International Journal of Intelligent Networks 4 (2023) 90–102

Contents lists available at ScienceDirect

International Journal of Intelligent Networks

journal homepage: www.keaipublishing.com/en/journals/international-journal-of-intelligent-networks

Solar irradiance forecasting models using machine learning techniques and digital twin: A case study with comparison

Neha Sehrawat [a], Sahil Vashisht [a,*], Amritpal Singh [b]

[a] Department of Computer Science, Shree Guru Gobind Singh Tricentenary University, India
[b] Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India

ARTICLE INFO

Index Terms:
Solar irradiance
Machine learning
Ensemble model
Photovoltaic
Linear regression
Digital twin
Renewable energy
Solar energy
Power generation
Stacking

ABSTRACT

The ever-increasing demand for energy and power consumption due to population growth, economic expansion, and evolving consumer choices has led to the need for renewable energy sources. Traditional energy sources such as coal, oil, and gas have contributed to global pollution and have adverse effects on human health. As a result, the use of renewable energy for power generation has increased tremendously. One such area of research is solar irradiation prediction, which utilizes Artificial Intelligence and Machine Learning techniques. With the use of real-time predicted data, the digital twins are intended to add value to the organization by identifying and preventing problems, predicting performance, and improving operations. This paper provides an overview of various learning methods used for predicting irradiance and presents a new ensemble solar irradiance forecasting model that combines eight machine learning models to ensure model diversity. The model's most critical factors for predicting irradiance include temperature, cloudiness index, relative humidity, and day of the week. To conduct a comprehensive analysis, the proposed 8-Stacking Regression Cross Validation (8 STR-CV) model was tested using data from three different climatic zones in India. The model's high accuracy scores of 98.8% for Visakhapatnam, 98% for Nagpur, and 97.8% for the mountainous region make it a valuable tool for future prediction in various sectors, including power generation and utilization planning.

1. Introduction

The production of energy is essential to a nation's ability to prosper economically. Compared to other production elements, the importance of energy is well acknowledged. Generally, the production of electrical energy in developing nations such as India majorly depends upon fossil fuels. The surge in fossil fuels prices which is further ignition by Ukraine and Russia war leads to a change in dependence to use of alternate source of energy. As per a report of Energy and power department, Government of India in 2014, 74% energy produce using fossil fuels as compared to the May 2022 where only 58% of energy is produce using fossil fuel. The surge in fossil fuels prices which is further ignition by Ukraine and Russia war leads to change on dependence to use of alternation source of energy. This trend shows that electricity generation from the various alternative renewable energy sources like solar, wind and hydro has been increased in last few years. This paradigm shift helps in one's economy to grow, regional self-sufficiency, environment protection, and emission reduction [1] (see Tables 1 and 2).

The unpredictable nature of solar and wind energy has resulted in difficult integration issues and dependability issues requiring expensive, technically complicated solutions. Additionally, the resources availability and production of electricity by using solar and wind resources is dependent heavily on regional climatic factors such the humidity, pressure, wind, and temperature of the atmosphere. Energy output fluctuation and unpredictability affect economic gains. Economic criteria like the expense of energy, and the out of the return, are difficult to determine when there is generational uncertainty. Additionally, considering PV farms which are connected to a grid, the power generation uncertainty might influence grid stability. In numerous countries, solar energy generation has a notable capacity in comparison to the wind energy, which helps to encourage the wide scale deployments of grid setups. This paper significantly focus on various machine learning models for prediction solar irradiance and how the ensembling of individual models results in optimal performance increase in the prediction results. There are various approaches that have been used to cope with the aforementioned challenges, but the majority of them are either expensive, like batteries, or perhaps unworkable in some circumstances (e.g. hybrid diesel generators). A practical solution is to improve

* Corresponding author.
E-mail addresses: neha_fet@sgtuniversity.org (N. Sehrawat), sahilvashist90@gmail.com (S. Vashisht), amritpal_bcet@yahoo.co.in (A. Singh).

https://doi.org/10.1016/j.ijin.2023.04.001
Received 15 November 2022; Received in revised form 13 April 2023; Accepted 13 April 2023
Available online 15 May 2023

photovoltaic predictions [2]. If the amount of potential solar energy can be anticipated precisely and with less uncertainty, it will assist grid operators in controlling power supply and demand and also creating solar systems more optimised. Grid stability would also rise with precise estimates [3]. The choice and refinement of forecasting models is the most crucial aspect of producing photovoltaic energy generation, which is why it is a crucial topic in the field of solar energy research.

The structure of the article is described as follows:

- The Introduction section serves as an opening to the paper and aims to provide the reader with an understanding of the problem statement. Along with a comprehensive review of the literature related to the research area. Briefly identify any gaps in the current knowledge and set the stage for the proposed work.
- The proposed approach section outlines the research design, the research questions, the data collection methods, and the data analysis techniques of the proposed work.
- Results and Discussion section discusses the implementation of the suggested model along with a comparison of the proposed model with existing models based on performance evaluation measures.
- The Conclusion section summarizes the main findings of the paper and discusses the implications of the research.

Table 1
Comparison table of solar irradiance forecasting using ML models.

Author	Model	R^2	RMSE	MAE	RAE
Huang et al. [19]	Gradient Boosting Regression Tree (GBRT)	.938	1.987	.899	NA
	Extreme Gradient Boosting (XGBoost)	0.944	1.131	0.87	NA
	Gaussian process regression (GPR)	0.941	1.811	0.899	NA
	Gaussian process regression (GPR)	0.941	1.811	0.899	NA
	Random forest	0.936	1.532	0.915	NA
V. Gunasekaran et al. [30]	Linear Regression(LR)	0.95	NA	14.73	NA
	Extreme Gradient Boosting(XGB)	0.982	0.87	5.39	NA
	Genetic Algorithm Optimization (GA)	0.984	NA	4.64	NA
Jumin*et al.* [21]	Neural Network Regression: Gaussian Normalizer	0.76603	0.14265	0.115	0.4233
	Neural Network Regression: Binning Normalizer	0.81802	0.12581	0.098	0.3694
	Boosted Decision Tree Regression	0.99956	0.00616	0.0237	0.08934
	Linear regression	0.81683	0.12622	0.0945	0.3552
Najibi et al. [22]	Gaussin Process Regression	NA	0.426	0.26	NA
Jun Qiu et al. [23]	Step wise Regression Method	0.841	3.093	2.474	NA
	Linear Correlation Analysis	0.823	4.123	2.41	NA
Shahin et al. [7]	Artificial Neural Network (ANN)	0.95	NA	NA	NA
Agbulut et al. [24]	Support Vector Machine(SVM)	0.898	2.651	1.698	NA
	K Nearest Neighbour (KNN)	0.89	2.748	1.87	NA
	Deep Learning (DL	0.915	2.443	1.505	NA
	Artificial Neural Network (ANN)	0.913	2.417	1.54	NA
Narvae et al. [6]	Neural Network (1 hl)	0.74	168	88	NA
	Neural Networks (3 hl)	0.75	167	85	NA
	Linear Regression	0.7	182	113	NA
	daBoost Decision Tree	0.57	219	172	NA
	Random Forest	0.99	66	33	NA
	Quantile Mapping Empirial Distribution	0.58	210	110	NA
Mauceri et al. [25]	N-SIM TSI	0.98	NA	NA	NA
	NRLTSI2	0.96	NA	NA	NA
	EMPIRE	0.95	NA	NA	NA
	SATIRE-S models	0.96	NA	NA	NA
Basaran et al. [26]	Support Vector Regression (SVR)	79.21	85.5	NA	NA
	SVR-Boosting	73.41	83.24	NA	NA
	SVR-Bagging	75.54	84.57	NA	NA
	ANN	75.13	73.5	NA	NA
	ANN-Boosting	68.65	70.24	NA	NA
	ANN-Bagging	65.54	68.39	NA	NA
	Decision tree (DT)	69.05	72.39	NA	NA
	Decision tree (DT) Bagging	69.75	66.85	NA	NA
	Decision tree (DT) Boosting	63.01	68.13	NA	NA
Jiaming Li*et al.* [29]	Hidden Markov Model(HMM)	64.6	NA	NA	NA
	Support Vector Machine(SVM)	73.1	NA	NA	NA
Guermoui et al. [30]	C-SVM	0.942	2.286	NA	NA
	Conventional SVM model	0. 936	2.575	NA	NA
Usman Munawar et al. [31]	Random Forest (RF)	0.93	6299.84	NA	NA
	Random Forest and PCA	0.91	8373.24	NA	NA
	Artificial Neural Network (ANN)	0.59	41686.89	NA	NA
	Artificial Neural Network (ANN) and PCA	0.81	18,633.53	NA	NA
	XGB	0.98	358.11	NA	NA
	XGB and PCA	0.99	2.49082	NA	NA
P Kumar*et al.*citekumari2021extreme	Support Vector Regression (SVR)	NA	45.12	NA	NA
	Random Forest (RF)	NA	43.68	NA	NA
	XGBOOST	NA	40.53	NA	NA
	DNN	NA	38.15	NA	NA
	XGBF-DNN	NA	35.47	NA	NA
Belaid barar et al. [33]	CLARA	0.96	NA	NA	NA
	ERA5	0.92	NA	NA	NA
	LR	0.97	NA	NA	NA
	RFR	0.97	NA	NA	NA

