessible, safe, and has the potential to slow the intensification of greenhouse effects [3,4]. Through harnessing solar radiation through technologies like solar panels and solar thermal systems, it can be converted and used in many ways, such as generating electricity, pumping water, and heating and purifying air and water [5]. Solar radiation at the planet's surface consists of three main components: direct radiation, which is sunlight that reaches Earth's surface without scattering or absorption; diffuse radiation, which is sunlight that has been scattered by the atmosphere and arrives from different directions; and additional radiation, resulting from the reflection and scattering of sunlight by the surrounding environment. Understanding these components is crucial for optimizing solar energy systems, designing efficient lighting in buildings, and studying microclimates. Insolation refers to the total ground radiation [6]. The world is dealing with serious issues such as industrial air pollution, global warming, and environmental destruction. Due to the detrimental environmental impact of using non-renewable resources derived from ancient organisms, such as coal, oil, and natural gas, which have harmed the environment, it has become crucial to seek alternative solutions. Thankfully, clean energy, such as solar energy, offers a pathway through which to address these issues effectively and create a more sustainable future. By transitioning to clean energy sources, we can mitigate climate change, improve air quality, enhance energy security, preserve ecosystems, alleviate energy poverty, and promote sustainable economic growth. Embracing clean energy technologies is crucial in tackling these pressing global problems and fostering a more resilient and environmentally friendly world [7]. Rs data must be available to assess the solar energy capacity of a specific area and integrate it into an electrical network [4]. Unfortunately, assessing the potential of Rs as a renewable energy source can be challenging for most weather stations worldwide. This is primarily due to the fluctuating and intermittent nature of the Rs resource and its lack of steadiness and non-controllability, which limits access to reliable data. Also, expensive equipment is required in order to measure it. These characteristics aggravate the situation. It makes the management of the grid more intricate, disturbs the balance between production and

consumption, causes variations in voltage, and raises concerns about quality and stability. The initial cost of installing solar power systems, although decreasing, can still be relatively high, posing a financial barrier for some individuals or businesses. Therefore, it is essential to evaluate Rs effectively using other meteorological factors, including relative humidity, ambient temperature, wind velocity, cloud cover, and other parameters [4,8]. For the purpose of estimating Rs from these readily available weather data, quite a few methods have been proposed, such as physical models, with complex structures due to complex conditions of the environment [9]. Empirical models generate a regression-based formula, either linear or not, that is simple with limited precision [10]. Statistical models, built on statistical correlation, are more accurate, but cannot fully express the nonlinear association between Rs and other factors [11,12]. ML models are of great interest to researchers worldwide because they can solve highly nonlinear problems with high accuracy compared to the other models [13–15]. All ML methods, mainly supervised models, typically require a compromise between model accuracy and complexity [16,17]. Accordingly, determining the optimum input aggregation for prediction models is indispensable. It avoids impertinent or extra information, while exclusively keeping the most required features. This mechanism is known as Feature Selection (FS) [8]. It lowers computational costs, improves performance and over-fitting issues, and enhances multicollinearity problems and model complexity [8,18,19]. The instructions for the FS technique include generating subsets, evaluating them, setting stopping criteria, and validating results [20]. Pertaining to this article, we propose a framework for feature selection (FS) aimed at categorizing different lag values. Our objective is to investigate how FS models can enhance forecasting quality in the field of Rs. Our research comprises two significant contributions:

- We employ an FS method with which to pick out essential feature sets from the initial feature sets using various models;
- We measure each model's feature importance score, RMSE, and R2 against the others by assessing their performance on an NCEP (National Centers for Environmental Prediction) dataset.

The choice of MVTS is driven by its ability to handle multiple variables and their complex relationships, capturing their interdependencies and improving accuracy. By leveraging MVTS analysis, we aim to contribute to the existing research and demonstrate its effectiveness in accurately estimating solar radiation.

The subsequent sections of this paper are structured as follows: Section 2 provides an overview of the field's historical background and its current state. Section 3 elucidates the methodology used in this research and presents the introduced model. Section 4 outlines the working environment, presents the findings, and engages in a thorough discussion. Lastly, Section 5 accentuates the key findings and outlines potential avenues for future research.

2. Related Works

Many researchers have proposed and published research studies on Rs estimation using ML models. However, only some studies have comprehensively examined the complete procedure of developing ML models. In particular articles, the discussion on FS methods was limited to a brief overview [1,8,16,21–28].

In their study, Diagne et al. [19] explored an approach that combined statistical, satellite-based, and numerical weather prediction (NWP) techniques. They also analyzed the proposed techniques' application conditions and spatial/temporal resolution ranges. Yadav and Chandel [20] presented forecasting models for different time horizons (short-term, medium-term, and long-term) of solar irradiation (Rs). They also evaluated methods used for selecting input parameters in these forecasting models. Kumar et al. [21] summarized widely utilized empirical regression models and the ANN model. Their findings unequivocally demonstrated that ANN defeated empirical regression models. Meenal, Selvakumar, and Pang et al. analyzed diverse ML models. They discovered that, although the ANN did not exhibit high predictive accuracy, it did propose a means for enhancing algorithm quality [16,22]. Voyant et al. examined the performance of ANN, SVM, and

tree-based models, noted that they produced comparable accuracy, and recommended using combined models [8]. In separate studies, both Chen et al. and Olatomiwa et al. used SVM for their research. The results indicated that its accuracy varied depending on the kernel functions used, and the optimized SVM demonstrated successful outcomes [23,29]. Mohanty et al. [26] conducted a study on the strengths and weaknesses of three models: an Adaptive Network-based Fuzzy Inference System (ANFIS), a radial basis function neural network (RBF-NN), and a multi-layer perception (MLP) model [30,31]. The ANFIS model was the focal point of discussion among the researchers, who emphasized its distinctive attributes as a hybrid intelligent model. The research team showcased its integration of conventional mathematical approaches, underscoring its uniqueness. It combines fuzzy logic and neural networks and improves learning and adaptability capabilities, yielding better results. Hedar et al. created an entirely new hybrid ML model that employs an auxiliary numerical data model with which to assess the accuracy of GHI predictions. The suggested hybrid approaches employ FS, classification, regression ML paradigms, and NWP models. Upon implementation on a dataset, the hybrid model lessened the RMSE [25]. Guermoui et al. [24] thoroughly investigated hybrid machine-learning techniques and defined five classifications: generalized, cluster-based, decomposition-based, decomposition-clustering-based, and ensemble learning methods incorporating evolutionary techniques. In their study, Ağbulut, Gürel, and Biçen [32] conducted a comparison of four ML algorithms (SVM, k-NN, DL, ANN) for predicting daily global solar radiation. The findings revealed that, in general, ANN outperformed DL, SVM, and k-NN in terms of prediction accuracy, while k-NN exhibited the least favorable performance among the algorithms assessed. Huang et al. developed a comprehensive ensemble of twelve ML methods, including a stacking model that combines the strengths of various algorithms in order to predict and compare daily and monthly Rs measurements accurately. According to the results, the XGBoost and stacking models, which combine RF, Gaussian Process Regression (GPR), GBRT, and XGBoost, exhibited superior predictive performance [1]. Guermoui, Bouchouicha, Bailek, and Boland [32] introduced a novel integrated model that combines a decomposition technique with an Extreme Learning Machine for predicting photovoltaic power generation. The performance of the proposed model was assessed using data from three distinct solar photovoltaic power plants situated in different locations with varying climatic conditions. The results indicated that the normalized error consistently remained below 10%, and the correlation coefficient exceeded 99% across the forecasting horizons. These findings demonstrate the effectiveness and accuracy of the proposed integrated model in forecasting photovoltaic power generation.

The following diagram (Figure 1) depicts the categorization of generalized models employed in forecasting Rs, while the Table 1 examines and assesses recent studies on Rs prediction utilizing machine learning techniques.

Table 1. Categorization of contemporary research pertaining to the prediction of Rs through the utilization of machine learning techniques.

