Predicting Solar Radiation with Various Machine Learning Algorithms

M.MOUNIKA NAGA BHAVANI

Asst Professor

Computer Science and Engineering Narasaraopeta Engineering College

(Autonomous) Narasaraopet, Andhra Pradesh [medurimounika4@gmail.com](mailto:medurimounika4@gmail.com)

Ganji Karthik Reddy Student

Computer Science and Engineering Narasaraopeta Engineering College

(Autonomous) Narasaraopet, Andhra Pradesh [karthikreddyganji@gmail.com](mailto:karthikreddyganji@gmail.com)

Kollapudi Dhanush Student

Computer Science and Engineering Narasaraopeta Engineering College

(Autonomous) Narasaraopet, Andhra Pradesh [Dhanush.banti@gmail.com](mailto:Dhanush.banti@gmail.com)

***ABSTRACT*--Solar radiation, serving as Earth's main energy source exerts a profound influence on various natural phenomena. Its impact extends across weather and climate patterns, hydrological cycles, photosynthesis in vegetation, and the equilibrium of surface radiation. For this reason, the solar industry and climate research depend heavily on precise solar radiation forecasts. To forecast and compare solar radiation values, we built twelve machine-learning models. Then, we used the best of these algorithms to create a stacking model that predicted solar radiation. The findings demonstrate the critical role that meteorological parameters—like temperature, humidity, time, wind direction, and pressure—play in Machine learning models. Research investigating the correlation between land surface temperature and solar radiation levels underscored the pivotal role of solar radiation in exacerbating catastrophic climate events. Various predictive models, including extreme gradient lifting (XGBoost), gradient boosting regression tree (GBRT), Extra tree regression, Gaussian process regression (GPR), Support vector machine(SVM), Bagging regressor, and random forest, were employed to forecast solar radiation levels. A stacking model combining GBRT, XGBoost, Bagging regressor, Extra tree regression, and random forest models outperformed individual models in solar radiation prediction. Nevertheless, The stacking model did not yield significant improvements over the Extra tree regression model specifically in solar radiation forecasting. Consequently, the stacking model and the Extra tree regression model emerge as the most effective predictors of solar radiation levels. Extra tree regression and stacking model have R2 values 0.932 and 0.930 respectively.**

***KEYWORDS: Climate extremes model comparison, machine learning models, stacking model, meteorological data, and solar radiation forecast***

1. INTRODUCTION

Solar radiation stands as Earth's primary energy source, profoundly shaped by interactions with the atmosphere, hydrosphere, and biosphere upon reaching the planet's surface [1]. The intricate relationship between solar radiation and the Earth's climate underscores the sensitivity of global temperatures to even minor fluctuations in solar energy output [2]. Variations in solar radiation reverberate through critical aspects of the planet's dynamics, including humidity, temperatures, and the occurrence of extreme climate events [3]. As research delves into solar-energy applications, construction materials, extreme weather phenomena, and the climate trends, precise monitoring and assessment of solar radiation's spatiotemporal variability become imperative [4]. Diverse methodologies, ranging

from theoretical parameter models to satellite-based data retrieval systems, have been developed to forecast solar radiation accurately [5,]. Pioneered by [6], the A-P model serves as a cornerstone for estimating solar radiation, while the BCM model, introduced by [7], enhances understanding through the analysis of solar radiation's correlation with daily temperature variations. Further refinements, like the hybrid model (YHM) proposed by [8], optimize predictions by integrating meteorological parameters, a process validated in the context of Japan. Building upon the YHM,

[9] refined estimates of solar radiation components through comparison with a climatological model (CYHM). Investigations by [12] underscore the importance of model selection, while [10] explore innovative prediction approaches leveraging back-propagation algorithms and support vector machines (SVM), respectively. Notably, advancements in tree methods, such as the random forest and GBRT algorithms, as explored by various researchers [11,12], showcase promising outcomes in solar radiation prediction. Recent comparative studies [14] shed light on the nuanced performance of Machine Learning algorithms, highlighting both the limitations and potential avenues for algorithmic enhancement, particularly in artificial neural networks (ANN). In parallel, deep learning techniques, including hybrid networks and WT-LSTM proposed by [16], exhibit substantial promise in enhancing short-term solar radiation predictions. Notably, [17] introduces the ACEEMDAN–CNN–LSTM model, demonstrating superior accuracy in hourly multi-region solar irradiance forecasting. Across disciplines, from image recognition to natural language processing [18], Machine Learning proves to be a potent tool, continually evolving to unlock deeper insights into complex phenomena like solar radiation dynamics.



Fig.1. The location of Moscow City's sun radiation monitoring station

Machine learning has gained popularity as a field of research and is now essential to the creation of models of solar radiation. Nevertheless, there aren't many in-depth analyses of the variations between these models because much research has been on developing one or more Machine Learning techniques. In our investigation of the diversity among solar radiation forecast models, we embarked on an analysis employing a dataset comprising meteorological and fundamental radiation elements. Following meticulous data processing, we curated a subset of variables utilizing the random forest algorithm, yielding a refined dataset for further analysis. Leveraging Machine Learning techniques, we embarked on constructing models aimed at predicting solar radiation levels. Through a meticulous comparison of the predictive performance of the twelve distinct Machine Learning models, we discerned those exhibiting the most promising prediction abilities. Subsequently, we synthesized a linear model by aggregating the top-performing models in a stacked framework. The anticipated outcomes of the stacking model, we then proceeded to evaluate its efficiency in comparison with individual models, assessing metrics such as mean squared error, mean absolute value, and R2 values. Upon comprehensive comparison of all twelve Machine Learning models, we observed that both the Extra tree regression and stacking models demonstrated notably high R2 values, at 0.932 and 0.931 respectively. Consequently, we prioritize the construction of the Extra tree regression model, recognizing its superior predictive capabilities over alternative Machine Learning algorithms.