1.1. Solar irradiance

Solar irradiance is the total quantity of solar energy that reaches the earth's surface over a particular period of time. Watts/m2 is the unit used to measure it. However, the names Direct Normal Irradiation (DNI), Diffuse Horizontal Irradiation (DHI), and Global Horizontal Irradiation (GHI) are the most frequently used kinds of solar radiation. Direct normal irradiance is the term used to describe the greatest direct (Ninety degrees) downward solar radiation for a certain place. The radiation is considered when computing diffuse horizontal irradiance, and it need not come from the sun directly; it may come from any direction. The total quantity of direct or indirect energy that a surface horizontal to the ground absorbs per area is known as global horizontal irradiance [4]. Artificial intelligence (AI)-based modelling, analysis, and prediction techniques are essential for managing renewable energy. The techniques used to model, regulate, or predict how effectively energy systems will perform are intricate and need powerful computers, extended computation durations, differential equations, and other mathematical constructs [5]. Rather than intricate regulations and mathematical formulas, AI techniques are used to comprehend the key information structures inside a diverse information domain. The design, administration, and performance of solar energy systems require a long sequence of climatic parameters, such as those regarding solar irradiance, heat, or wind. For the majority of the interest areas, such long-term measures are frequently absent, or when they are there, they have a variety of drawbacks (e.g. data of poor quality, not enough lengthy series, etc.) AI approaches seem to be one of the best options for solving these issues. The study provides a general review of the prominent AI approaches in the solar energy area, focusing on supervised machine learning algorithms and ensemble techniques like bagging, boosting and stacking [6]. Fig. 1 clearly illustrates that solar irradiance levels in a particular area can affect photovoltaic power generation directly, and it can be directly used as a source for on-grid as well as off-grid power supply conventionally.

One of the key components of solar energy generation and one of the "hottest" research areas is improving predicting models. Future solar power generating statistics can be predicted using software tools called solar energy prediction models. The anticipated amount of energy would have some level of uncertainty and inaccuracies, just like any system that makes predictions about the future. Future values can be predicted using a good prediction model with the fewest errors and uncertainties [7].

1.2. Ensemble learning

In order to improve predictive performance, ensemble learning, a comprehensive conceptual framework for machine learning, integrates predictions from various models. There are three techniques that dominate the field of ensemble learning, irrespective of the fact that you can generate an apparently endless number of ensembles to address your predictive modelling problem. The detailed study of algorithms has given rise to numerous more specific techniques. The techniques are bagging, stacking, and boosting, and it is critical to grasp each technique in-depth and take it into account in a predictive modelling project [8].

1.3. Digital twin and AI

The term "Digital Twin" (DT) refers to a method combining traditional data-gathering techniques with Machine Learning and Artificial Intelligence to create innovative models and analyses. A DT is a living, breathing simulation model that combines big data analytics, IoT, AI, ML, and other technologies to represent its behaviour in close to real-time. A DT model regularly absorbs information from various sources and updates itself. Fig. 2 shows relation between all the major terminologies of this digital era. It clearly depicts how the IoT devices are generating the big data which is extensively used in ML and AI, to finally develop a statistical model that serves as a digital twin for a real-world issue. The Digital twins are supposed to provide business value by detecting and preventing issues, forecasting performance, and optimising operations using real-time data. One of the major applications of DT is the optimization of positive energy districts(PED). A city that generates at least as much energy annually as it uses is known as a positive energy district (PED). The ability to pool resources, manage energy systems effectively across multiple buildings, and achieve economies of scale are the driving forces for developing entire positive energy districts rather than individual buildings. The digital PED twin has four essential components: a digital model, sensor system integration, big data and analytics, and a client layer. If districts and buildings are simulated, assessed, and tested before construction, they can be built to be more active, efficient, and robust [9].

Table 2
The Performance Comparison of different models on testing Dataset for Shimla, Nagpur and Vishakhapatnam.

Location	Model	R^2	RMSE	MSE	MAE
SHIMLA	Stacking Regressor	0.9780	13.5401	183.3343	10.1010
	XGBoost Regressor	0.9759	14.1699	200.7869	11.1884
	Random Forest Regression	0.9758	14.2018	201.6915	11.3362
	XG Boosted Random Forest Regressor	0.9710	15.5454	241.6611	12.8699
	Cat Boost Regressor	0.9685	16.2069	262.6644	11.0042
	Linear Regression	0.9619	17.816	317.4120	14.3004
	Decision Tree Regressor	0.9484	20.7313	429.7901	15.7725
	Light GBM Regressor	0.5445	61.6424	3799.7881	18.1018
NAGPUR	Stacking Regressor	0.9818	12.7340	162.1570	9.5762
	XGBoost Regressor	.9808	13.2632	170.3712	9.4223
	Random Forest Regression	0.9799	13.3636	178.5877	9.4145
	Linear Regression	0.9790	13.6771	187.0654	10.3721
	Cat Boost Regressor	0.9787	13.7590	189.3118	0.4284
	XGBoost Random Forest Regressor	0.9778	14.0611	197.7169	10.3503
	Decision Tree Regressor	0.9710	16.0585	257.8785	11.9369
	Light GBM Regressor	0.91989	26.7278	714.3771	13.9006
VISHAKHAPTNAM	Stacking Regressor	0.9895	9.3274	87.0013	7.0547
	XG Boost Regressior	0.9882	9.8920	97.8517	7.4961
	Cat Boost Regressor	0.9882	9.8920	97.8517	7.4961
	Decision Tree Regressor	0.9813	12.4504	155.0143	9.1756
	XG Boosted Random Forest Regressor	0.9667	16.6497	77.2153	8.3887
	Random Forest Regression	0.9185	26.0366	677.9078	8.2664
	Linear Regression	0.8751	32.2356	1039.1387	10.7896
	Light GBM Regressor	0.7298	47.4256	2249.1955	15.7158

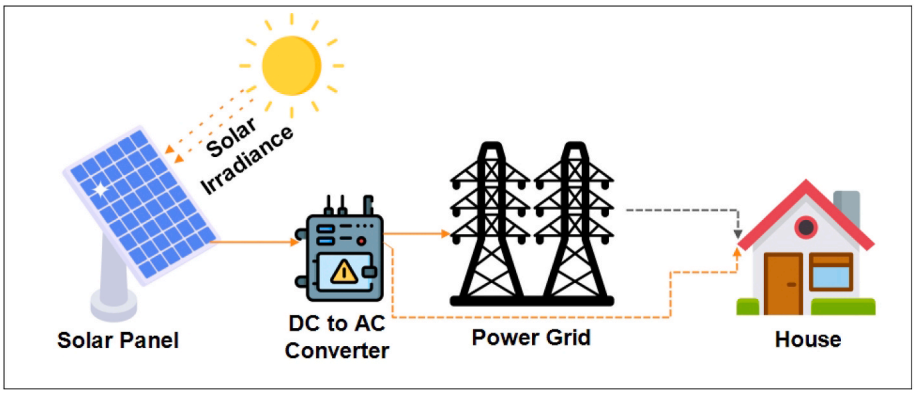

Fig. 1. Solar powered electricity generation.

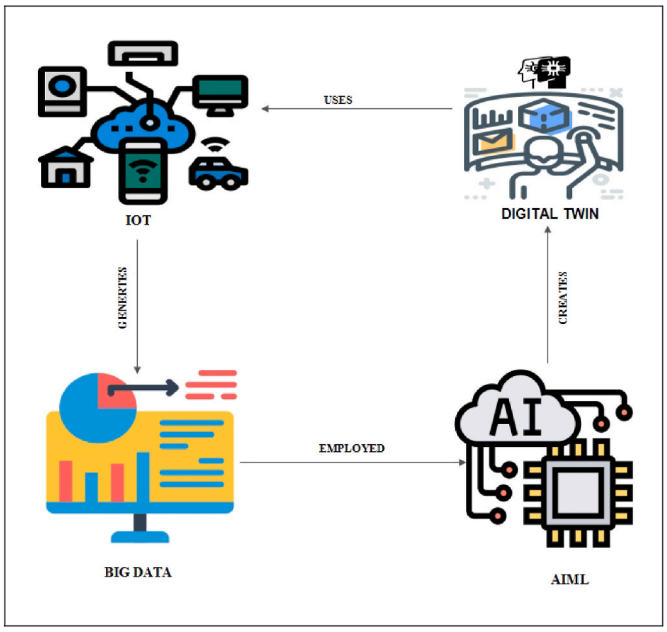

Fig. 2. Relation between digital twin, IoT, big data and AIML

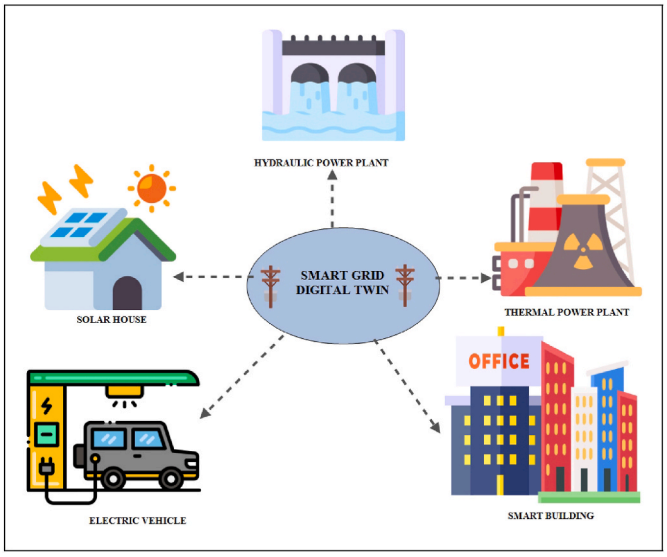

Fig. 3. Renewable energy and digital twin.