Contribution	Date	Forecasting Model	Geographical Position	Optimal Model	Findings
[33]	2022	The hybrid CXGBRFR framework integrates deep learning CNN, XGB (Extreme Gradient Boosting) + RF, and bird-inspired models like HHD-BN (Harris Hawks Deep Belief Network), DNN (Deep Neural Network), ANN, ELM (Extreme Learning Machine), and MARS (Multivariate Auto-Regressive Spline models)	Australia daily	deep hybrid CXGBRFR	Correlation coefficient (r): deep hybrid CXGBRFR: 0.941–0.962 ANN/ELM: 0.934–0.956/0.954 DBN: 0.495–0.911 DNN: 0.922–0.941 MARS: 0.928–0.935 Legate's and McCabe's Index: deep hybrid CXGBRFR: 0.943–0.962, ANN: 0.933–0.958 ELM: 0.931–0.955 DBN: 0.493–0.911 DNN: 0.922–0.941 MARS: 0.928–0.942

Table 1. *Cont.*

Contribution	Date	Forecasting Model	Geographical Position	Optimal Model	Findings
[34]	2022	Seasonal Auto-Regressive Integrated Moving Average (SARIMA), K-Nearest Neighbors (KNN) Recursive Neural Network-Long Short-Term Memory (RNN-LSTM)	UAE daily	RNN-LSTM and KNN	RNN-LSTM and KNN outperform SARIMA. RNN-LSTM and KNN perform similarly RNN-LSTM slightly outperforms KNN
[27]	2022	SVM and Corrected-SVM	Ghardaia, Algeria	C-SVM	RMSE = 11.35% rRMSE = 1.713 MJ/m^2, MABE = 1.623 MJ/m^2 r = 12.61%
[28]	2022	LM, SCG, and RP	6 locations from Tamil Nadu, India	LM	LM: R = 0.9376 for training data, 0.9340 For testing data.
[35]	2022	ANN, CNN, RNN, SVR, PR RF	4 locations in Nigeria	RNN	Deep learning outperforms RNN: r = 0.9546, RMSE= 82.22 W/m^2, MAE = 36.52 W/m^2
[36]	2022	DL, SMGRT, and ANFIS	Isparta, Turkey	SMGRT	SMGRT is the best MSE = 1.878 R2 = 0.960 MBE = 0.156 RMSE = 1.371
[37]	2022	RNN, LSTM, and GRU	5 cities Bangladesh	GRU	MAPE = 19.28%
[1]	2021	GPR, RF GBRT, XGBoost {RF, GPR, XGBoost, GBRT}	12 sites in China	{GBRT, XGBoost, GPR, RF} XGBoost	Daily predictions: Stacking model outperforms Monthly predictions: Comparable performance
[5]	2021	MLP (XE MLP) SVR MLR LightGBM	Fez, Morocco	LightGBM SVR	LightGBM: Coefficient of determination R2 = 0.9377, RMSE = 0.4827 kWh/m^2 MAE = 0.3614 kWh/m^2
[38]	2021	22 empirical models RF, MLP, bagged trees, boosted trees	5 locations in Morocco.	RF	r ranges from 0.8753 to 0.9620, normalized mean absolute error (nMAE) ranges from 5.84 to 11.81%, negative root mean square error (nRMSE) ranges from 7.85 to 15.33%.
[39]	2021	SVM RF	India	RF	RF: MSE = 0.750 R2 = 0.97 SVM: MSE = 0.867—R2 = 0.9385
[4]	2021	K-Nearest Neighbors (k-NN) ANN, DL, SVM	4 Turkish stations	ANN	MBE = 0.195 MJ/m^2 RMSE = 2.157 MJ/m^2—rRMSE = 14.10% T statistic = 1.280 MJ/m^2—Mape = 15.92% MABE = 1.597—R2 = 0.9320%
[22]	2020	ANN, RNN	Tuscaloosa, Alabama in the USA	RNN with higher computational costs than ANN.	RNN: better prediction results Cloud cover impacts GSR prediction. RMSE = 7.64%, Normalized Mean Bias Error (NMBE) = 0.2%
[40]	2020	MLPd and RBF	Ghardaia in Algeria	MLP	MLP demonstrates slightly superior performance.
[41]	2020	ANN	Sapporo, Tateno, Fukuoka, Ishigakijima, and Minamitorishima in Japan	NA	Monthly diffuse, direct, and GRS forecasts are extremely accurate. All locations have a R2 of 0.988 or higher.

Table 1. *Cont.*

Contribution	Date	Forecasting Model	Geographical Position	Optimal Model	Findings
[42]	2020	M5Tree, CatBoost, and XGBoost SVM, RF	15 provinces in China	SVM	CatBoost outperforms.
[43]	2019	Naive Bayes 2 days ahead global horizontal irradiance	Austin, TX in USA	Naive Bayes	Various weather type: MBE = 2.73%, r = 86.33% Clear days: RMBE = 1.49%, r = 99.85%.
[44]	2019	RF, M5, MARS, CART	India (Gorakhpur side)	RF	RF: highest accuracy, CART: lowest accuracy.
[45]	2019	SVR, ANN, and DT	4 provinces in turkey	NA	Boosting improves prediction performance RMSE between 4.6 and 14.6%
[46]	2019	SVR, GPR, MLP, and Extreme Learning Machines (ELM)	Toledo in Spain	ELM	Satellite measurements improved predictability by increasing input parameters. ELM: RMSE = 60.60 W/m^2, r2 = 96%
[47]	2019	SP, ANN, and R	Odeillo in France	RF	nRMSE: 19.65% (GHI—first hour ahead), 27.78% (GHI—sixth hour ahead), 34.11% (Beam Normal Irradiation—first hour ahead), 49.08% (Beam Normal Irradiation—sixth hour ahead), 35.08% (Diffuse Horizontal Irradiation—first hour ahead) 49.14% (Diffuse Horizontal Irradiation—sixth hour ahead).
[48]	2019	SVR	Gurugram in India	NA	Performance SVR is influenced by the air temperature (the most significant parameter) RMSE = 14.3 MJ/m^2
[49]	2019	ANN, k-NN, empirical models	Fez -Morocco	KNN Hybrid model	k-NN: rRMSE = 0.2027 R2 = 0.9663. Hybrid model (k-NN—ANN): rRMSE = 0.1785, R2 = 0.9750.
[50]	2018	Radial basis function (RBF) MLP GPR	Ghardaïa—Algeria. Daily	GPR	MBE = 0.1861 kWh/m^2 nRMSE = 5.2%, r = 0.9842 RMSE = 0.3194 kWh/m^2,
[51]	2018	ANN Regression Analysis	4 stations in Turkey Monthly	ANN	ANN: R2 = 0.961, RMSE =0.14
[52]	2018	SVM XGBoost	China Daily	XGBoost	RMSE = 0.9238 kWh/m^2 R2 = 0.7530, XGBoost MAE = 0.6925 kWh/m^2—training phase = 3.02 s—testing phase = 0.05 s
[2]	2018	GPR	Mashha Iran Daily, Monthly	NA	Daily: RMSE = 0.16 MAPE = 1.97%, Model Efficiency (EF) = 0.99
[16]	2018	SVM ANN	India Monthly	SVM	SVM > ANN ANN: more accurate with long training time for large dataset. R2(ANN) = 0.9968 R2 (SVM) = 0.9912
[53]	2017	ANFIS, SVM, ANN	6 provinces in Mexico daily	SVM	RMSE = 2.578, R2 = 0.689 MAE = 1.97
[54]	2017	ANN	13 different stations	ANN	rMBE < 4% R2 = 0.64 r = 0.800 rRMSE = 13%

Table 1. *Cont.*

Contribution	Date	Forecasting Model	Geographical Position	Optimal Model	Findings
[55]	2017	MLP ANFIS SVM DT	Egypt Daily	MLP	MLP > ANFIS > SVM > DT
[56]	2016	ANN	Italy Monthly	ANN	MAPE = 1.67% to 4.25% based on the type and number of inputs
[57]	2016	Generalized Regression Neural Network (GRNN) Radial Basis Neural Network (RBNN) MLP	12 Sites (China) Daily	MLP	RBNN > GRNN > MLP R2 = 0.86 MAE = 0.425 kWh/m^2 RMSE = 0.5388 kWh/m^2
[58]	2016	ANN, ANFIS Gene Expression Programming (GEP)	Karmen, Iran Daily	ANN	R2 = 0.935
[7]	2015	SVR, Empirical	2 provinces (Iran)	SVR	RMSE = 0.4515 kWh/m^2 R2 = 0.9330
[21]	2015	Regression models. ANN	(1 month)	ANN	ANN > Regression models.
[59]	2015	Linear techniques SVM	Italy 1 Day	SVM	SVM > linear model
[60]	2015	k-NN	USA 30 min	K-NN	k-NN > persistence Enhancements in the forecast between 10% and 25%
[61]	2015	ANN—k-NN—SVR Autoregressive models persistence	Italia hourly	SVR	SVR > ANN > AR > k-NN > persistence