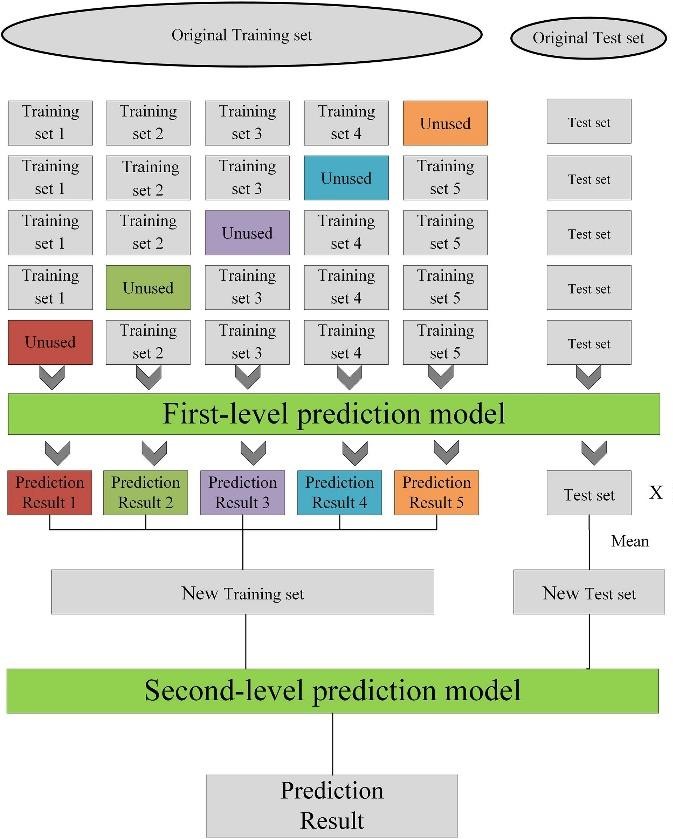


Fig. 2. The stacking model's framework

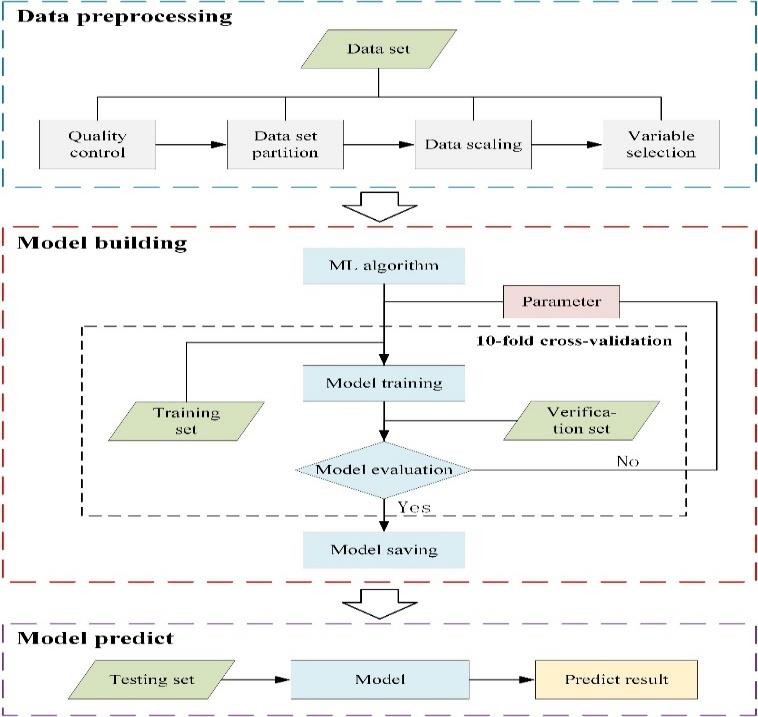


Fig. 3. A flowchart illustrating the machine learning methods employed for sun radiation estimation.

1. LITERATURE SURVEY

Li, Wang, and Sung et al.[15] introduced a novel approach termed AdaBoost with SVM-based component classifiers, which combines the strengths of the AdaBoost Machine Learning with the Support Vector Machine (SVM) classifiers. This methodology was proposed to enhance classification performance by leveraging the robustness of SVMs within the AdaBoost framework. The authors conducted experiments to evaluate the effectiveness of their approach, demonstrating its applicability in various engineering and artificial intelligence tasks. Through their research, they provided insights into the potential of integrating SVMs into ensemble learning paradigms like AdaBoost, offering improved accuracy and robustness in classification tasks. Their findings contribute to the ongoing exploration of innovative techniques for enhancing Machine Learning performance in practical applications.

Meenal et al. [16] conducted an in-depth investigation into the phenomenon of XGBoost's dominance in Machine Learning competitions through his Ph.D. thesis titled "Tree Boosting With XGBoost: Why Does XGBoost Win Every Machine Learning Competition?" from the “Norwegian University of Science and Technology”. In his research, Nielsen explored the underlying principles and mechanisms behind XGBoost's exceptional performance, focusing particularly on its tree-boosting algorithm. Through a comprehensive analysis of XGBoost's design, optimization strategies, and empirical evaluations, he elucidated the factors contributing to its effectiveness and widespread adoption in various Machine Learning tasks. Nielsen's work provided valuable insights into the reasons behind XGBoost's success, offering a

detailed examination of its features and capabilities that set it apart as a leading algorithm in the field of Machine Learning.