1.4. Digital twin and renewable energy power generation

Digital twins are one of the most promising techniques which are used when implementing applications based on renewable energy [10]. These powerful statistical models are increasingly being deployed to tackle challenges associated with the performance and management of renewable investments such as wind farms and solar plants in order to enhance their efficiency and extend their operational life [11]. Fig. 3 shows how various renewable energy sources can be utilized for creating smart city elements like solar-powered houses, smart buildings, and electric vehicles which can be potentially dependent on digital twins of grids for their smooth operation and management [12] Each renewable energy project has distinct characteristics that are related to the environmental conditions and terrain of its individual location. No two projects are alike, and each asset owner has unique problems based on the assets and systems they own [13] Therefore, there isn't a strategy that works for everyone and adaptability is essential when implementing such a system. For positive energy districts (PED) both the startup cost of renewable energy installations and the lifetime cost of energy-saving technologies are considered. Savings and revenues should be considered in the business model for these technologies, together with investment and maintenance expenses [14]. Renewable energy is more "uncertain" than conventional energy since it is "natural", which adds complexity to the maintenance and operation of the plant. In remote, difficult-to-reach areas, on-site maintenance and repair work can significantly increase operating costs and reduce profitability [15]. Although businesses that produce various forms of renewable energy employ analytics, it is typically descriptive and does not provide real-time insights. The efficiency of legacy monitoring methods is declining as plants get huge, grid management gets complex, and more people get involved in renewable energy systems [16]. As a result, industry leaders are aggressively researching innovative technological solutions: Another technology that is very helpful for renewable energy is digital twinning. It enables the creation of virtual replicas of real-world assets, processes, and systems. A digital twin is a computer-based simulation of a wind turbine that continuously receives updates from its physical counterpart via networked systems. The twin's prediction algorithm handles all incoming data and can be utilized to model various usage situations [17].

It helps to

- Calculate profitability and performance under various circumstances.
- Ensure digital inspections and real-time monitoring.
- Enable dynamic maintenance and fault prediction to assist and enhance the work of the engineers on-site.

Given the significant interconnection of financial and energy flows, it is imperative to have a solid understanding of the model for the energy

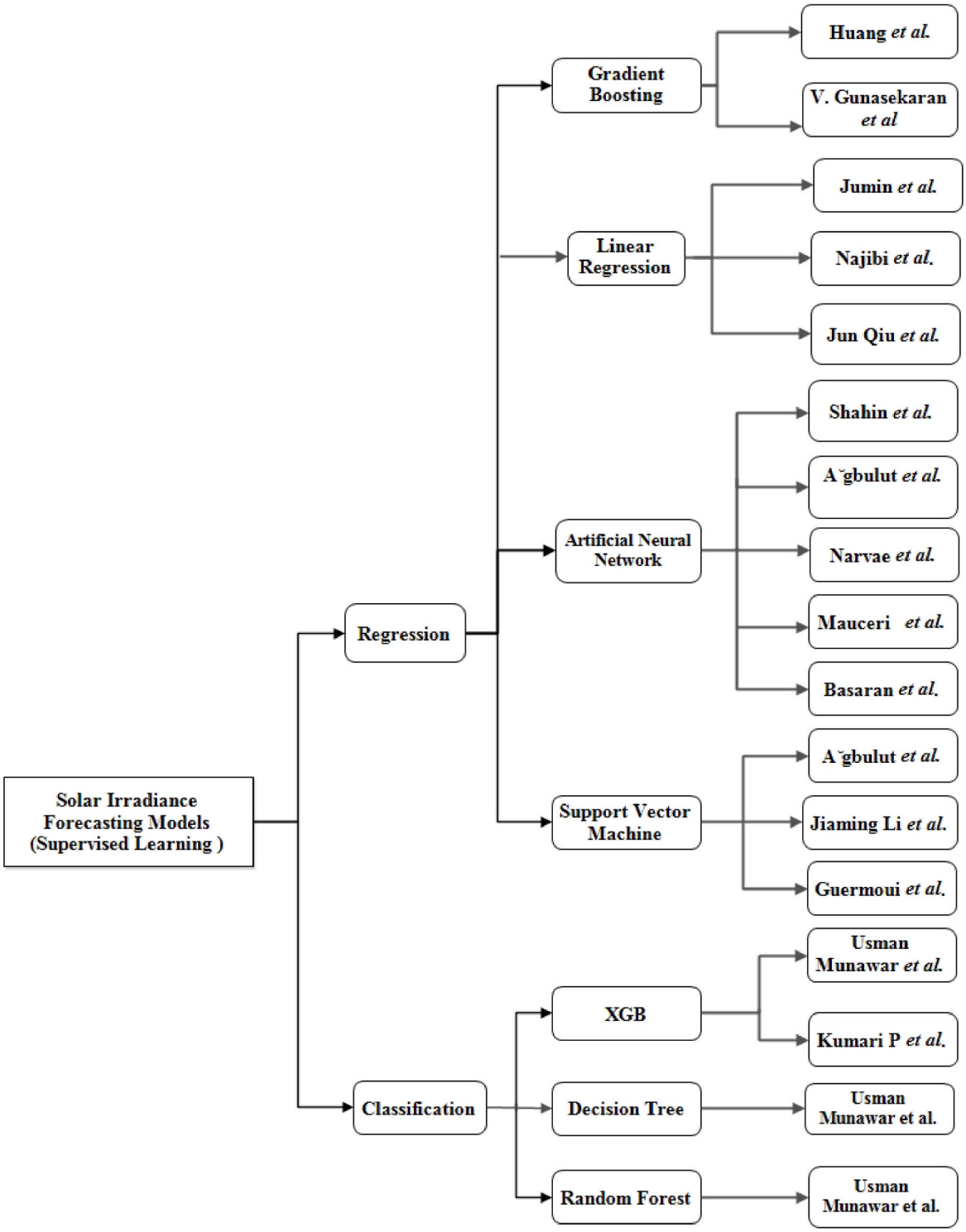

Fig. 4. Machine learning algorithms for solar irradiance forecasting.

flows in a power grid. Due to the various business models that these technologies propose, such as peer-to-peer, trading platforms, net energy metering, and electricity contract provisions, researchers have begun to investigate the interactions between prosumers (producer + consumer) inside an energy generating district [18]. The deployment of a DT can offer a number of advantages from the design to operation phases, permitting the selling and redistribution of energy depending on various economic models based on the power generation system's instruments and simulators which are deployed based on the requirements. Using a DT in the design stage will make it easier to forecast performance in the operational stage. The profit of energy investments can be increased, and investment risk can be decreased through ongoing learning [9].

1.5. Motivation

Solar power is a well-known Renewable Energy Source (RES) with many potential rights now. But because of this reliance on outside

meteorological circumstances, this technology is occasionally unpredictable and undesirable from an energy and cost standpoint. In the case of RES, these artificial intelligence techniques are used to forecast weather patterns as well as the amount of energy produced in a power plant. These projections are quite significant in the energy market. It is a well-known issue that the solar irradiance that reaches the Earth's surface is variable and unpredictable. As a result, a precise forecast of this variable can help with better planning and operation of power distribution at the economic or energy production process, either by establishing alternate plans for conventional power and overall schedules or by allocating the appropriate amount of energy resources and reserves in order to lower the costs associated with running the power system. A good prediction technique can also help to reduce the consequences of PV output unpredictability in addition to these benefits. Increase the level of PV System penetration while maintaining and improving the grid's power quality. These tools forecasts might also be helpful for researching the viability of solar power plants in a certain area.

2. Literature review

Solar irradiance forecasting has long been crucial in producing renewable energy. Predictions of solar irradiance positively impact a nation's economic progress with regard to solar power generation. ML techniques can be used to make prediction in various ways.

Fig. 4 illustrates the various learning techniques separated into supervised and unsupervised learning order. Supervised learning creates a function to predict a particular label from the input data. It can either involve identifying data (classification challenge) or predicting a result (regression algorithms). Regression models anticipate a continuous output, whereas classification models identify an object's category. This literature review reviews different classification and regression models to predict solar irradiance, considering various meteorological factors affecting it. The following section provides a glimpse of regression models.

2.1. Gradient boosting

One well-liked boosting approach is gradient boosting (GB). It is an ML technique for regression and classification problems that constructs a powerful estimation prototype from a series of weak prediction model results, frequently a decision tree. It expands the concept by allowing any variational loss function to be implemented, just like existing boosting techniques do [1]. The gradient boosting node uses a partitioning method to find the best possible data partition for a particular target variable.

As mentioned by Huang et al. [19], the XGBoost model outperformed with prediction result, other models were also studied and compared by results of correlation coefficient like for the Random Forest (R2 = 0.94), and the GPR (R2 = 0.941), GBRT (R2 = 0.938), Moreover, models performed well in making predictions. The prediction accuracy of the K-nearest neighbour (R2 = 0.900) and decision tree (R2 = 0.901) models was comparatively low. With an RMSE of 1.131 MJ/m2, among all the available models, the XGBoost had the best level of estimation. In terms of predicting daily sun irradiation, the stacking model—which incorporated the combination of G-BRT, XGBoost, GPR, and RF models—performed better than the standalone minature. But when it came to predicting monthly solar radiation, it had no use over the XGBoost model.

Study by gunasek et al. [20] shows that global solar irradiation is predicted using ML algorithms like LR, XGB, and GA, and the accuracy of each algorithm's prediction is evaluated. The outcomes unequivocally demonstrate that GA outperforms other ML approaches to forecast solar irradiation. XGB gave an accuracy of 98.4% as compared to GA with 98.2 and LR with 95.5%.