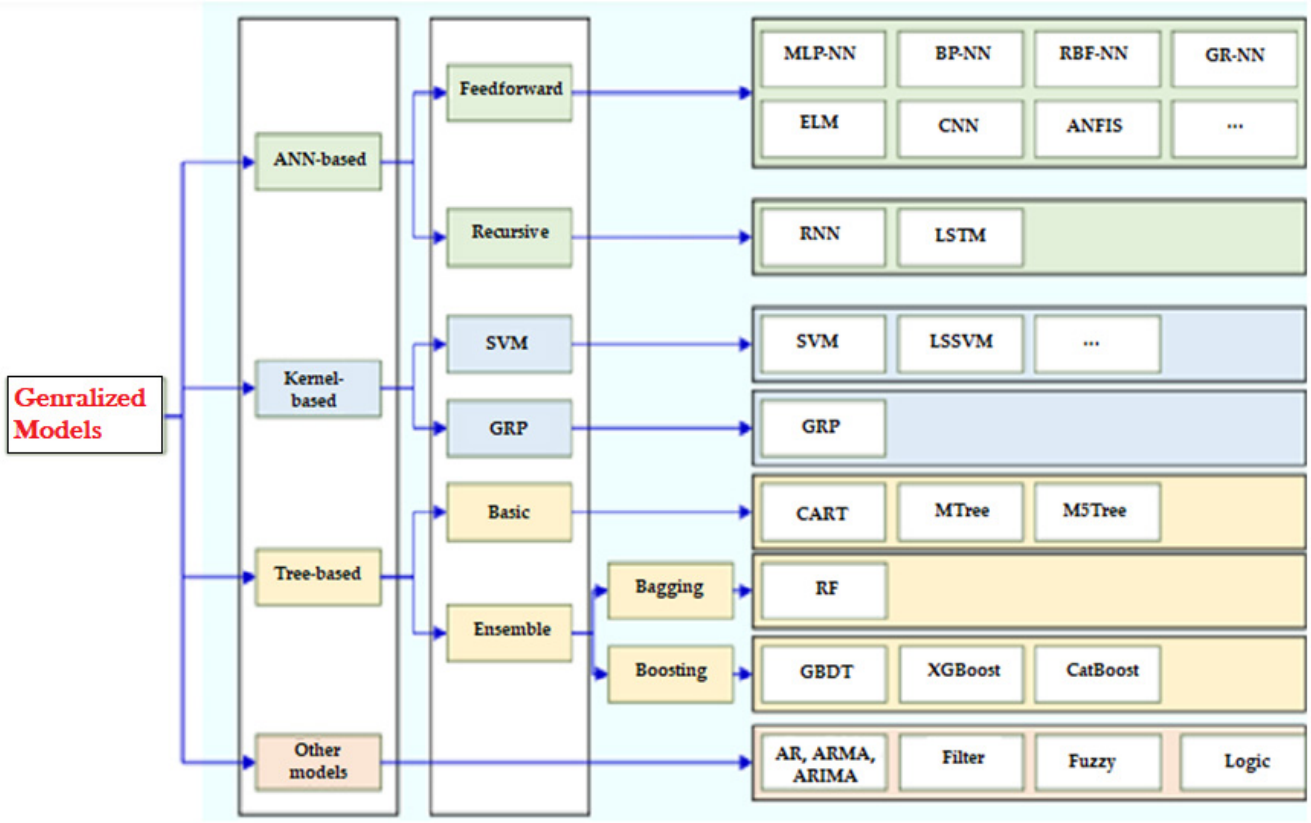

Figure 1. Categorization of the generalized models employed in predicting Rs.

Finally, the comparison of various models reveals interesting insights into their performance. The deep hybrid CXGBRFR model consistently demonstrates high correlation coefficients, ranging from 0.941 to 0.962, outperforming other models such as ANN/ELM

(0.934–0.956/0.954), DBN (0.495–0.911), DNN (0.922–0.941), and MARS (0.928–0.935). RNN-LSTM and KNN exhibit comparable performance, with RNN-LSTM slightly outperforming KNN. Corrected SVM shows accurate predictions, with an RMSE of 11.35% and rRMSE of 1.713 MJ/m^2, while LM exhibits a correlation coefficient (R) of 0.9376 for training data and 0.9340 for testing data. Deep learning models, particularly RNN, outperform others, with an RMSE of 82.22 W/m^2 and an MAE of 36.52 W/m^2. Additionally, other models, such as SMGRT, GRU, LightGBM, RF, SVM, MLP, GPR, XGBoost, GHI, ANFIS, SVR, and Naive Bayes, show varying levels of performance across different evaluation metrics. While the accuracy of the ANN and ARIMA approaches is nearly equal, ANN has the advantage of being more flexible. Merged and generalized models surpassed conventional empirical models. Additionally, combined models tended to be more accurate than generalized models, using the same input parameters. Single-stochastic algorithm methods, such as ANN and ARIMA, are progressively becoming less relevant. Because the accuracy of the predictions is contingent on the integrity of the training data, choosing the best input is critical for FS. Through removing unnecessary or redundant information and retaining only the most important features, FS reduces computing costs and solves overfitting problems. It also aids in multicollinearity problems. There are three FS methods: filter, wrapped, and embedded [15,62]. Overall, this comparative analysis of the models highlights their strengths and weaknesses, providing valuable insights for understanding their capabilities in analyzing the given data, as well as for future model selection and application.

3. Our Proposed Approach

In this section, we will introduce the proposed model, as depicted in Figure 2, and an overview of the methodology used in this study.

We begin by identifying the data requirements and collecting the data. We analyze the information for both quantity and quality. Second, we standardize the data from different formats, correct errors, and expand it, adding more dimensions if necessary. We reduce noise and ambiguities, sample from large databases, select attributes that identify the most significant attributes, and reduce dimensions by implementing various strategies (feature engineering). As we possess a model based on time series, the initial step in transforming the data involves framing it as a supervised learning problem using the sliding window technique.

The steps for processing in our study are described as follows and shown in the flowchart in Figure 2:

1. Import records from CSV files;
2. To ascertain the crucial correlation among all the features employed in the training process, we first need to determine the number of features in the training set;
3. After determining the number of features in the previous step, the Recursive Feature Elimination (RFE) technique is applied to identify the features from the CSV files that exhibit the strongest correlation;
4. In order to divide the dataset into distinct folds for training and testing, we must indicate the number of folds (in this case, ten folds are chosen);
5. Divide the dataset into numerous folds, with one allocated for testing and the remainder for training, through k-fold cross-validation;
6. Train the data using various algorithms (RF, DT, LR, Pr, GBM) to then train the model using the created training dataset. It is then used to test the rest of the dataset compared to the randomly selected feature;
7. In the next step, the trained model is applied to the test dataset, and various metrics are computed in order to assess the accuracy and efficacy of the model;
8. After calculating the scoring metrics, the final results are displayed and graphed.

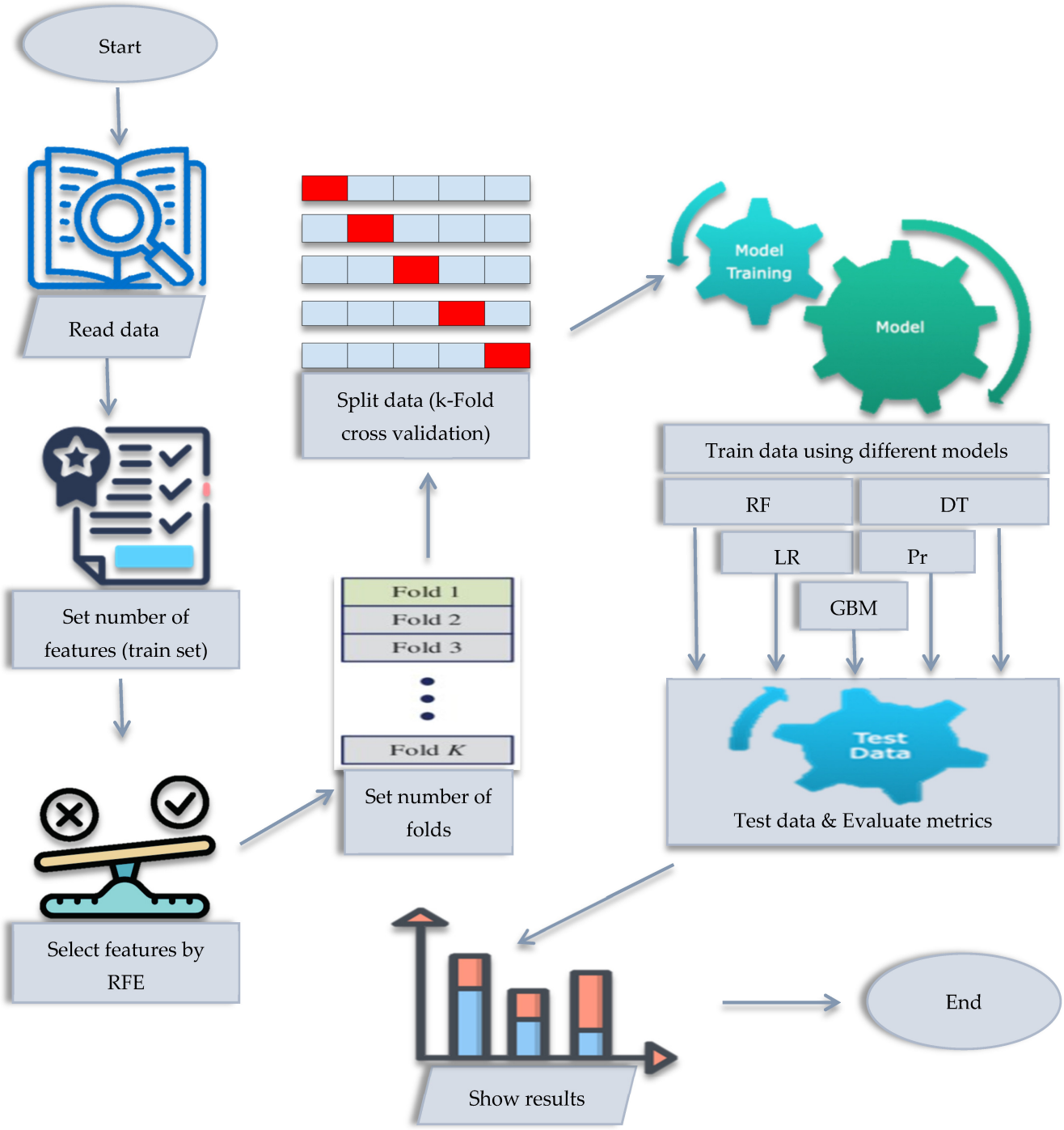

Figure 2. Our proposed model.