Agarwal et al. [20] proposed innovative ensemble learning techniques, namely A-stacking, and A-bagging, as adaptive versions tailored specifically for spoof fingerprint detection in their study published in Expert Systems with Applications. Their research aimed to address the challenges posed by spoof attacks in fingerprint recognition systems by enhancing the robustness and performance through ensemble learning methodologies. By integrating adaptive mechanisms into traditional ensemble learning frameworks, such as stacking and bagging, the authors introduced novel approaches tailored to the unique characteristics of spoof fingerprint detection tasks. Through empirical evaluations and comparative analyses, Agarwal and Chowdary demonstrated the effectiveness of A-stacking and A-bagging in improving detection accuracy and resilience against spoof attacks, thereby contributing to the advancement of security measures in biometric authentication systems.

Kapwata et al. [27] conducted a study focusing on the spatial malaria transmission model in Mpumalanga Province, South Africa, utilizing random forest variable selection techniques. Published in Geospatial Health, their research aimed to identify key environmental and socio-economic factors influencing malaria transmission in the region. By employing random forest variable selection, the authors sought to determine the most significant predictors among a multitude of variables, thereby enhancing the accuracy and efficiency of spatial malaria modeling. Through their investigation, Kapwata and Gebreslasie provided insights into the complex interplay between environmental conditions and malaria transmission dynamics, contributing valuable knowledge towards the development of effective malaria control strategies tailored to the specific context of Mpumalanga Province.

Ebden et al.[26] presented a concise introduction to Gaussian processes in his preprint titled "Gaussian processes: a quick introduction," available on arXiv. This work aimed to provide a clear and accessible overview of Gaussian processes, a powerful tool in Machine Learning and statistics. Through a systematic explanation of fundamental concepts and mathematical formulations, Ebden elucidated the principles underlying Gaussian processes, including their probabilistic nature and applications in regression, classification, and optimization tasks. As a result, a book is a useful tool for practitioners and novices in the fields of statistical modeling and Machine learning. The author's ability to simplify difficult mathematical ideas into clear explanations made it easier for readers to understand the principles of Gaussian processes.

1. ALGORITHMS FOR MACHINE LEARNING AND DATA

A. *Study Area and Dataset*

Moscow City is located in Russia at 55.7558° N and 37.6173°E. Over half of the Russian land receives daily average sun radiation over 3,5 kWh/m2 or annual solar radiation exceeding 1278 kWh/m2. The average temperature is 49-52 degrees Fahrenheit.

The Kaggle website is the source of the dataset[28]. To estimate solar radiation, we take data from 2016 into consideration. Variables including unixtime, temperature,

pressure, humidity, wind direction, speed, time, time sunrise and sunset times, and total solar radiation are all included in the dataset.

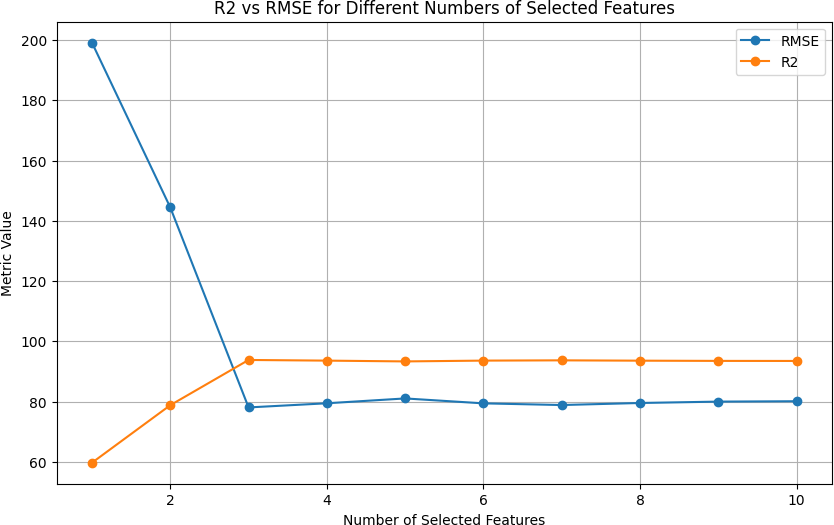


Fig. 4. Variable selection using the random-forest model's predictive performance (measured by RMSE and R2).

The quality control of data was crucial because of how long the inquiry took and since, according to our observations, the device had intrinsic flaws. We subjected our dataset to data quality control after eliminating anomalous and missing values. There are 32,686 data records in total in our data set. The dataset was further divided into the training set, which had 90% of the information, and a test set, which included 10% of the total data. We used 29,177 training sets and 3,509 test sets to create our final sample.

1. PREDICTIVE ALGORITHMS FOR MACHINE LEARNING AND

STACKING TECHNIQUES

1. *Algorithms for Machine Learning:*

More and more academics are utilizing Machine Learning to forecast solar radiation as a result of advancements in the field. We looked into 12 distinct Machine Learning forecasting algorithms: multiple-linear regression [19]. the K-nearest neighbor model and the Extra tree regression[20]. the back-propagation neural network [27], the decision tree [21], the extreme learning machine (), SVM regression[22], Gaussian process regression(GPR) [23] the GBRT [24], adaptive boosting (AdaBoost) [25], extreme gradient lifting (XGBoost)[26], random forest regression[15], and Extra tree regression model[13] is a full description of Machine Learning techniques

1. *Stacking Model*

According to Agarwal et al.[20], stacking technology represents a versatile integration technique that amalgamates high-level learners with multiple lower-level learners, thereby enhancing overall performance. Typically, the K-fold cross-validation methodology is employed to derive prediction outcomes after the training and testing phases of the Machine Learning models. The stacking model is meticulously crafted to mitigate generalization errors and is then

created by combining the prediction outcomes that each model has produced. Typically, the stacking model has two layers. The first training set serves as the input for the base learner, which is the first layer. The final results are obtained by employing the second layer, which is trained with the input data source being the output data from the first layer.