2.2. Linear regression

Linear Regression is indeed a learning-based algorithm. LR uses independent variables to model a targeted predicted value. Its primary application is to establish a link between variables and forecasts. Several investigations have used linear regression to predict solar irradiance, and all these studies are mentioned underneath. The authors in Ref. [21] investigated and compared four prediction models conjured up of boosted decision tree regression, linear regression, NN, Gaussian normalizer, and NN binning normalizer with R2 values of 0.89125 and 0.90183, respectively, All other models were outscored by enhanced decision tree regression utilizing two distinct data splitting techniques (80–20% and 75–25%). The authors Najibi et al. [22], address how to utilize the Gaussian Process Regression to simulate solar irradiation forecasting. As a kernel function, Matern 5/2 is used to establish a function that connects the above-chosen characteristics that produce solar energy. Five solar power plants located in various locations are used to test our methodology, and the results are contrasted with those obtained using other methods. In more detail, two separate techniques are employed to validate the proposed model: (1) a 5-fold cross-validation; and 2) keeping test data for 30 randomly selected days. On each of the four clusters, we use our framework 30 times separately to verify the model's accuracy. According to a normal distribution, the average error has values between 1.6% and 1.4% with a 95% confidence level.

Qui et al. [23], describes how to implement the two major techniques of machine learning concept, the step-wise regression technique(SRM) and linear correlation Analysis (LCA), to assess and identify the influencing elements that could have an impact on solar irradiation during various seasons on the Qinghai-Tibet Plateau. The input parameters for an ANN model were combinations of multiple factors in various seasons. The output was a prediction of the daily sun irradiation for the following day with a forecast period of each day. Based on the outcomes of the predictions for the various stations, the results demonstrate that the motivating factors generated from the SRM were more successful in constructing an accurate year-round prediction model than those from the LCA. Using the Golmud station as an example, the coefficient correlation R between the observation and the forecast was 0.841 for the SRM and 0.823 for the LCA, respectively.

2.3. Artificial neural Network (ANN)

A computationally effective method for identifying an empirical, potentially non-linear relationship between a number of inputs and outputs is to use an ANN. The ANN model is an intelligent system employed in several applications, like pattern recognition, prediction, modelling, clustering, simulation, and others, to solve complex issues. The ANN structure includes three layers, the input layer, maintains gathered data. The output layer, which delivers computed data, and one or more convolution neurons, can integrate the input and output levels. A neuron is an essential part of a neural network (NN) that gathers inputs and produces output. After each input has been multiplied by connection weights, their products, and biases, the activation function provides the output. A lot of work has been done to predict solar radiance using neural Networks for power generation analysis [3].

The authors Shahin et al. in Ref. [7], have explored an ANN-based mode with feed forward and back propagation techniques applied for forecasting of solar irradiance. Eight significant variables have been incorporated as independent input variables to forecast daily solar irradiance, including ambient temperature, wind velocity, rainfall, humidity, pressure difference, irradiance clearness index, and earth skin temperature. The assumptions made by the proposed model had a satisfactory accuracy of 0.95%.

In Agbulut et al. [24], The authors have looked into the possibility of four distinct ml algorithms (DLL, ANN, SVM and KNN) to predict daily global sun radiation. A few of the input factors include solar radiation,

daytime, cloud data, lowest and max temperature, the study evaluates the four different sites in Turkey (Krklareli, Tokat, Nevs ehir, and Karaman). Seven measures (Regression coefficient, Root mean squared error, Mean absolute error, t-statistics) are investigated in this work to evaluate the performance of the machine learning algorithms. The results varied depending on the input variables, data set size, lost data feature selection, tropical variances, changing the parameter values, etc. The ANN technique utilized in this work produced the optimal prediction outcomes.

The authors Narvaez et al. in Ref. [6], Site-adaptation is a strategy based on artificial intelligence that combines the best characteristics of the two data sources to generate a higher spatiotemporal resolution. Deep learning is also used to create very precise predicting solar radiation models. The author illustrates the advantages of applying the suggested methodology to combine data from many sources and create accurate solar radiation forecasting models through a real data study. The research claims that conventional machine learning models for site alteration achieve an accuracy of up to 38%. The NN model with three hidden layers scoring best with an accuracy of 75%. As mentioned in Ref. [25], According to citation, the author recently recorded and examined sun spectrum irradiance (SSI) measurements utilizing a novel techniques for deterioration repair in order to create the brand-new model. The NN for Solar Irradiance (SI) Modeling (NN-SIM) used in this research was developed, and it had a higher solar variation than other models and a 98% accuracy rate. The NRLTSI2, SATIRE-S model and EMPIRE model were also investigated. Their respective correlation coefficients were R = 98%, R = 96%, and R = 92%. In Ref. [26], the authors evaluated the fundamental learner models (SVR, ANN, and DT) and their associated bagging ensemble models and boosting models using hourly SI data from five cities in Turkey. The findings show that ANN and its ensemble models delivered 75.13, 68.65, and 65.54% accurate predicted results, but SVR, SVR-Boost, and SVR-Bagg models performed at 79.20, 73.42, and 75.54% respectively. DT and its teams for bagging and boosting: DT, DT-Boost, and DT-Bagg predict accuracy averages out to 69.05, 69.75, and 63.01%, respectively. Combinations of ANN-Bagging and ANN-Boosting provide extremely appreciated overall model performance.

2.4. Support vector machine (SVM)

SVM have recently gained a lot of popularity and are now widely used in a variety of engineering fields, including biomedical engineering [27], text classification, picture recognition [28], and regression and forecasting problems. The theoretical underpinnings of statistical learning and the structural risk minimising principle are the basis for the SVM algorithm. In Ref. [29], the author proposed a short-term solar irradiance forecasting algorithm based on SVM regression, Hidden Markov Model, and machine learning techniques. Regression using SVM performs better than HMM in any weather. Overall, SVM regression outperforms HMM with a 92% accuracy rate for sunny day and long-term predictions.

In [30], the authors proposed a unique approach based on SVM for daily horizontal global sun radiation estimates taking into account air and temperature conditions. The given method (Corrected-SVM) has been tested at the Ghardaa site in Algeria (2012–2015) for three years. In order to calculate the global horizontal irradiation, the SVM-model is corrected. For the suggested model, the RMSE was 11.35, and the Relation of coefficient R2 was 94.20%.

2.5. Extreme gradient boosting

A decentralized, extensible boosted decision gradient tree(GBDT) computational framework is called Extreme Gradient Boosting (XGBoost). It enables parallel tree boosting and is the ideal machine learning programme for prognostication, segmentation, and grouping challenges. The authors in Ref. [31], proposed an approach for determining the best models, feature selection techniques, and permutations for forecasting short-term solar output. In order to support and indicate that the ensemble of the XGBoost model and the PCA approach has the best performance, extreme gradient boosting (XGBoost) was investigated and tested using actual solar data. This results in the highest r2 score of 99% and the lowest root mean square error of 2.49082.

2.6. Light GBM

Built on decision trees, LightGBM is a gradient-boosting framework that enhances model performance while consuming less memory. It addresses the issues with the histogram-based technique, the primary method deployed in all GBDT (Gradient Boosting Decision Tree) frameworks. It integrates different innovative strategies: gradient-based one-side sampling and exclusive feature bundling (EFB). They empower the model to perform well and give it a distinct advantage over other GBDT frameworks. In Ref. [32], the author has implemented an ensemble machine learning model combining XGBF and DNN, two advanced base models, to estimate hourly global horizontal irradiance. This ensemble model had the highest prediction accuracy. The best results for XGBF-DNN have been seen in the autumn in Jaipur RMSE 1435.47 when the atmosphere is primarily clear. Compared to the above, when the sky conditions are significantly changing during the monsoon in Gangtok, XGBF-DNN has demonstrated the lowest performance RMSE 129.83.

2.7. Random forest

Supervised machine learning approaches like random forest are commonly used in classification and regression problems. It generates decision trees utilizing random samples, employing their average for categorization and majority vote for regression.

In [33] estimates of surface solar irradiance from the Cloud, Albedo Radiation data collection (CLARA-A2) and ECMWF Reanalysis 5 (ERA5) Edition served as the input for a random forest regression (RFR) model. The sky-stratification experiment shows that the given model produces improved results in all-sky conditions, with significant gains in the intermediate cloud. Under identical climatic conditions, the suggested regression model was proven reliable, improving surface solar irradiance estimates to a comparable extent as it did in Norwegian locations.

2.8. CatBoost regressor

Yandex created the CatBoost Regressor in 2017, a relatively recent open-source machine learning method. The ability of CatBoost to blend several data types, such as images, audio, and text features, into a single framework is one of its main advantages. It provides a revolutionary approach to managing categorical data, requiring only a small amount of categorical feature transformation, in contrast to the majority of other ML techniques, which are unable to accommodate non-numeric values. CatBoost avoids this phase, which from the perspective of feature extraction, can be a very time-consuming and complex procedure when converting from a non-numeric state to numeric values. It integrates decision tree theory and gradient boosting. The Gradient boosting allows the fitted trees to learn from the errors of earlier trees and hence minimise errors by fitting the decision trees one at a time. The number of functions gradually increases until the selected loss function can no longer be minimised. CatBoost, on the other hand, generates uninformed trees by requiring all nodes located at the same level to test the exact predictor under the same conditions. The CatBoost does not create its decision trees using similar gradient boosting strategies. The tree structure is a regularisation to locate the best solution and prevent overfitting, while the blind tree technique provides a straightforward fitting strategy and CPU efficiency.