4. Experimental Study and Results

The proposed ML approach focuses on identifying the most relevant features that contribute significantly to solar radiation estimation. Through iteratively eliminating less informative features, the RFE algorithm helps in building a more accurate predictive model. Through utilizing only the most influential features, the model can better capture the underlying patterns and relationships in the data, leading to improved accuracy compared to traditional methods that may consider all features. It helps in mitigating the risk of overfitting, which occurs when a model performs well on the training data but fails to generalize to new, unseen data. Through eliminating less informative features, RFE prevents the model from being overly complex and overly sensitive to noise in the

training data. This reduction in overfitting enhances the model's ability to provide accurate estimations on unseen data, improving its overall reliability. The reduction in the number of features not only simplifies the modeling process, but also improves computational efficiency. With a smaller feature space, the model requires fewer computational resources and less training time, making it more efficient compared to other methods that consider all available features. This provides transparency and interpretability in the Rs estimation process. In identifying the subset of features that contribute most to the prediction, it helps in understanding the key factors driving the solar radiation patterns. This interpretability allows domain experts to gain insights and make informed decisions based on the selected features. These advantages make it a promising approach when compared to existing methods, leading to more accurate and efficient predictions of Rs.

The conversion from time series to lag values, as employed in our research, offers several advantages over existing methods. It helps capture temporal dependencies and patterns in the data, incorporating past observations as predictors. This enables the model to better capture the historical context and temporal dynamics of the data, leading to improved prediction accuracy compared to methods that do not consider lag values. Additionally, the conversion to lag values facilitates feature engineering through creating additional informative variables. By including lagged features as predictors, we provide the model with a richer set of inputs, capturing the relationship between current and past observations and enabling more accurate predictions. Furthermore, the conversion to lag values can enhance computational efficiency through reducing the dimensionality of the data. Transforming the time series into a matrix of lagged features streamlines the modeling process and reduces computational complexity, leading to improved efficiency compared to methods that consider the entire time series. This approach also provides flexibility in handling various time series data. It allows for the incorporation of different lag lengths or time intervals, enabling the model to capture different time dependencies present in the data. This adaptability enables us to tailor the lagged feature representation to different temporal patterns and achieve more accurate predictions for specific time series datasets. Moreover, the conversion to lag values enhances the interpretability of the model. By analyzing the importance and influence of past observations on current predictions, we gain insights into the temporal dynamics and relationships within the time series. This interpretability enables better understanding and interpretation of the model's predictions.

The model parameters are carefully selected and tuned in order to achieve the best performance in estimating solar radiation. We employ a systematic approach where we define a list of hyperparameter values and test them one by one. The model is trained and evaluated using the chosen evaluation metric for each hyperparameter value. The hyperparameter configuration that yields the best performance metric is selected as the optimal choice. This iterative process of parameter tuning and evaluation allows us to optimize the model and ensure accurate and reliable predictions. We place great importance on this step, in order to enhance the overall effectiveness of our developed model.

4.1. Environment Description

For this ongoing investigation, we performed examinations utilizing data compiled over a period of 36 years (1979–2014), sourced from The National Centers for Environmental Prediction. This dataset encompasses 12,988 entries of daily humidity, minimum and maximum temperatures, longitude, latitude, elevation, wind, precipitation, and sunlight. Our investigation was conducted and completed on a portable computer, provided with a Core-i5 3437U CPU (2.4 GHz)—DDR3 memory capacity of 16 GB; the operating system employed during the experiment was Windows 10 Professional (64-bit), and the model was trained using Python version 3.9.7. In order to evaluate it, k-fold was repeated three times, with ten folds across all repetitions; cross-validation, a statistical technique, employs limited samples for resampling purposes. The k-fold cross-validation procedure blindly divides the dataset into k non-overlapping folds, as shown in Figure 3. The held-back test set ensures that each fold is used only once. At the same time, the remaining sections are

continuously merged to compose the training set. Performance metrics are computed and preserved on the test set. At the same time, the remaining folds are repeatedly joined to form the training set. Performance metrics are calculated and saved on the test set.

Figure 3. K-fold cross-validation method.

Performing the method several times for the designated number of folds is essential in k-fold cross-validation. The average performance metrics are provided after fitting and evaluating k models on the corresponding hold-out test sets. This technique offers the benefit of minimizing common mistakes and enhancing the anticipated model performance.

The MAE is a metric that quantifies the difference in inaccuracies between two instances of an identical event taking place. It compares the anticipated outcome with the observed outcome, denoted as X vs. Y, where the values of X and Y are identical. The computation of MAE is as follows:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \quad (1)$$

It computes the average absolute deviation observed between anticipated and target values. An upward pattern implies a reduction in the measure. The feedback data substantially vary from the training data. A decreasing trend indicates an increase in the metric. This indicates that the model's training is efficient. The importance of the feature assigns a rating to each feature in the dataset according to its significance. The ratings explain the "importance" of each feature. A better score signifies that the characteristic holds greater significance and will exert a more potent influence on the model. There are multiple methods with which to calculate the importance of features. This write-up presents an outline of the Gini importance technique utilized in Scikit-learn for evaluating the impurity of nodes. The weight of a node is determined using the proportion of samples that arrive at it relative to the total number of samples. The reduction in the impurity of a node is known as feature importance, which is equivalent to the probability of the node. When a decision tree has two child nodes, the formula is as follows:

$$\mathrm{ni_j} = \mathrm{w_j C_j} - \mathrm{w_{left(j)} C_{left(j)}} - \mathrm{w_{right(j)} C_{right(j)}} \quad (2)$$

where:

ni_j represents the importance of node j, w_j is the weighted count of samples reaching node j, C_j measures the impurity measure at node j, left(j) signifies the left child node of node j, and right(j) the right child node of node j.

Formula (2) evaluates the feature importance for every DT through considering the significance of the node *j* (*nj*). An exclusive attribute can be utilized in every branch of the tree. Hence, we assess the importance of features by making use of Equation (3).

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \tag{3}$$

where:

fi_i: Importance of feature i

ni_j: Importance of node j

Through dividing each feature's importance by the total importance value, these values can be standardized to a numerical range that falls between 0 and 1. This can be achieved by utilizing Equation (4).

$$norm\ fi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \tag{4}$$

The Min–Max normalization technique is employed in order to avoid the negative impact of heavy weights. This method is a linear approach that maintains the associations among the initial data points. It recognizes the smallest and largest values of characteristic X as (min_X) and (max_X). The process involves calculating the value v'i of X within the range [new_min_X, new_max_X] through transforming the value vi using Equation (5). Equation (3) demonstrates the normalization formula utilized for range transformation.

$$v'_i = \frac{v_i - min_X}{max_X - min_X}(new\_max_X - new\_min_X) + new\_min_X \tag{5}$$

If the normalization of a following input instance goes beyond the range of the primary data for *X*, an "out-of-bounds" error is indicated [63]. The functioning of RFE involves developing prediction models, evaluating features, eliminating those with minor significance, and repeating this process until the desired number of features is attained. RFE is a wrapper-based FS method that uses a filter-based FS method within its internal process. Fundamentally, it utilizes unique ML algorithms. RFE encompasses and applies these algorithms in order to aid in the feature selection process. The FS method based on filters evaluates each characteristic and selects those with the most elevated (or lowest) ranking.

RFE employs a technique to choose a subset of features from the training set through eliminating irrelevant features until the optimum number of features is obtained. This involves the following steps [64]:

- Training the ML algorithm implemented in the heart of the model;
- Ranking the features based on their importance;
- Eliminating the insignificant features and providing the model with further training;
- Continuing the procedure until the intended quantity of features is chosen;
- Creating a metric of importance for variables that sorts the predictors according to their relevance once the entire model has been built;
- In every cycle, the model is reconstructed after eliminating the least significant predictors.

4.2. Discussion of Results

As illustrated in the chart below, we are dealing with an MVTS model, where we can observe periodicity in each parameter. A collection of datasets where two or more variables are observed each time refers to an MVTS. While most time series analysis techniques focus on univariate data, which is simpler to comprehend and handle, MVTS analysis, on the

other hand, is often more challenging to manipulate and model. It concurrently deals with multiple time series, typically more intricate than univariate analysis.

The initial data preprocessing phase involves understanding each column's data type and identifying missing values, duplicates, and errors. Several preprocessing steps can be taken. Firstly, the missing values need to be identified and handled. Secondly, duplicates should be identified and removed. Thirdly, it is crucial to identify and manage errors, such as outliers, using statistical methods, and decide whether to replace or remove them. Fourthly, data should be stored in the appropriate data type, and data types should be converted if necessary. Fifthly, columns should be renamed with meaningful names to facilitate analysis. Sixthly, irrelevant columns that are not necessary for analysis should be removed. Finally, numeric variables are selected and standardized using the Standard Scaler method from Scikit-learn. The selected columns include MaxTemperature, MinTemperature, Precipitation, Wind, RelativeHumidity, and Solar. The appropriate transform method fits the scaler to the data and transforms the numeric variables, ensuring the data are ready for analysis.