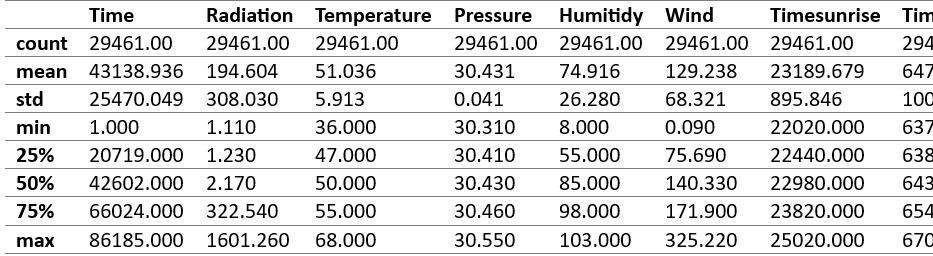
Training each model with five-fold cross-validation is a step in the construction of the stacking model, as plotted in Figure 1. With four sections designated for training data and one for test data, the training set is made up of five components. The trained model predicts the test set data to acquire another prediction result (b), whereas the test data from each of the four training sets is predicted to produce a prediction result (a). The averaged prediction result (b) following five training sessions is shown as B, while the individual prediction results from each of the five runs are combined into a single column as A. This yields new datasets A and B, where the number in A corresponds to the number of training sets, but the data is represented in only one dimension. Following the creation of N single models, NA and NB datasets are generated. These datasets are then merged to form a new test set and training set. The second layer employs a simple linear model to evaluate the new test set and train with the updated training set.

TABLE 1. THE TRAINING DATASET'S MODEL VARIABLES' DESCRIPTIVE STATISTICS

used the five-fold cross-validation method for choosing the parameters. A detailed description of the model building can be found in the Section "Model Building". For the model prediction stage, the model that was saved during the model creation procedure is utilized. The step was utilized with the test dataset to forecast the sun's radiation. After that, we save the analysis and anticipated outcomes. The particular steps of the experiment went as follows:

1. Gathering and preprocessing the data;
2. choosing a Machine-Learning algorithm to estimate solar radiation from the 12 available algorithms;
3. assessing the solar radiation's predictive capability according to different criteria;
4. save the model upon achieving the highest level of prediction accuracy;
5. After all 12 machine-learning algorithms have been exposed to the creation of machine-learning models, return to step 2 and choose a different Machine Learning Algorithm;
6. Preprocessing dataset is input;
7. projected results are saved and analyzed;

*B. Variable Selection*

Building Machine Learning models requires careful consideration of the variable selection process. The genetic algorithm, Tabu search, particle swarm optimization, and random forest algorithm are the mainstream variable selection strategies used today. To choose the data variables, we applied the random forest approach. The random forest model was built and trained using normalized data, and its significance was determined. The purpose of the data preprocessing was to examine the effects of variable changes on the predictive ability of the model and confirm the significance of the variables in a particular model. Upon completion of the data quality control process, the following steps are undertaken:

1. The dataset undergoes division into a training data set and a testing data set;
2. Model is trained and preserved using the training set, followed by computation of the R2value(R2) and the root mean square error (RMSE) of a saved model;
3. Variable elimination commences, starting with the removal of the least significant variable based on their relative relevance;
4. Steps 2 and 3 continue iteratively until there are only two variables left, which is the least amount needed to compute.

.

1. MATERIALS AND METHODS
   1. *Prediction of the Flow of Solar Radiation*

Three components made up our experiment (Figure 2): creating the model, predicting the model, and preparing the data. Four steps comprised the preprocessing of the data: variable selection, splitting, scaling, and quality-checking of the dataset. Section “Study Area and Data Set” describes data quality control, data set splitting, and data scaling, whereas the Section “Variable Selection” describes variable selection. The main steps in the model-building process that we follow are choosing the machine-learning method, choosing the parameters, developing the model, and saving the model. We

The temperature and the time are important features or variables in the dataset to predict solar radiation which is shown in Figure 5. The temperature has the highest importance when compared to the time.

1. Model Building

In Python 3.12 experiments, third-party libraries like NumPy, Pandas, Xgb, and scikit-learn (Sklearn) were utilized. Twelve machine-learning methods were employed for model creation, considering the unique characteristics of each algorithm for setting initial parameters. For instance, neural network architecture followed empirical methods and design principles to determine neuron count and hidden layer structure [. Parameter adjustment techniques were applied across various algorithms to define relevant parameter ranges. This process involved parameter selection for each of the 12 machine-learning models. Following this, the best model was saved. A stacking model's initial layer

consisted of multiple models exhibiting superior prediction abilities, with parameters retained from previous selections. The second layer employed multiple linear regression for model construction. Upon achieving optimal parameters, Using the train set, the model was trained, and finished model is stored. The training time is the time required to develop model, where the model memory is the final model

size. Following the process of construction of the model, the

testing set was utilized to generate prediction outcomes, specifically in building the GPR regression model, which took more time when compared to the other regression models and the R2 value for Extra tree regression and the random forest regression model is almost equal and there is a slight difference in their mean squared error. we want to preprocess the dataset before constructing a model. we construct the model by importing some required models and also using some Machine Learning algorithms.

1. *Statistical Metrics*

Four indicators were used to assess the models: BIAS, R2, RMSE, and MAE.



-(1)

 (2) ------------------------------(3)

 (4)

where the average of the expected and observed is represented by ym and yo. solar radiation, n denotes the amount of data, and ymt and yot stand for the predicted and observed solar radiation, respectively outcomes, in that order. The observed and expected values are highly associated if R2 is near 1. The proximity of RMSE/MAE values to 0 indicates a stronger alignment between predicted and observed values. Model performance is commonly evaluated using various metrics such as mean absolute error and R-squared error combined.