3. Proposed approach

The proposed ensemble framework's methodology for predicting daily solar irradiance is explained in the sections that follow. The historical weather data is pre-processed in the preliminary phase. The most significant input features are then implicitly chosen by the proposed model from the large-scale dataset. Moreover, the optimal structure of eight regression models is merged with ensemble techniques, including ridge regression cross validation and the LR, XGBRF, RF, XGB Boost, CatBoost, DT, GB, and LGBM models, to create a unique ensemble model stacking regression (8-STR CV). On the basis of meteorological data from three places with various climatic conditions, the model is validated. Additionally, the viability of the recommended ensemble model is assessed in comparison to other traditional learning models in regard to prediction accuracy, durability, and processing time (i.e. Gradient boosting, LR. RF and DT). Although it has been demonstrated that individual algorithms are capable of predicting solar irradiation, differences in their effectiveness allow us to compare, assemble, and select the ideal model for the best forecast.

We have chosen for our research geographically different locations in mountainous, plane and coastal regions in India. These areas are Shimla, Nagpur, and Vishakhapatnam. These locations always contribute significantly to our economy. In comparison to other nearby cities, the electricity consumption in these areas is very high. Future solar panel power predictions could benefit from the improved performance of our suggested model. This concept will also have a great chance of working in any isolated part of India. Some important weather parameters are linked to solar irradiance prediction. There are a lot of weather factors that matter such as temperature, humidity, wind direction, shortwaves, UV rays, wind speed, latitude, longitude, season, and time, which are strongly related to solar irradiance and can impact's its forecasting severely [5].

3.1. The 8 stacking regressor cross validation (STR-CV) algorithm

The data set is divided into learning and evaluation data. before the algorithm receives a collection of input values with varied weather factors. It runs multiple regression models by creating a simple pipeline which will scale the data, perform feature selection and applying PCA which prevents the predictive algorithms from overfitting. Moreover, the optimal structure of eight regression models is merged with ensemble techniques, including ridge regression cross validation and the LR, XGBRF, RF, XGB Boost, CatBoost, DT, GB, and LGBM models, to create an unique ensemble model (8-STR CV) stacking regressor model by passing values to pipeline and giving output in form of a data frames with a prediction accuracy score.

Algorithm 1. Stacking Regressor(8-STR CV) Algorithm

Algorithm 1: Stacking Regressor(8-STR CV) Algorithm

Input: Data input S =$(x_1, y_1), (x_2, y_2), ...(X_m, Y_m)$
Data: Base algorithm BL_T (t = 1,2,3,....T); Meta learning algorithm BL
Process
For i = 1,2....I;
$H_i = BL(S_i)$;
Train a base Learner H_i applying the level 0
end Data: Testing set$D$
S'= Φ;
New Data set For f = 1,2,3…F $Z_i f = H_f(X_f)$;
Use H_f to ensemble voting the Training Examples X;
end
S' =$S'U(Z_{if}, Z_i f, Z_i f.)..Y_f)$; # A new data set is finished
end
$H' = BL(S')$; Train Meta Learner H' by using the level 1
output
$O(X)$= $H'(H_1(x), H_2(X1), H_t(x))$

3.2. Historical characteristics and solar irradiance forecasting problems

Solar irradiance statistics are crucial for the sizing of freestanding photovoltaic systems used for solar energy conversion and renewable energy applications. Additionally, factors like ambient temperature, wind speed, pressure, and humidity have a significant impact on the availability of resources and the production of electricity using solar and wind energy in a given region [34]. Economic gains are impacted by erratic and fluctuating energy output. However, installing pyranometers in places with adverse climatic and topographic circumstances is exceedingly challenging and expensive in terms of upkeep. Therefore, the proposed algorithm may be readily applied to process the climatic data collected by satellites in order to forecast solar irradiance and therefore the power production at such sites. Machine learning has proven to be an effective investigative tool in a variety of fields, including natural language processing, picture recognition, and forecasts. A well-liked area of research is the application of ML for the development of solar irradiation models.

Meteorological parameters are derived from GEOS 5.12.4 FP-IT and NASA's GMAO MERRA-2 assimilation model. The Goddard Earth Observing System (GEOS) Data Assimilation System of NASA is known as MERRA-2. The grid resolution of GEOS version 5.12.4 is the same as that of MERRA-2 (and the same model physics less selected observations and surface rain gauge normalised precipitation). The POWER project team processes daily GEOS version 5.12.4 data and provide low-latency products that are added to the final of the MERRA-2 daily time series and are often available within two days following real-time. The MERRA-2 data in the resulting daily time series are frequently updated every few months. A detailed description of the selected locations is given below.

I. **Shimla** The capital and largest city of the Himachal Pradesh state in northern India is Shimla. It is situated on the southern Himalayan ranges at 31.61°North 77.10°East. It extends along a ridge with seven spurs and is 2,206 m (7,238 feet) above mean sea level on average. From east to west, the city is roughly 9.2 km (5.7 mi) long. Shimla's climate is usually cold in the winter and moderately warm in the summer. Temperatures often fluctuate from 4 °Celsius (25 °Fahrenheit) to 31 °Celsius (88 °Fahrenheit) throughout the course of a year.
II. **Nagpur** Nagpur is the third-largest city in India. Nagpur is situated close to the quadrilateral geometric centre at the precise

centre of the Indian subcontinent. The hottest month of the summer, which lasts from March through June, is May. Temperatures below 10 °C (50 °F) are common during winter, which lasts from November through February. Annually, it receives about 355 mm of rain.

III. **Visakhapatnam** The coastal region Visakhapatnam city is located in the range of the Eastern Ghats and Bay of Bengal. The city is located at 17.7041 N and 83.2977 E latitude and longitude. The city has a 682 km^2 area. The elevation is 45 m on average. The climate is tropical. The maximum and minimum temperatures alternately occur in May and January, with the yearly mean temperatures falling between 24.7 and 30.6 °C, 1,118.8 mm of rain have been observed on average annually.

3.3. Proposed workflow model

The various phases in the PAWM proposed scheme are discussed in below sections and depicted in Fig. 5:

3.3.1. Data collection

The Data set is acquired from the official website of NASA, of three geographically separated locations in India, namely Shimla, Nagpur, and Vishakhapatnam dividing the locations categorically as mountainous, plains, and coastal region respectively for 3 years, day-wise (2019–2021).

3.3.2. Data preprocessing

Only historical solar data will be used in this project to forecast solar irradiance at any selected period. This work uses data sets from three Indian locations— Shimla, Nagpur, and Visakhapatnam—for 5 years (from 2018 to 2022) to run the simulation. While the last 2 years' worth of input is utilized to validate and test the built models, the initial three years' worth of data is used for training. It is not necessary to use all of the collected data because the sun irradiation is low during the early morning and nighttime hours. As a result data input of 9hrs (i.e. the hours between 8:00 and 16:00 each day) of irradiance data are taken into account for training and testing models. Also, the input set to make the model learn contains time and the most recent hourly value of meteorological parameters. The model's output is the All-Sky Surface Downward for each day. The data was prepossessed by removing the missing values, performing normalization, and using a filter-based feature selection module to identify the most important features that will affect the prediction of the target variable by using K-select features technique. The feature scaling method standardises the independent features that are present in the data within a specified range. We also employed the dimensional reduction method known as probabilistic component analysis (PCA). All these data prepossessing techniques are used in the pipeline on the data set.

3.3.3. Proposed approach- integrated ensemble model

A series of regression and numerous boosting models make up the proposed ensemble model. The proposed model's pseudo-code is presented in the workflow diagram. The input features are processed independently by each base model to provide a unique prediction. These forecasts are then input into the ridge regressor cross-validation as metadata to produce the final predictions. The final result of the ridge regressor cross-validation concerning the input is the weighted average of each base model's unique output. To ensure base model diversity, the eight models are chosen. To be more specific, 80% of the training data for each base model are chosen randomly to make sure the data is unique. Each base model is trained on a distinct set of hyper-parameters and a different training data collection. The results of a few base models could have significant practical correlations, even though the aforementioned methods are sufficient to ensure that the suggested framework is diverse. The linear regression hypothesis, which contends that each prediction should be independent, may conflict with this. Ridge regression, which successfully addresses the issue of combining the forecast from base models to assure optimal outcomes, is employed along with the cross-validation technique and XGBoost Regressor, which outperforms all the individual models.