Standardization is a data preprocessing technique that transforms data into a standard format in order to allow for a fairer comparison between them. It is helpful for many ML techniques, as it can improve model performance and reduce biases introduced by variables with different scales. This technique involves centering the data around zero and scaling them to the same range. Specifically, for each variable, Standardization includes subtracting the mean of the variable from each observation and dividing the outcome by the variable's standard deviation. This process transforms the variable into a distribution centered on 0 with a variance of 1 [65]. We used the Standard Scaler method from the Sklearn preprocessing module to perform Standardization.

Seasonal adjustment is a common technique used in time series analysis in order to remove the effects of seasonal patterns from a time series dataset. Seasonal patterns are recurring patterns within a fixed period, such as a month, a quarter, or a year. By performing a seasonal difference on the time series data, we can eliminate the seasonal pattern and focus on the data's underlying trends and irregular components. In this case, the code uses a lag of 12 months or one year to perform the seasonal difference, subtracting each value from the value 12 months prior. This will help to remove any recurring patterns that occur yearly. Thereafter, we trim off the first year of empty data (since the first 12 months of differenced data will be NaN) and save the differenced dataset. After performing the seasonal difference, the differenced data for the variables from July 2013 to July 2014 is plotted in the line graphs below (Figure 4). This allows us to inspect the data and see the seasonal adjustment visually. Overall, seasonal adjustment is an essential step in time series analysis, as it removes the effects of seasonal patterns from the time series data. This enables us to gain a deeper insight into the fundamental trends and patterns within the data, thereby enhancing the precision of our forecasts and predictions.

As we have an MVTS problem, transforming data involves converting a time series, which follows a chronological order, into a supervised learning task that includes input and output patterns (X, Y) using the sliding window method. This allows an algorithm to understand how to anticipate the output based on the input patterns. The sliding window technique involves utilizing the preceding to anticipate the succeeding time steps. It is sometimes referred to as the window method in specific texts. In statistics, it is known as a lag or lagging method. It proves to be beneficial in decreasing the time complexity of particular issues. The approach applies to solving almost any problem that satisfies the condition of being capable of adding items consecutively or simultaneously into a single variable. The sliding window strategy is adaptable for both univariate and MVTS analysis (Figure 5).

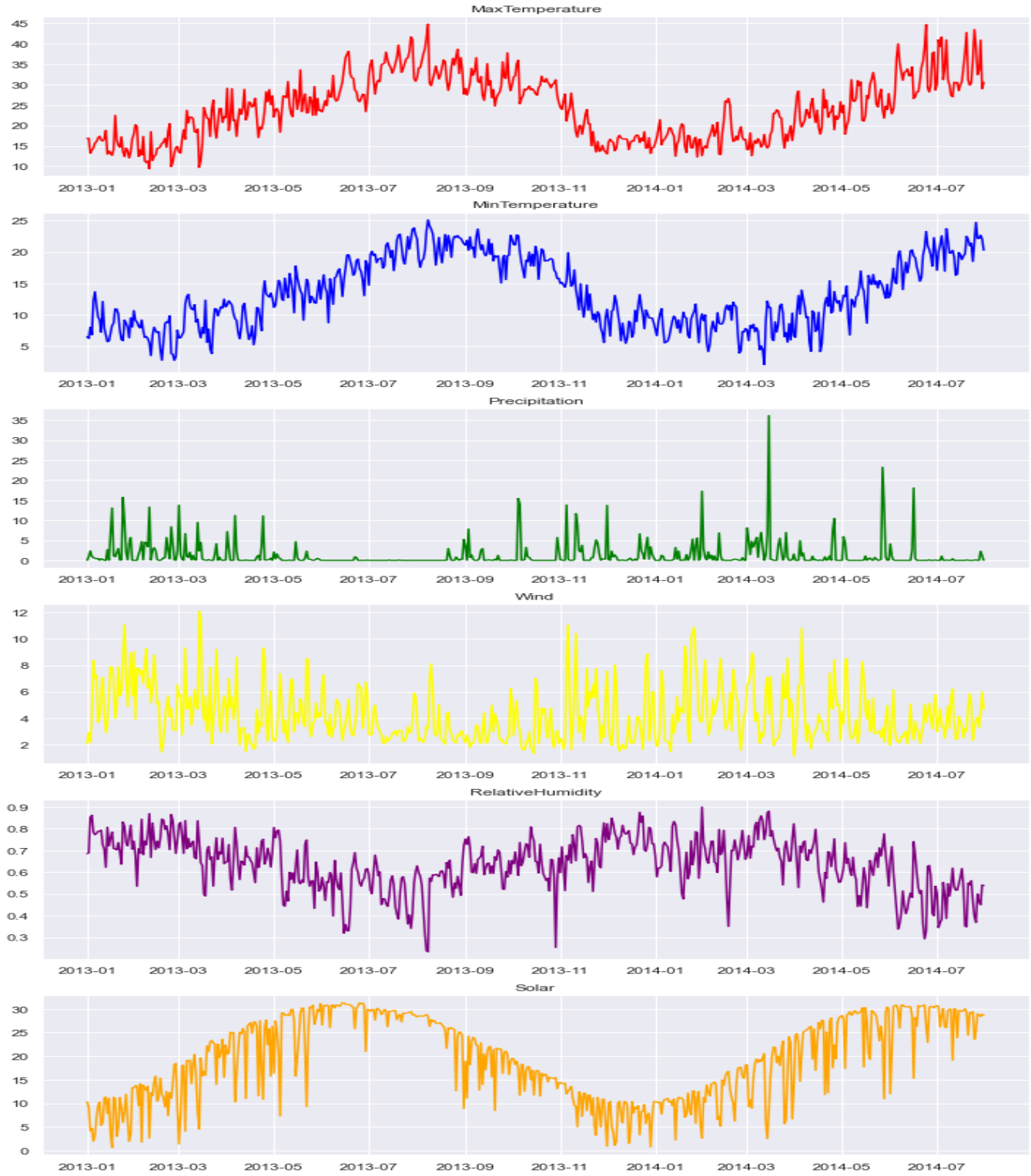

Figure 4. Graphical representation of environmental variables.

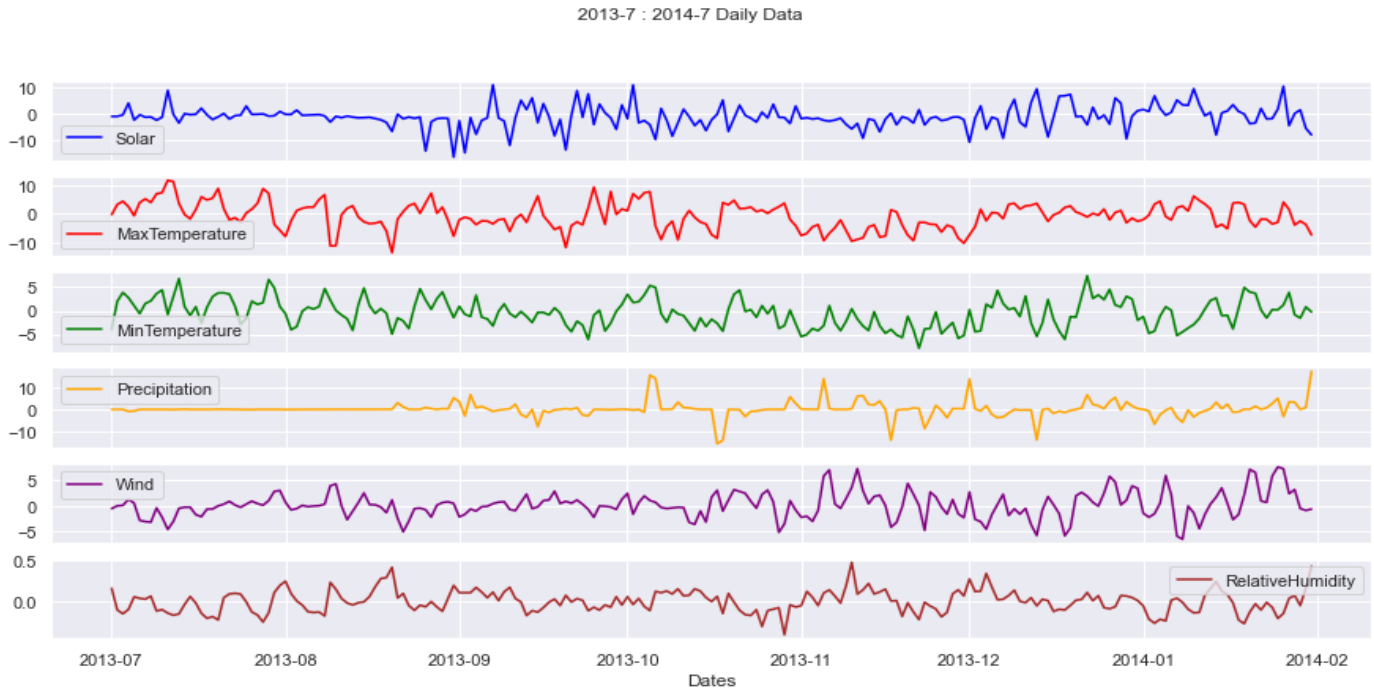

Figure 5. The differenced data for the different variables from July 2013 to July 2014.