1. RESULTS
2. *Description and Selection of Variables*

The mean radiation level is 194.60 MJ/m2. The standard deviation for radiation is relatively high at 308.03 units. The minimum radiation level recorded is 1.11 units, while the maximum is 1601.26 units. The mean atmospheric pressure is around 30.43 units. The standard deviation for pressure is relatively low at 0.042 units, indicating relatively consistent pressure levels across the dataset. The minimum pressure recorded is 30.31 units, while the maximum is 30.55 units. The mean wind direction is approximately 129.24 degrees. The standard deviation for wind direction is 68.32 degrees, The minimum wind direction recorded is 0.09 degrees, while the maximum is 325.22 degrees. The mean temperature is around

51.04 degrees. The standard deviation for temperature is 5.91

degrees. The minimum temperature recorded is 36.00 degrees, while the maximum is 68.00 degrees. The average humidity is approximately 74.92%. The standard deviation for humidity is relatively high at 26.28%. The minimum humidity recorded is 8.00%, while the maximum is 103.00%. The mean time of sunrise is 23189.68 seconds. The mean time of sunset is 64717.89 seconds.

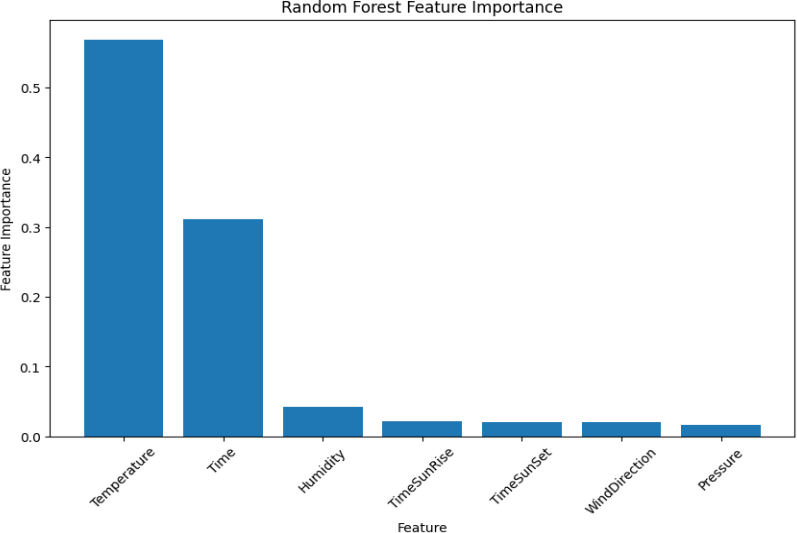


Fig.5 Factors to consider when utilizing the random forest model to predict solar radiation

1. *Predictive Performance for the Solar Radiation*

Twelve Machine Learning models showed considerable variance in their ability to estimate solar radiation, with the majority of the models producing acceptable results. The R2 values varied in terms of prediction accuracy, ranging from 0.62 to 0.93. Top-performing models including GBRT, Bagging regressor, XG Boost, Extra tree regression, and random forest achieved R2 values of 0.89, 0.92, 0.934, 0.934, and 0.934, respectively. Conversely, models like decision tree, SVM, Kmeans, and Linear regression showed lower precision, with R2 values ranging from 0.62 to 0.80. RMSE values varied between 1.0 and 6.19 MJ/m2, with GBRT exhibiting the lowest root mean square error(RMSE) at 1.0 MJ/m2 and random forest displaying the highest (6.19 MJ/m2). Mean absolute error (MAE) ranged from 28.4 to 140 MJ/m2, with random forest demonstrating the lowest variance (28.4 MJ/m2). Some models showed more consistent performance than others, with varying levels of prediction bias observed. Construction times varied, with random forest, GBRT, and GPR taking longer to build due to their underlying principles, while XG Boost boasted quicker construction times. Overall, The Extra tree regression model outperformed other algorithms in solar radiation prediction, albeit with variations in construction times and predictive accuracy among the models.