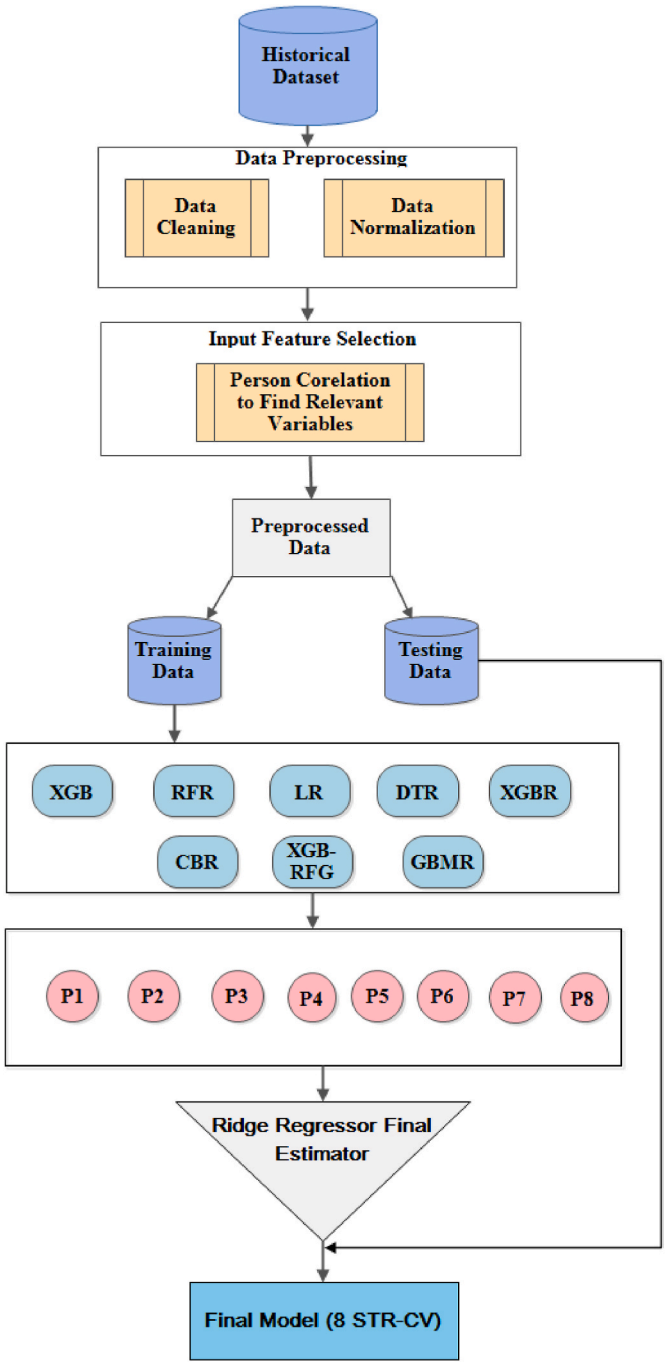

Fig. 5. 8 stacking regression cross validation model.

3.3.3.1. Pseudocode. Input:

- Step 1: Split data set D1 into T (training data set), V (validation data set)
- Step 2: Initiate a Scalar, PCA, Kselector techniques with the Regressor models
- Step 3: Train the training data T on the pipeline
- Step 4: Predict the target for validation input

Compute the prediction and find the error and accuracy Output is the ensemble forecasting error and accuracy in form of performance metrics. In this study, 8 models were tested and compared using historical sun radiation for different months of the year as the input parameter in order to determine the most precise irradiance prediction model for the optimum power generation. The Stacking Regressor method 8 STR-CV is proposed with all the techniques in the pipeline and will be an ensemble method which will be an integration technique for the forecasting of solar irradiation where factors including heat, peak sun hrs, wind velocity, latitude, longitude, rainfall, month, and year are taken into consideration and the prototype is given instructions to estimate the irradiation. The various type of difficulties will be investigated, as choosing the appropriate indicators for solar energy maximum power output, running the provided model based on the data set, and contrasting the outcomes with those of current learning methods.

The following steps have been followed for the study and analysis of the data.

- Step 1: Identify your Target Variable (y) and feature variables(X)
- Step 2: Scaling the data using Standard Scalar, applying PCA, K-selector in the pipeline for data prepossessing
- Step 3: Applying Pearson co-relation for feature selection
- Step 4: Divide the data input into the learning set and evaluation set
- Step 5: Fit the data in the created pipeline of 8 models
- Step 6:Predict the Testing data

In this study, 8 individual models have been tested and compared to the proposed ensemble model comprising these 8 models in the pipeline of utilizing historical sun radiation with different years as the input parameters.

- I. Random Forest Regressor
- II. XGB Regressor
- III. Decision Tree Regressor
- IV. Gradient Boosting
- V. XGB Random Forest Regressor
- VI. Linear Regression
- VII. Catboost regressor
- VII. Light GBM regressor

With three alternative normalizers, and different climatic parameters based on the study, different metrics including coefficient of determination (R2), root mean square error (RMSE), and mean absolute error (MAE), are used to evaluate and compare the prediction accuracy the models simulation results were evaluated along with similarities and debate. As the result, the best model comes out to be with the proposed ensemble model 8-STR CV for both regions.

4. Result and discussion

4.1. Pearson correlation coefficient

The assessment of the Pearson correlation coefficient is the linear relation between two variables. Its value falls between −1 and 1, with a range that includes: • A value of −1 indicates a negative linear relationship between two data input • A 0 value shows that there is no linear association between the two data input. • A value of 1 denotes a perfect positive linear relationship between two data inputs. The association between the two variables is stronger the farther the coefficient of correlation is from zero. However, there are times when we need to comprehend the relationship between more than one set of related variables. In these instances, a relationship between the variables, which is a square table depicting the correlation coefficients between different factors, can be generated. Fig. 6 shows the correlation of all the features variables with our target variable for the mountainous region (see Figs. 7 and 8)

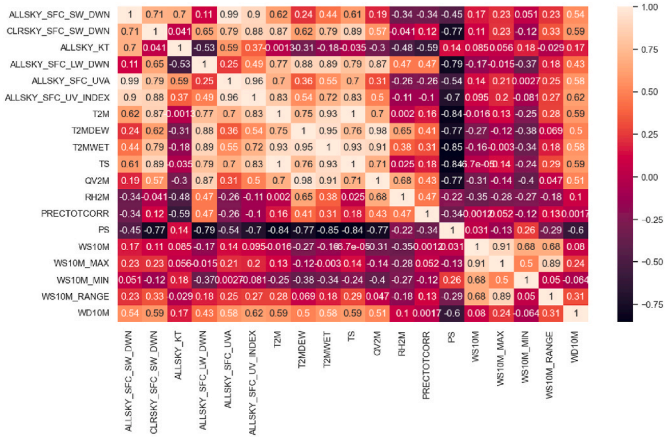

Fig. 6. Mountainous solar irradiance correlation.

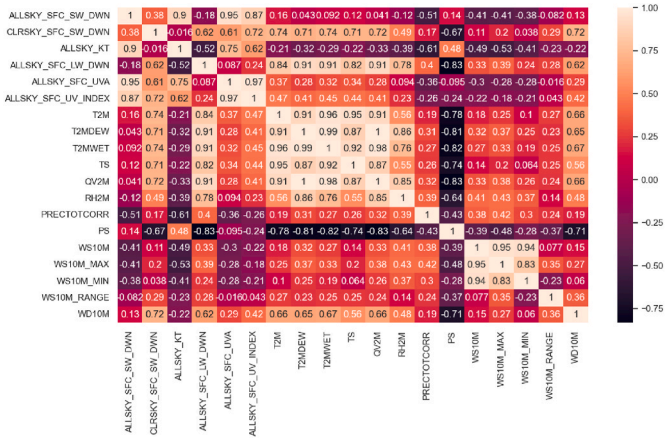

Fig. 7. Plains/coastal region solar irradiance correlation.

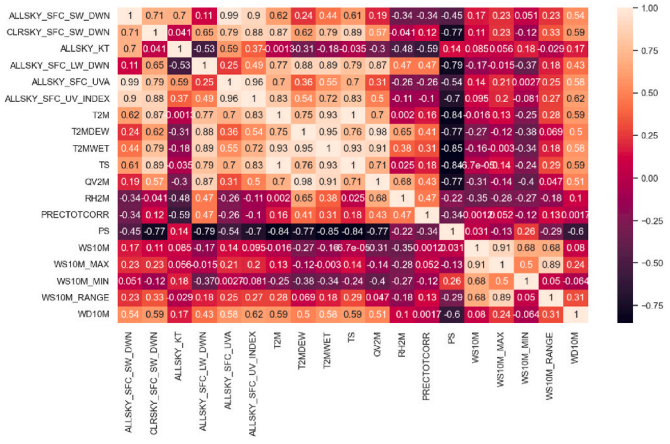

Fig. 8. Mountainous Region Solar Irradiance Correlation using proposed model.

4.2. Performance evaluation metric

The ML pipeline allows data to be processed and correlated into a model, which can then be examined to produce outputs, automating the machine learning workflow. The effectiveness of the developed miniature is assess using the mean bias error (MBE) and root mean square error. Higher model performance is correlated with a low RMSE value.

- R2 It assesses how well a model conforms to the facts. How well the Statistical evaluation of the accuracy of regression predictions fits the actual data points is called the R2 coefficient of determination. When the R2 value is 1, the regression predictions are completely accurate.

$$R^2 = 1 - \frac{\Sigma(x_i - y_i)^{2^2}}{\Sigma(x_i - \overline{X_i})} \tag{1}$$

- RMSE It offers details on the prediction models' immediate performance. It is preferred for its value to be near to zero and is always positive.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{y_i - x_i}{\sigma_i}\right)^2} \tag{2}$$

- MAE Mean absolute error is a metric for deviations between paired observations depicting the same occurrence (MAE). Examples of Y vs X comparisons include analyzing expected data to observed data, subsequent time to original process, By dividing the total absolute errors by the sample size, the MAE is calculated.

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}\frac{|y_i - x_i|}{n} \tag{3}$$

- MSE The mean squared error, also referred to as the estimator's mean squared deviation in statistics, is a measurement of the average of the squares of the errors, or the average of the squares difference between the two values that were estimated and the value that was actually obtained. A risk function is MSE, which stands for the estimated value of the squared error loss.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - x_i) \tag{4}$$

4.3. Results

The results of the suggested 8 STR-CV Model are presented and compared in this section with a standard of ensemble modeling and various conventional regression models, such as SVR, RF, XGBoost, DT, LGBM, Catboost Regressor for three different climatic areas in India (Vishakhapatnam, Nagpur, and Shimla). Fig. 9 graph depicts the accuracy score R^2 versus machine learning models for the three locations. For the mountainous region Shimla, the graph depicts the accuracy R^2 as a function of model performance. The Stacking Regressor (8 STR CV) has a value of 0.978024, the highest accuracy score achieved among all the compared models. The XG Boosted Regressor models yield values of 0.975932. The value for the XG Boosted Random Forest Regressor models is 0.975823. The values for the Light GBM Regressor models are 0.544520, the lowest of all the compared models. For the Plains region Nagpur, the stacking Regressor (8 STR CV) performs best, scoring 0.981818. The score value presented by the XGBoosted Regressor model is 0.980897, performing better than other models. Random Forest Regressor models yield a score of 0.979975. Light GBM Regressor models yield a score of 0.919899, the lowest performance score, making it the worst performer. Similarly, for the coastal region Visakhapatnam, the accuracy score bars depict that the stacking Regressor (8-STR CV) model returns a score of 0.989551, which outperforms all the models. The Cat Boosted Regressor and XG Boosted Regressor models show identical scores of 0.988248. The Light GBM Regressor models show scores of 0.729865, giving the worst performance.