Feature selection is an important process in preparing data. In this case, feature selection is performed using the RF Regressor algorithm. The selected features are as follows:

- MaxTemperature of months 12, 10, 4, and 1;
- MinTemperature of months 12, 11, 10, 6, 4, 3, 2, and 1;
- Precipitation of months 12, 11, 10, 7, and 1;
- Wind speed of months 12, 11, 7, 6, and 1;
- Relative humidity of months 11 and 3.

These features were selected using the RF Regressor algorithm, to build a model that uses a set of DTs to predict continuous values. The algorithm analyzes feature importance to evaluate the effect of each feature on the target variable and select the most critical features for prediction. Feature importance ranking using RF is shown in Figure 6:

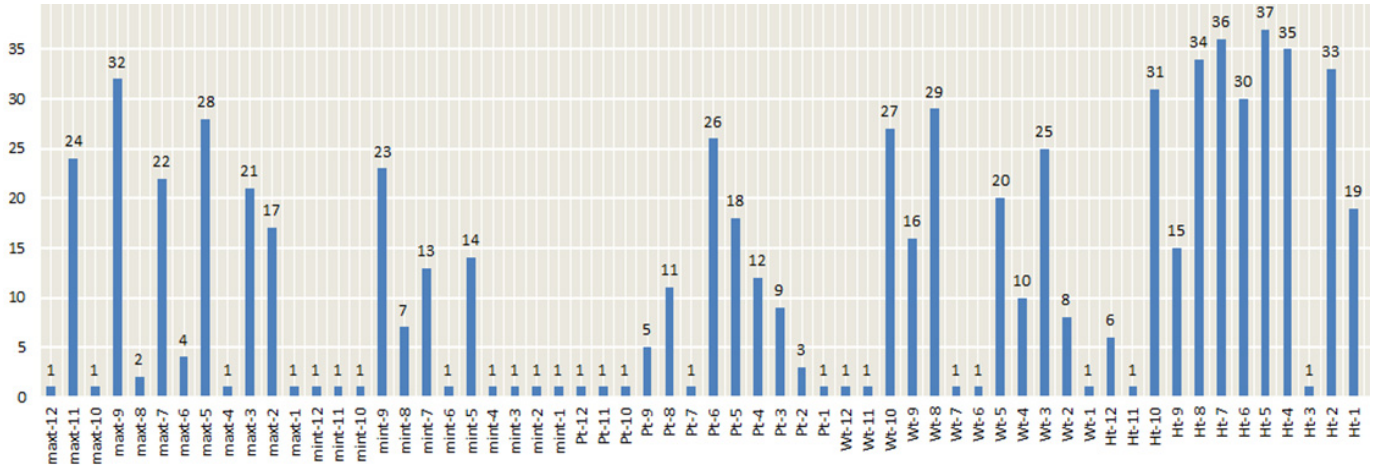

Figure 6. Feature importance ranking using RF Regressor.

The DT algorithm is an ML model that builds a tree-like structure in order to classify or forecast a target variable based on input features. The algorithm recursively splits the data based on the feature that results in the highest information gain, which measures the

reduction in entropy after the split. The features with the highest information gain are considered the most important for prediction. In this case, the DT algorithm was used for FS, and the picked features are:

- MaxTemperature of months 12, 10, 4, and 1;
- MinTemperature of months 12, 10, 6, 5, 4, 3, 2, and 1;
- Precipitation of months 12, 9, 8, 6, and 1;
- Wind speed of months 12, 9, 6, and 1;
- Relative humidity of months 11, 9, 5, and 2.

It is interesting to note that some features picked with the DT algorithm differ from those selected with the RF Regressor algorithm. In conclusion, the DT algorithm selected essential features for predicting solar radiation based on their information gain. However, the selected features may differ from those selected according to the algorithm used in the core of RFE, and it is important to compare and assess the performance of different techniques and FS methods.

In order to compare the lr, RF, DT, Pr, and GBM performance models, a box and whisker plot is presented for the RMSE and R2 evaluation metrics (Figure 7).

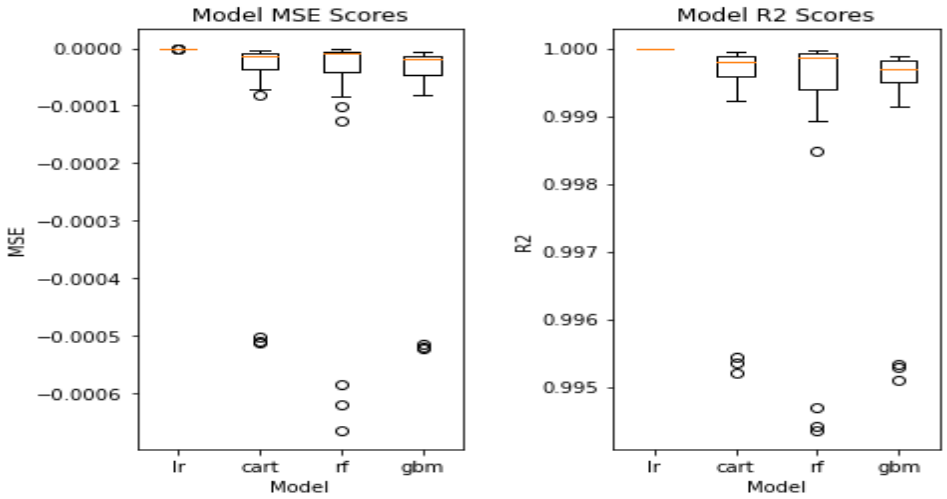

Figure 7. Lr, RF, DT, Pr and GBM performance models for MSE and R2 evaluation metrics.

Based on the RMSE and R2 scores, the LR model appears to have the most outstanding performance, of 0.003 (0.002). Nevertheless, the other models also show impressive performance, with RMSE scores ranging from 0.006 to 0.007 and consistent R2 scores of 0.999.

5. Conclusions

FS is a critical phase in preparing data for ML models; because selecting irrelevant or redundant features can lead to overfitting or poor model performance, choosing the right features is critical for building accurate and robust models.

Recursive Feature Elimination is used for FS. It recursively removes features from the dataset and constructs a model using the remaining features until the desired number of features is reached.

The approach used in our research work offers several advantages: It improves the model interpretability and enhances model performance. It focuses on the most informative features, leading to more accurate predictions, and takes into account feature interactions and dependencies. It can handle multicollinearity issues through iteratively eliminating redundant features, ensuring that the final feature set is independent and representative.

RFE offers flexibility in algorithm selection; it is a versatile technique that can be used with different ML algorithms. It is not limited to a specific algorithm and can be applied with various models, such as linear regression, support vector machines, or random forests.

It also automates the FS process. It provides a ranking of the importance of each feature, allowing researchers to gain insights into the relative significance of different variables in the model's performance. In this case, RFE is used in conjunction with the different algorithms to identify the most essential features for prediction. The advantage of using RFE is that it considers the interaction between features rather than simply evaluating each feature in isolation.

The RF Regressor algorithm analyzes feature importance in order to evaluate the effect of each feature on the target variable and select the most critical features for prediction. On the other hand, the DT algorithm is a model that builds a tree-like structure in order to classify or estimate a target variable derived from a set of input features. The features with the highest information gain are considered the most important for prediction. As we can see from the results, the two algorithms selected different sets of features, which may be due to the different methods used for evaluating feature importance. However, we have observed common patterns in the FS across different models, indicating their significance in accurately estimating solar radiation. Some of the key features consistently identified as important for accurate Rs estimation include:

- MaxTemperature of months 12, 10, 4, and 1: The maximum temperature during these months likely captures seasonal variations and their impact on solar radiation levels;
- MinTemperature of months 12, 10, 6, 4, 3, 2, and 1: The minimum temperature during these months provides insights into the daily temperature range, which can influence solar radiation patterns;
- Precipitation of months 12 and 1: The amount of precipitation during these months may affect cloud cover and atmospheric conditions, impacting solar radiation levels;
- Wind speed of months 12, 6, and 1: The wind speed during these months is an indicator of atmospheric dynamics, which can influence the dispersion of clouds and affect solar radiation availability;
- Relative humidity of month 11: The relative humidity in month 11 likely represents a critical period for moisture content in the air, which can affect solar radiation absorption and scattering.

These features highlight the importance of considering meteorological factors, such as temperature, precipitation, wind speed, and relative humidity, in accurately estimating solar radiation. By incorporating these influential features, our approach captures the relevant environmental dynamics and improves the precision of solar radiation estimation.