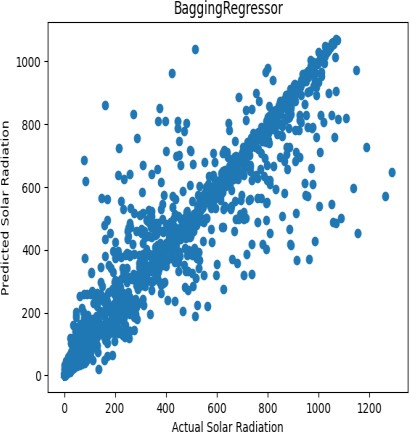
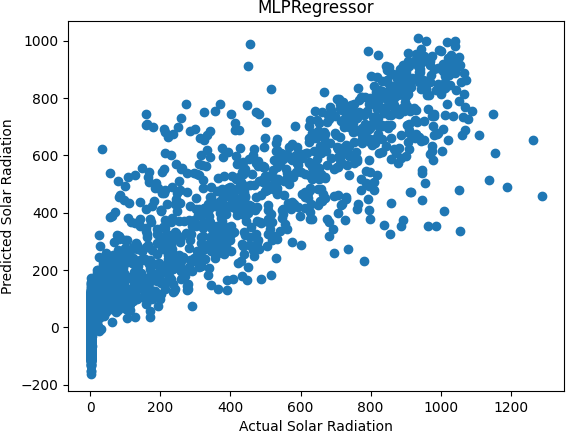
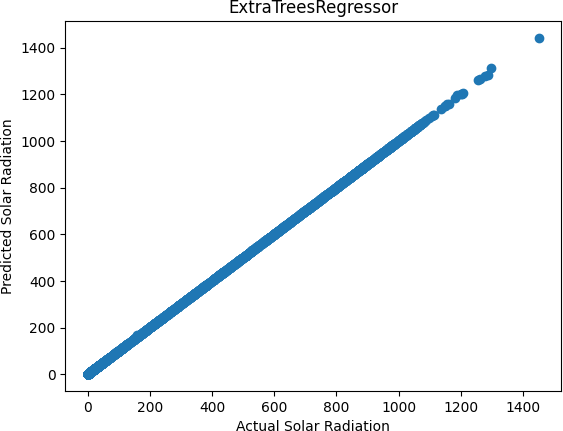
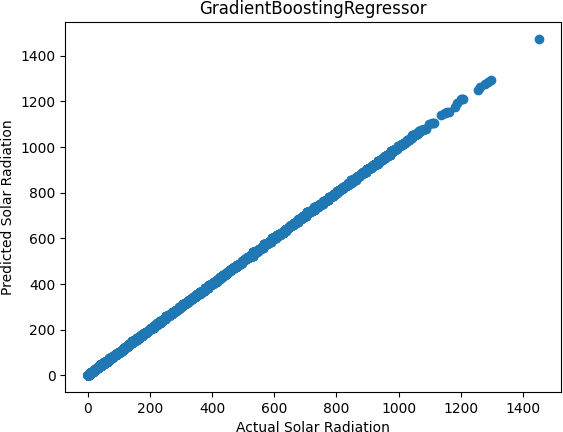
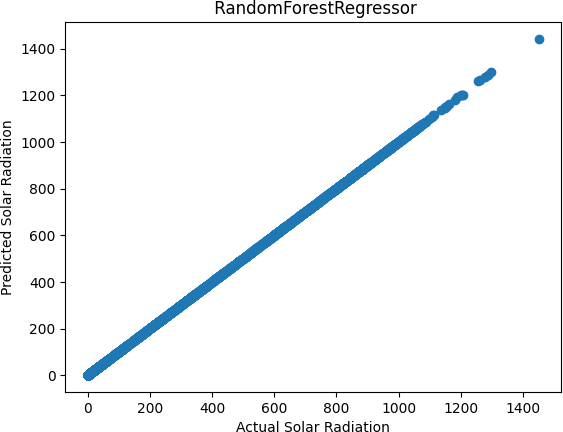