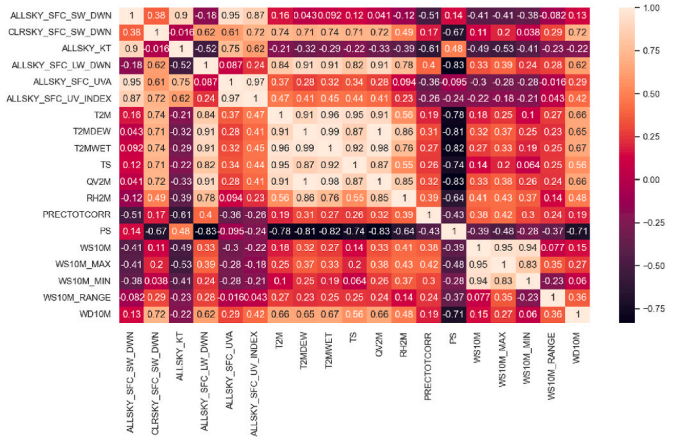

Fig. 9. Plains/Coastal Region Solar Irradiance Correlation using proposed model.

Fig. 10 graph with error rate (RMSE) versus models displays the values based on the performance of the models for the three geographically separated locations of India. For mountainous region i.e. Shimla, the stacking regressor (8-STR CV) model produce the best results, with values of 13.540102. The XG Boost Regressor models returns a value 14.169931. The values for the XG Boosted Random Forest Regressor models are 15.54545. While Light GBM Regressor models return 61.642422 giving the worst performance. For Nagpur region the Stacking Regressor(8-STR CV) model for Nagpur yield values of 12.734088 which indicates the best performance. Random Forest Regressor models exhibit value 13.263273. XGB Regressor models show value 13.363671. Light GBM Regressor models show values of 26.727834 delivering the worse performance. For Vishakapatnam the Stacking Regressor (8-STR CV) model for Vishakhaptnam indicate values of 9.327450 delivering excellent performance. Meanwhile XG Boosted Regressor models and Cat Boost Regressor models gave 9.892003 and 9.892003 RMSE score respectively performing moderately fair. Light GBM Regressor methods generate a very high error score of 47.425684, which is once again the poorest performance.

In Fig. 11, the MAE comparison with the machine learning models for the three locations is shown in the graph.

The mean absolute error (MAE) loss function is used in regression. When performing regression when you don't want outliers to have a significant impact it use MAE. A value close to 0 is more correct. For the moutaineous region, Shimla the Stacking Regressor models gives the best score of 10.101040. Value presented by XG Boosted Regressor models is 11.336299. The value for XG Boosted Random Forest Regressor models is 12.869961 and the worst score is given by Light GBM Regressor models of 18.101884. For plains region, Nagpur the MAE

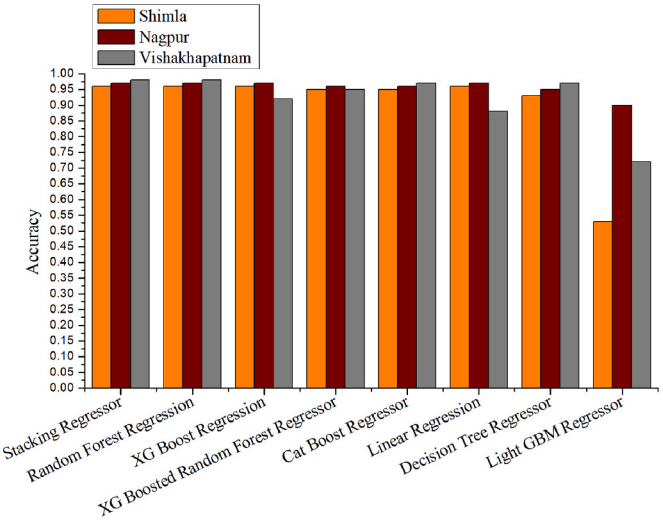

Fig. 10. R^2 comparison Graph plot for Solar Irradiance for Shimla, Nagpur, Vishkhapatnam.

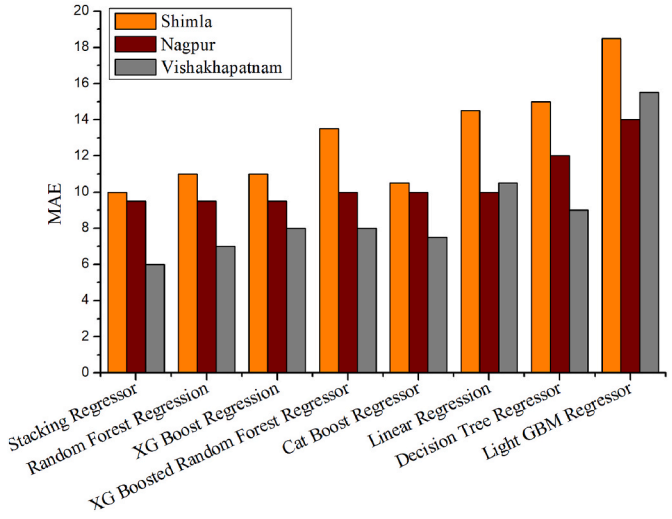

Fig. 11. MAE Comparison Graph Plot for solar Irradiance for Shimla, Nagpur, visakhapatnam.

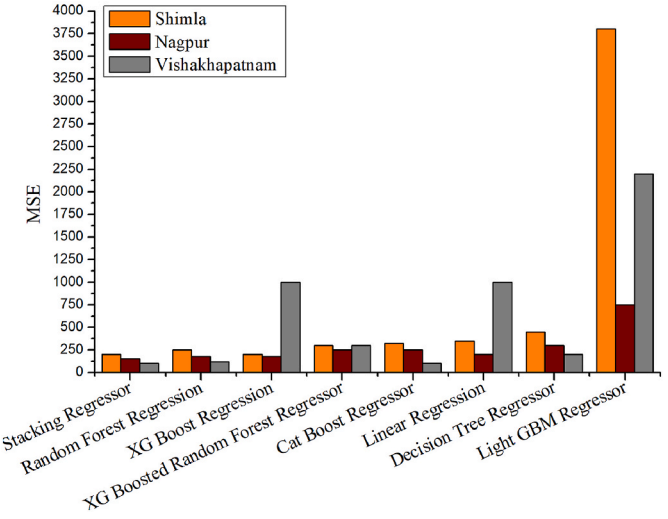

Fig. 13. MSE Comparison Graph Plot for solar Irradiance for Shimla, Nagpur, visakhapatnam.

score for the Stacking Regressor (8-STR CV) models gives a best score of 9.576206. The XG Boosted Regressor models shows values 9.422308 which is close to the proposed model and Light GBM Regressor models gives the highest score of 13.900620. And for the coastal regions Vishkhapatnam the MAE score is 7.054701 for Stacking Regressor (8-STR CV) model which is the best score available. The XG Boosted Regressor model and Cat Boost Regressor model shows values 7.496128 and 7.496128 respectively. The Light GBM Regressor models shows values 15.715831.

Figs. 12 and 13 the graph depicts the MSE comparison with the machine learning models for the three locations. For shimla the Stacking Regressor (8-STR CV) model shows best value of 183.334352. And Light GBM Regressor model shows the highest score of 3799.788173. For Nagpur the score goes to 162.157009 for the Stacking Regressor (8-STR CV) model and Light GBM Regressor models shows gives a score of 714.37711. For vishkhapatnam the Stacking Regressor models shows values 87.001319 which is lowest among all the compared models and Light GBM Regressor models shows values 2249.195548 which is extremely high.

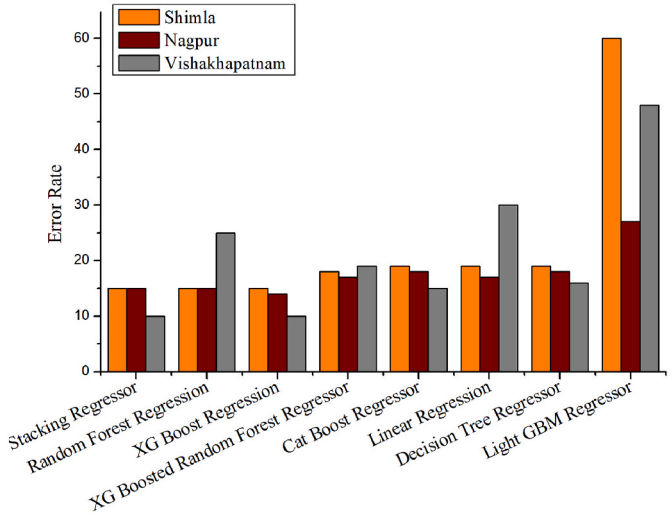

Fig. 12. RMSE Comparison Graph Plot for solar Irradiance for Shimla, Nagpur, visakhapatnam.