In order to assess and contrast the effectiveness of various models, the RMSE and R2 evaluation metrics were used, and a box and whisker plot was created to visualize the results. LR had the top RMSE and R2 scores (0.003, 0.002), followed closely by other models, such as RF, DT, Pr, and GBM (RMSE scores ranging from 0.006 to 0.007 and consistent R2 scores of 0.999). This suggests that an ensemble of regression models can help improve the accuracy of predictions for complex problems. Compared to other research, Sivanandam and Deepa discovered that an ensemble of regression models, including linear regression, random forest, and Gradient Boosting, outperformed individual models in predicting housing prices [66]. Additionally, Kumar and Singh compared machine learning models for predicting stock prices and found that the Gradient Boosting and random forest models outperformed other models [67].

The novelty of the proposed approach lies in several aspects. Firstly, it introduces a novel framework that integrates various models, offers insights into previously unexplored aspects, and challenges existing theories. It emphasizes the importance of feature selection and model evaluation within the context of RFE, shedding light on the factors influencing feature rankings and prediction performance. This contributes to a deeper understanding of the underlying mechanisms of feature selection. Secondly, the study provides insights into the suitability of LR, RF, DT, CART, and GBM models for MVTS analysis, which expands the knowledge base for solving complex problems in this domain. Lastly, by showcasing the effectiveness of RFE as a feature selection technique, the research offers a practical approach to enhancing the performance of predictive models in complex problem domains.

In conclusion, this study presents novel insights, contributes to existing knowledge in feature selection and model evaluation, and provides a practical approach to addressing challenges in MVTS analysis.

Our approach holds significant potential for real-world applications in various areas related to solar energy systems. It can assist in the planning and design of solar energy systems. Thus, the developers can optimize the placement and capacity of solar panels, maximizing energy production and ensuring optimal system efficiency. This leads to more cost-effective and sustainable solar energy installations. It can contribute to providing accurate solar radiation forecasts at different spatial and temporal resolutions. This information allows grid operators to balance the fluctuating solar energy supply with demand, optimize grid stability, and reduce reliance on conventional energy sources. Understanding solar radiation patterns and variability is essential for environmental impact studies. Reliable solar radiation estimation enables both researchers and policymakers to assess the environmental effects of solar energy systems, analyze their impact on ecosystems, and develop mitigation strategies.

Before widely adopting our approach, it is important to consider its limitations and potential drawbacks. These include the requirement for accurate and comprehensive input data, sensitivity to the chosen model, potential limitations in generalization to different geographical locations and climate conditions, the dynamic nature of solar radiation that may not be fully captured, the trade-off between interpretability and accuracy, the need for sufficient computational resources, and the necessity for validation and benchmarking against existing methods. Addressing these limitations will contribute to the robustness and suitability of our approach for real-world applications in solar radiation estimation.

To further enhance the validity and applicability of our findings, future research should consider additional evaluation metrics, explore alternative feature selection techniques, and investigate the generalizability of our approach to different datasets and problem contexts. By continuing to advance the feature selection and model evaluation field, we can improve the accuracy and robustness of ML models in solving real-world challenges.

Author Contributions: H.H. is the main author, who manages the contribution and gives the detailed description of the research work. S.B. writes the abstract and introduction, and analyzes the related works section. A.G. evaluates the results obtained from implementation and illustrates the figures. M.A. and A.B.-H. participate in the implementation of the model, prepare the final manuscript, and correct the English language. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, L.; Kang, J.; Wan, M.; Fang, L.; Zhang, C.; Zeng, Z. Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events. *Front. Earth Sci.* **2021**, *9*, 596860. [CrossRef]
2. Rohani, A.; Taki, M.; Abdollahpour, M. A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I). *Renew. Energy* **2018**, *115*, 411–422. [CrossRef]
3. Zhang, Y.; Cui, N.; Feng, Y.; Gong, D.; Hu, X. Comparison of BP, PSO-BP and statistical models for predicting daily global solar radiation in arid Northwest China. *Comput. Electron. Agric.* **2019**, *164*, 104905. [CrossRef]
4. Ağbulut, Ü.; Gürel, A.E.; Biçen, Y. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renew. Sustain. Energy Rev.* **2021**, *135*, 110114. [CrossRef]
5. Chaibi, M.; Benghoulam, E.M.; Tarik, L.; Berrada, M.; Hmaidi, A.E. An Interpretable Machine Learning Model for Daily Global Solar Radiation Prediction. *Energies* **2021**, *21*, 7367. [CrossRef]
6. Boutahir, M.K.; Farhaoui, Y.; Azrour, M.; Zeroual, I.; El Allaoui, A. Effect of feature selection on the prediction of direct normal irradiance. *Big Data Min. Anal.* **2022**, *5*, 309–317. [CrossRef]

7. Piri, J.; Shamshirband, S.; Petković, D.; Tong, C.W.; ur Rehman, M.H. Prediction of the solar radiation on the Earth using support vector regression technique. *Infrared Phys. Technol.* **2015**, *68*, 179–185. [CrossRef]
8. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.-L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [CrossRef]
9. Rigollier, C.; Lefèvre, M.; Wald, L. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. *Sol. Energy* **2004**, *77*, 159–169. [CrossRef]
10. Huertas-Tato, J.; Aler, R.; Galván, I.M.; Rodríguez-Benítez, F.J.; Arbizu-Barrena, C.; Pozo-Vázquez, D. A short-term solar radiation forecasting system for the Iberian Peninsula. Part 2: Model blending approaches based on machine learning. *Sol. Energy* **2020**, *195*, 685–696. [CrossRef]
11. Shadab, A.; Said, S.; Ahmad, S. Box–Jenkins multiplicative ARIMA modeling for prediction of solar radiation: A case study. *Int. J. Energy Water Res.* **2019**, *3*, 305–318. [CrossRef]
12. Alsharif, M.; Younes, M.; Kim, J. Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea. *Symmetry* **2019**, *11*, 240. [CrossRef]
13. Hocaoğlu, F.O. Novel analytical hourly solar radiation models for photovoltaic based system sizing algorithms. *Energy Convers. Manag.* **2010**, *51*, 2921–2929. [CrossRef]
14. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *J. Clean. Prod.* **2019**, *216*, 288–310. [CrossRef]
15. Bouzgou, H.; Gueymard, C.A. Minimum redundancy—Maximum relevance with extreme learning machines for global solar radiation forecasting: Toward an optimized dimensionality reduction for solar time series. *Sol. Energy* **2017**, *158*, 595–609. [CrossRef]
16. Meenal, R.; Selvakumar, A.I. Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renew. Energy* **2018**, *121*, 324–343. [CrossRef]
17. Yadav, A.K.; Malik, H.; Chandel, S.S. Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India. *Renew. Sustain. Energy Rev.* **2015**, *52*, 1093–1106. [CrossRef]
18. Zhou, Y.; Liu, Y.; Wang, D.; Liu, X.; Wang, Y. A review on global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Convers. Manag.* **2021**, *235*, 113960. [CrossRef]
19. Diagne, M.; David, M.; Lauret, P.; Boland, J.; Schmutz, N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renew. Sustain. Energy Rev.* **2013**, *27*, 65–76. [CrossRef]
20. Yadav, A.K.; Chandel, S.S. Solar radiation prediction using Artificial Neural Network techniques: A review. *Renew. Sustain. Energy Rev.* **2014**, *33*, 772–781. [CrossRef]
21. Kumar, R.; Aggarwal, R.K.; Sharma, J.D. Comparison of regression and artificial neural network models for estimation of global solar radiations. *Renew. Sustain. Energy Rev.* **2015**, *52*, 1294–1299. [CrossRef]
22. Pang, Z.; Niu, F.; O'Neill, Z. Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons. *Renew. Energy* **2020**, *156*, 279–289. [CrossRef]
23. Chen, J.-L.; Liu, H.-B.; Wu, W.; Xie, D.-T. Estimation of monthly solar radiation from measured temperatures using support vector machines—A case study. *Renew. Energy* **2011**, *36*, 413–420. [CrossRef]
24. Guermoui, M.; Melgani, F.; Gairaa, K.; Mekhalfi, M.L. A comprehensive review of hybrid models for solar radiation forecasting. *J. Clean. Prod.* **2020**, *258*, 120357. [CrossRef]
25. Hedar, A.-R.; Almaraashi, M.; Abdel-Hakim, A.E.; Abdulrahim, M. Hybrid Machine Learning for Solar Radiation Prediction in Reduced Feature Spaces. *Energies* **2021**, *14*, 7970. [CrossRef]
26. Mohanty, S.; Patra, P.K.; Sahoo, S.S. Prediction and application of solar radiation with soft computing over traditional and conventional approach—A comprehensive review. *Renew. Sustain. Energy Rev.* **2016**, *56*, 778–796. [CrossRef]
27. Guermoui, M.; Abdelaziz, R.; Gairaa, K.; Djemoui, L.; Benkaciali, S. New temperature-based predicting model for global solar radiation using support vector regression. *Int. J. Ambient. Energy* **2022**, *43*, 1397–1407. [CrossRef]
28. Geetha, A.; Santhakumar, J.; Sundaram, K.M.; Usha, S.; Thentral, T.T.; Boopathi, C.S.; Ramya, R.; Sathyamurthy, R. Prediction of hourly solar radiation in Tamil Nadu using ANN model with different learning algorithms. *Energy Rep.* **2022**, *8*, 664–671. [CrossRef]
29. Olatomiwa, L.; Mekhilef, S.; Shamshirband, S.; Mohammadi, K.; Petković, D.; Sudheer, C. A support vector machine–firefly algorithm-based model for global solar radiation prediction. *Sol. Energy* **2015**, *115*, 632–644. [CrossRef]
30. Guezzaz, A.; Benkirane, S.; Azrour, M.; Khurram, S. A Reliable Network Intrusion Detection Approach Using Decision Tree with Enhanced Data Quality. *Secur. Commun. Netw.* **2021**, *2021*, 1230593. [CrossRef]
31. Guezzaz, A.; Azrour, M.; Benkirane, S.; Mohy-Eddine, M.; Attou, H.; Douiba, M. A Lightweight Hybrid Intrusion Detection Framework using Machine Learning for Edge-Based IIoT Security. *Int. Arab. J. Inf. Technol.* **2022**, *19*, 5. [CrossRef]
32. Goliatt, L.; Yaseen, Z.M. Development of a hybrid computational intelligent model for daily global solar radiation prediction. *Expert Syst. Appl.* **2023**, *212*, 118295. [CrossRef]
33. Ghimire, S.; Deo, R.C.; Casillas-Pérez, D.; Salcedo-Sanz, S. Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms. *Appl. Energy* **2022**, *316*, 119063. [CrossRef]