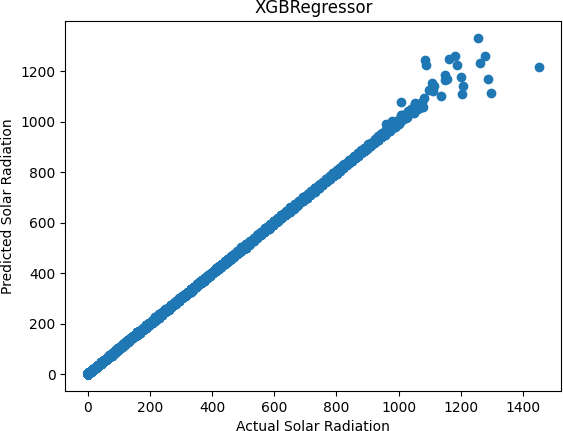
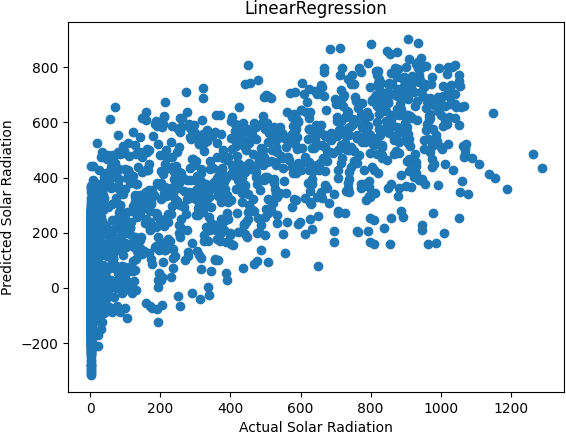
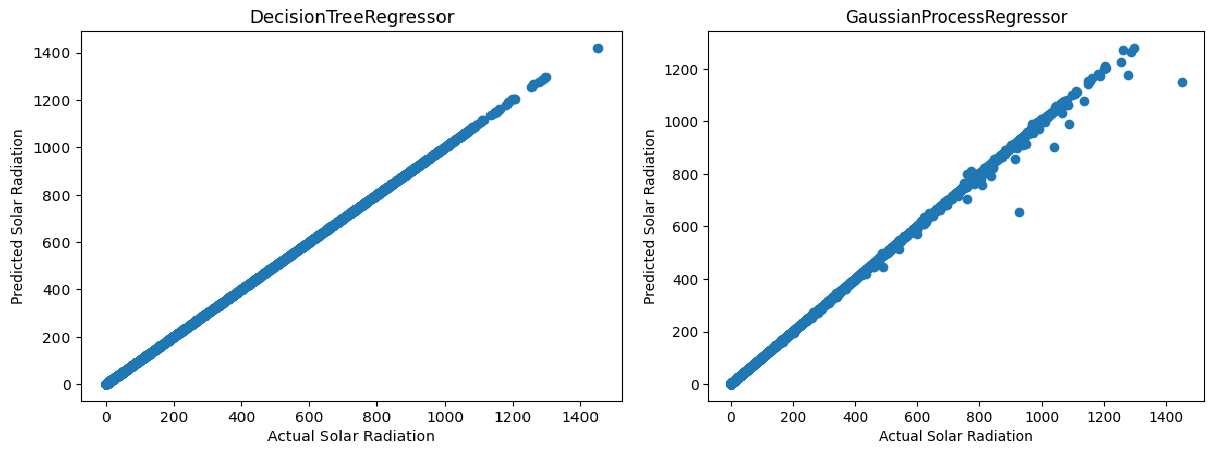
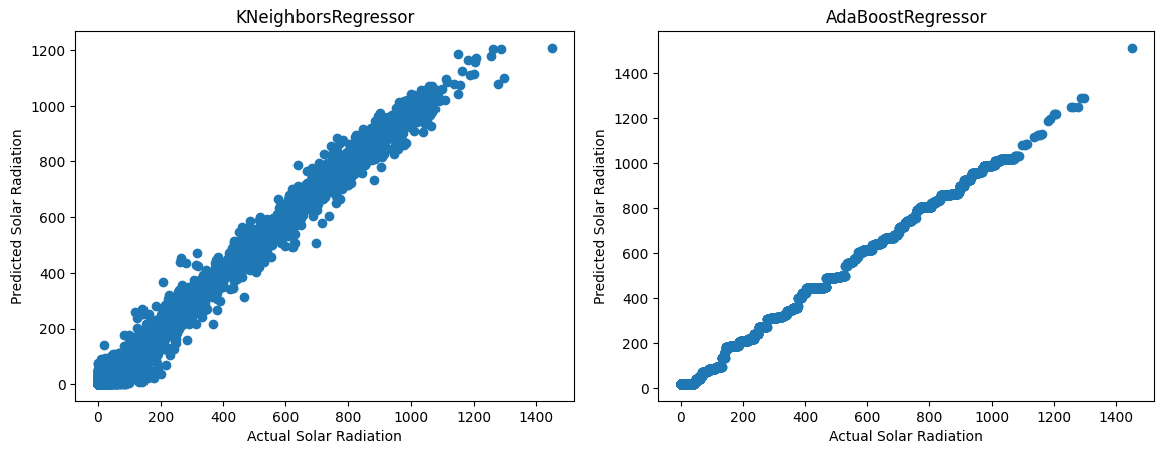
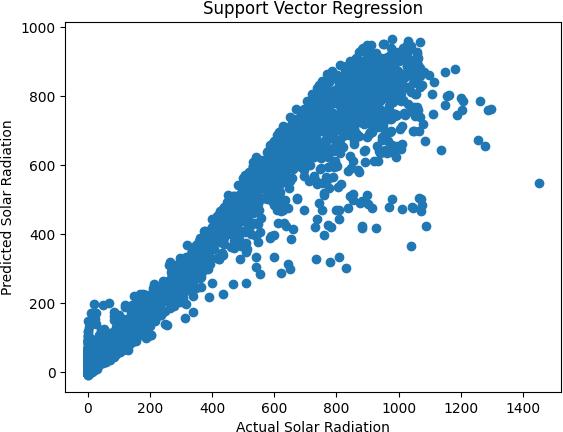