4.4. Findings

However, Fig. 9 displays the performance of the forecasting models on a three-year testing data set for the mountainous region and the coastal/plains region in terms of R2. The RMSE values shown in Fig. 10 indicate that there is a significant variation in the solar irradiance prediction errors throughout the coastal regions when employing different ML models. Coastal/plains shows the high errors than mountainous region. As demonstrated by Figs. 9 and 10 the Light GBM, the model displays the worst results across all sites. For Nagpur and Vishakhapatnam XG Boost Regressor outperformed all other individual results the highest accuracy R^2 of 0.9808 and RMSE score of 13.05 also for Vishakhapatnam XG Boost Regressor has an accuracy of 0.988 and RMSE score 9.892. Light GBM Regression showed the worst performance with 0.7298 accuracy and 47.42 RMSE score. However, our suggested model 8-STR-CV surpassed them all, delivering the best performance with accuracy R2 scores of 0.9895 and 9.32 for Vishakhapatnam and accuracy R2 scores of 0.9818 and 12.73 for Nagpur. For Mountainous region also the XGBoost Random Forest Regressor showcased the best performance with accuracy score of 0.975 and RMSE score 14.01 among the individual models and Light GBM gave the worst performance with 0.544 and with a RMSE score of 61.642. Also the proposed model gave a very high accuracy of 0.9780 and RMSE score 13.54.

5. Conclusion

This article formulated an ensembled stacking regressor model based on a combined approach of the different machine-learning models to predict solar irradiance for optimal power generation. The outcome of this research work shows XGBoost Regressor outperformed all other machine learning models when used as a meta-regressor or final estimator in this study, achieving an accuracy of 98.8% for all three geographically selected locations. Additionally, adopting the proposed concept can support the development of digital twin models for positive energy districts and smart buildings. Although the model is presenting promising results in various geographical locations, there is still a substantial amount of scope for work to improve the performance and operational life of renewable assets like wind farms and solar plants. Therefore, the developed model has the potential to contribute significantly to the renewable energy industry by providing reliable predictions for various decision-making processes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Ziane, A. Necaibia, N. Sahouane, R. Dabou, M. Mostefaoui, A. Bouraiou, S. Khelifi, A. Rouabhia, M. Blal, Photovoltaic output power performance assessment and forecasting: impact of meteorological variables, Sol. Energy 220 (2021) 745–757.

[2] H. Sharadga, S. Hajimirza, R.S. Balog, Time series forecasting of solar power generation for large-scale photovoltaic plants, Renew. Energy 150 (2020) 797–807.

[3] M. Rana, A. Rahman, Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling, Sustain. Energy Grids Networks 21 (2020), 100286.

[4] H.K. Yadav, Y. Pal, M.M. Tripathi, Short-term pv power forecasting using empirical mode decomposition in integration with back-propagation neural network, J. Inf. Optim. Sci. 41 (1) (2020) 25–37.

[5] B.K. Puah, L.W. Chong, Y.W. Wong, K. Begam, N. Khan, M.A. Juman, R. K. Rajkumar, A regression unsupervised incremental learning algorithm for solar irradiance prediction, Renew. Energy 164 (2021) 908–925.

[6] G. Narvaez, L.F. Giraldo, M. Bressan, A. Pantoja, Machine learning for site-adaptation and solar radiation forecasting, Renew. Energy 167 (2021) 333–342.

[7] M.B.U. Shahin, A. Sarkar, T. Sabrina, S. Roy, Forecasting solar irradiance using machine learning, in: 2020 *2nd International Conference On Sustainable Technologies For Industry 4.0 (STI)*. Plus 0.5em Minus 0.4emIEEE,, 2020, pp. 1–6.

[8] M. Moreira, P. Balestrassi, A. Paiva, P. Ribeiro, B. Bonatto, Design of experiments using artificial neural network ensemble for photovoltaic generation forecasting, Renew. Sustain. Energy Rev. 135 (2021), 110450.

[9] A. Rasheed, O. San, T. Kvamsdal, Digital twin: values, challenges and enablers from a modeling perspective, IEEE Access 8 (2020) 21, 980–22 012.

[10] S. Aslam, H. Herodotou, S.M. Mohsin, N. Javaid, N. Ashraf, S. Aslam, A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids, Renew. Sustain. Energy Rev. 144 (2021), 110992.

[11] A. Ozbek, A. Yildirim, M. Bilgili, Deep learning approach for one-hour ahead forecasting of energy production in a solar-pv plant, Energy Sources, Part A Recovery, Util. Environ. Eff. (2021) 1–16.

[12] H.A. Kazem, J.H. Yousif, M.T. Chaichan, Modeling of daily solar energy system prediction using support vector machine for Oman, Int. J. Appl. Eng. Res. 11 (20) (2016), 10 166–10 172.

[13] M. Javaid, A. Haleem, R.P. Singh, R. Suman, Enhancing smart farming through the applications of agriculture 4.0 technologies, Int. J. Intell. Networks 3 (2022) 150–164.

[14] M. Shahzad, M.T. Shafiq, D. Douglas, M. Kassem, Digital twins in built environments: an investigation of the characteristics, applications, and challenges, Buildings 12 (2) (2022) 120.

[15] T. Saravanan, S. Saravanakumar, Enhancing investigations in data migration and security using sequence cover cat and cover particle swarm optimization in the fog paradigm, Int. J. Intell. Networks 3 (2022) 204–212.

[16] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, S. Mishra, Decision tree and svm-based data analytics for theft detection in smart grid, IEEE Trans. Ind. Inf. 12 (3) (2016) 1005–1016.

[17] A. Althobaiti, A. Jindal, A.K. Marnerides, Scada-agnostic power modelling for distributed renewable energy sources, in: 2020 *IEEE 21st International Symposium On" A World Of Wireless, Mobile and Multimedia Networks"(WoWMoM)*. Plus 0.5em Minus 0.4emIEEE, 2020, pp. 379–384.

[18] A. Agouzoul, M. Tabaa, B. Chegari, E. Simeu, A. Dandache, K. Alami, Towards a digital twin model for building energy management: case of Morocco, Proc. Comput. Sci. 184 (2021) 404–410.

[19] L. Huang, J. Kang, M. Wan, L. Fang, C. Zhang, Z. Zeng, Solar radiation prediction using different machine learning algorithms and implications for extreme climate events, Front. Earth Sci. 9 (2021), 596860.

[20] V. Gunasekaran, K. Kovi, S. Arja, R. Chimata, Solar Irradiation Forecasting Using Genetic Algorithms, 2021 *arXiv preprint arXiv:2106.13956*.

[21] E. Jumin, F.B. Basaruddin, Y.B. Yusoff, S.D. Latif, A.N. Ahmed, et al., Solar radiation prediction using boosted decision tree regression model: a case study in Malaysia, Environ. Sci. Pollut. Control Ser. 28 (21) (2021), 26 571–26 583.

[22] F. Najibi, D. Apostolopoulou, E. Alonso, Enhanced performance Gaussian process regression for probabilistic short-term solar output forecast, Int. J. Electr. Power Energy Syst. 130 (2021), 106916.

[23] J. Qiu, X.-J. An, Z.-G. Wu, F.-F. Li, Forecasting solar irradiation based on influencing factors determined by linear correlation and stepwise regression, Theor. Appl. Climatol. 140 (1) (2020) 253–269.

[24] Ü. Ağbulut, A.E. Gürel, Y. Biçen, Prediction of daily global solar radiation using different machine learning algorithms: evaluation and comparison, Renew. Sustain. Energy Rev. 135 (2021), 110114.

[25] S. Mauceri, O. Coddington, D. Lyles, P. Pilewskie, Neural network for solar irradiance modeling (nn-sim), Sol. Phys. 294 (11) (2019) 1–30.

[26] K. Basaran, A. Özçift, D. Kılınç, A new approach for prediction of solar radiation with using ensemble learning algorithm, Arabian J. Sci. Eng. 44 (8) (2019) 7159–7171.

[27] P.K. Chahal, S. Pandey, S. Goel, Hybrid approaches for brain tumor detection in mr images, in: *Advances In Computing And Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part I 3*. Plus 0.5em Minus 0.4em, Springer, 2019, pp. 264–274.

[28] J. Singh, D. Thakur, F. Ali, T. Gera, K.S. Kwak, Deep feature extraction and classification of android malware images, Sensors 20 (24) (2020) 7013.

[29] J. Li, J.K. Ward, J. Tong, L. Collins, G. Platt, Machine learning for solar irradiance forecasting of photovoltaic system, Renew. Energy 90 (2016) 542–553.

[30] M. Guermoui, R. Abdelaziz, K. Gairaa, L. Djemoui, S. Benkaciali, New temperature-based predicting model for global solar radiation using support vector regression, Int. J. Ambient Energy 43 (1) (2022) 1397–1407.

[31] U. Munawar, Z. Wang, A framework of using machine learning approaches for short-term solar power forecasting, J. Electric. Eng. Technol. 15 (2) (2020) 561–569.

[32] P. Kumari, D. Toshniwal, Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance, J. Clean. Prod. 279 (2021), 123285.

[33] B. Babar, L.T. Luppino, T. Boström, S.N. Anfinsen, Random forest regression for improved mapping of solar irradiance at high latitudes, Sol. Energy 198 (2020) 81–92.

[34] P. Rocha, J. Fernandes, A. Modolo, R. Lima, M. da Silva, C. Bezerra, Estimation of daily, weekly and monthly global solar radiation using anns and a long data set: a case study of fortaleza, in brazilian northeast region, Int. J. Energy Environ. Eng. 10 (3) (2019) 319–334.