34. Etxegarai, G.; López, A.; Aginako, N.; Rodríguez, F. An analysis of different deep learning neural networks for intra-hour solar irradiation forecasting to compute solar photovoltaic generators' energy production. *Energy Sustain. Dev.* **2022**, *68*, 1–17. [CrossRef]
35. Bamisile, O.; Oluwasanmi, A.; Ejiyi, C.; Yimen, N.; Obiora, S.; Huang, Q. Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions. *Int. J. Energy Res.* **2022**, *46*, 10052–10073. [CrossRef]
36. Üstün, İ.; Üneş, F.; Mert, İ.; Karakuş, C. A comparative study of estimating solar radiation using machine learning approaches: DL, SMGRT, and ANFIS. *Energy Sources Part A Recovery Util. Environ. Eff.* **2022**, *44*, 10322–10345. [CrossRef]
37. Faisal, A.N.M.F.; Rahman, A.; Habib, M.T.M.; Siddique, A.H.; Hasan, M.; Khan, M.M. Neural networks based multivariate time series forecasting of solar radiation using meteorological data of different cities of Bangladesh. *Results Eng.* **2022**, *13*, 100365. [CrossRef]
38. Bounoua, Z.; Chahidi, L.O.; Mechaqrane, A. Estimation of daily global solar radiation using empirical and machine-learning methods: A case study of five Moroccan locations. *Sustain. Mater. Technol.* **2021**, *28*, e00261. [CrossRef]
39. Meenal, R.; Michael, P.A.; Pamela, D.; Rajasekaran, E. Weather prediction using random forest machine learning model. *Indones. J. Electr. Eng. Comput. Sci.* **2021**, *22*, 1208. [CrossRef]
40. Khelifi, R.; Guermoui, M.; Rabehi, A.; Lalmi, D. Multi-step-ahead forecasting of daily solar radiation components in the Saharan climate. *Int. J. Ambient. Energy* **2020**, *41*, 707–715. [CrossRef]
41. Kurniawan, A.; Shintaku, E. Estimation of the Monthly Global, Direct, and Diffuse Solar Radiation in Japan Using Artificial Neural Network. *Int. J. Mach. Learn. Comput.* **2020**, *10*, 253–258. [CrossRef]
42. Fan, J.; Wang, X.; Zhang, F.; Ma, X.; Wu, L. Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data. *J. Clean. Prod.* **2020**, *248*, 119264. [CrossRef]
43. Kwon, Y.; Kwasinski, A.; Kwasinski, A. Solar Irradiance Forecast Using Naïve Bayes Classifier Based on Publicly Available Weather Forecasting Variables. *Energies* **2019**, *12*, 1529. [CrossRef]
44. Srivastava, R.; Tiwari, A.N.; Giri, V.K. Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. *Heliyon* **2019**, *5*, e02692. [CrossRef]
45. Basaran, K.; Özçift, A.; Kılınç, D. A New Approach for Prediction of Solar Radiation with Using Ensemble Learning Algorithm. *Arab. J. Sci. Eng.* **2019**, *44*, 7159–7171. [CrossRef]
46. Cornejo-Bueno, L.; Casanova-Mateo, C.; Sanz-Justo, J.; Salcedo-Sanz, S. Machine learning regressors for solar radiation estimation from satellite data. *Sol. Energy* **2019**, *183*, 768–775. [CrossRef]
47. Benali, L.; Notton, G.; Fouilloy, A.; Voyant, C.; Dizene, R. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renew. Energy* **2019**, *132*, 871–884. [CrossRef]
48. Bhola, P.; Bhardwaj, S. Estimation of solar radiation using support vector regression. *J. Inf. Optim. Sci.* **2019**, *40*, 339–350. [CrossRef]
49. Marzouq, M.; Bounoua, Z.; El Fadili, H.; Mechaqrane, A.; Zenkouar, K.; Lakhliai, Z. New daily global solar irradiation estimation model based on automatic selection of input parameters using evolutionary artificial neural networks. *J. Clean. Prod.* **2019**, *209*, 1105–1118. [CrossRef]
50. Guermoui, M.; Gairaa, K.; Rabehi, A.; Djafer, D.; Benkaciali, S. Estimation of the daily global solar radiation based on the Gaussian process regression methodology in the Saharan climate. *Eur. Phys. J. Plus* **2018**, *133*, 211. [CrossRef]
51. Yıldırım, H.B.; Çelik, Ö.; Teke, A.; Barutçu, B. Estimating daily Global solar radiation with graphical user interface in Eastern Mediterranean region of Turkey. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1528–1537. [CrossRef]
52. Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F.; Yu, X.; Lu, X.; Xiang, Y. 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 Convers. Manag.* **2018**, *164*, 102–111. [CrossRef]
53. Quej, V.H.; Almorox, J.; Arnaldo, J.A.; Saito, L. ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *J. Atmos. Sol.-Terr. Phys.* **2017**, *155*, 62–70. [CrossRef]
54. Marzo, A.; Trigo-Gonzalez, M.; Alonso-Montesinos, J.; Martínez-Durbán, M.; López, G.; Ferrada, P.; Fuentealba, E.; Cortés, M.; Batlles, F.J. Daily global solar radiation estimation in desert areas using daily extreme temperatures and extraterrestrial radiation. *Renew. Energy* **2017**, *113*, 303–311. [CrossRef]
55. Hassan, M.A.; Khalil, A.; Kaseb, S.; Kassem, M.A. Potential of four different machine-learning algorithms in modeling daily global solar radiation. *Renew. Energy* **2017**, *111*, 52–62. [CrossRef]
56. Alsina, E.F.; Bortolini, M.; Gamberi, M.; Regattieri, A. Artificial neural network optimisation for monthly average daily global solar radiation prediction. *Energy Convers. Manag.* **2016**, *120*, 320–329. [CrossRef]
57. Wang, L.; Kisi, O.; Zounemat-Kermani, M.; Salazar, G.A.; Zhu, Z.; Gong, W. Solar radiation prediction using different techniques: Model evaluation and comparison. *Renew. Sustain. Energy Rev.* **2016**, *61*, 384–397. [CrossRef]
58. Mehdizadeh, S.; Behmanesh, J.; Khalili, K. Comparison of artificial intelligence methods and empirical equations to estimate daily solar radiation. *J. Atmos. Sol.-Terr. Phys.* **2016**, *146*, 215–227. [CrossRef]
59. De Felice, M.; Petitta, M.; Ruti, P.M. Short-term predictability of photovoltaic production over Italy. *Renew. Energy* **2015**, *80*, 197–204. [CrossRef]

60. Pedro, H.T.C.; Coimbra, C.F.M. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. *Renew. Energy* **2015**, *80*, 770–782. [CrossRef]
61. Lazzaroni, M.; Ferrari, S.; Piuri, V.; Salman, A.; Cristaldi, L.; Faifer, M. Models for solar radiation prediction based on different measurement sites. *Measurement* **2015**, *63*, 346–363. [CrossRef]
62. Demirhan, H. The problem of multicollinearity in horizontal solar radiation estimation models and a new model for Turkey. *Energy Convers. Manag.* **2014**, *84*, 334–345. [CrossRef]
63. Al Shalabi, L.; Shaaban, Z. Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix. In Proceedings of the 2006 International Conference on Dependability of Computer Systems, Szklarska Poreba, Poland, 25–27 May 2006; pp. 207–214. [CrossRef]
64. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013.
65. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
66. Sivanandam, S.N.; Deepa, S.N. Hybrid models using support vector regression for stock price prediction. *J. Appl. Res. Technol.* **2014**, *12*, 205–214.
67. Singh, S.; Madan, T.K.; Kumar, J.; Singh, A.K. Stock Market Forecasting using Machine Learning: Today and Tomorrow. In Proceedings of the 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, India, 5–6 July 2019; pp. 738–745. [CrossRef]