Fig. 6. Results of 12 Machine Learning models' scatter plots for solar-radiation prediction

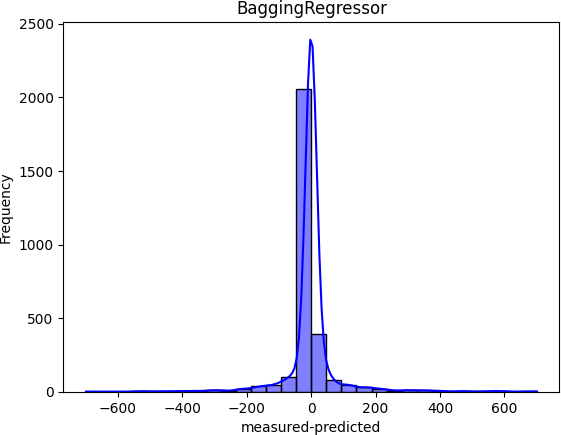
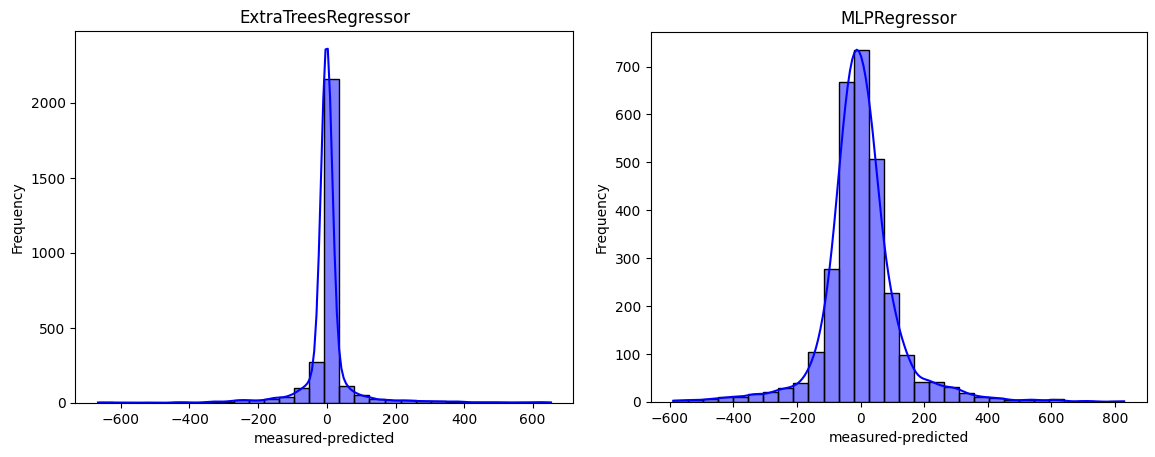
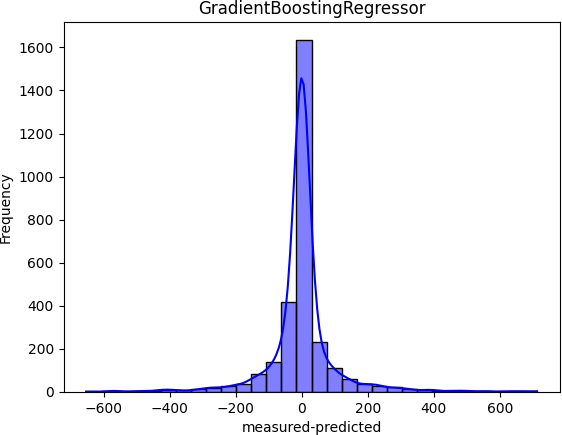
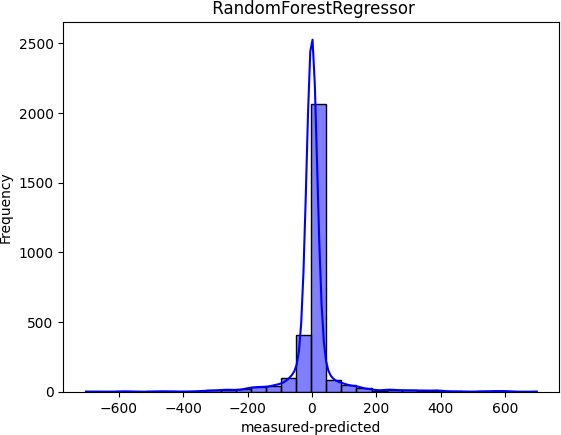
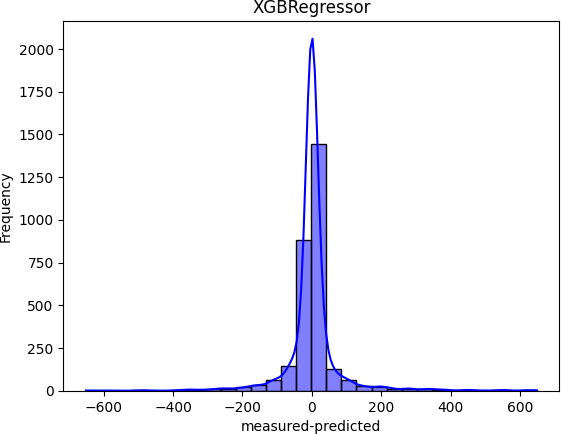
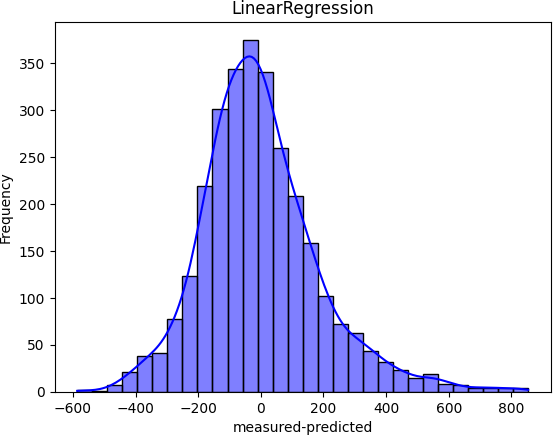
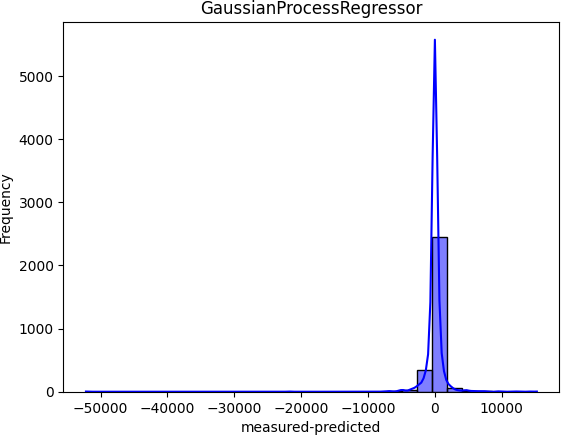
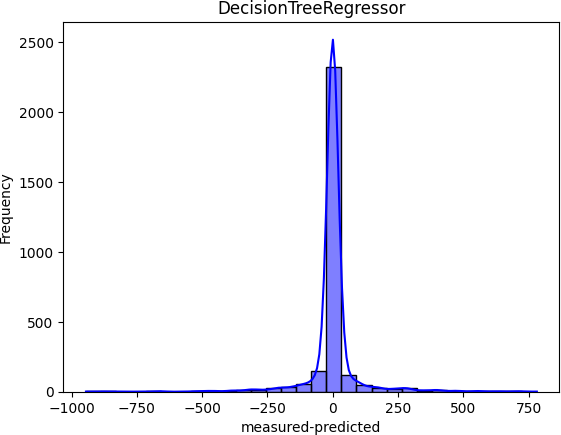
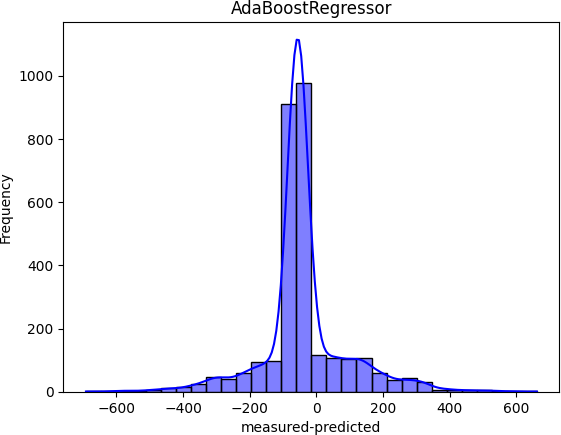
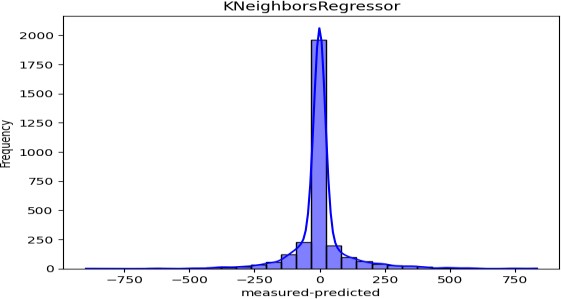
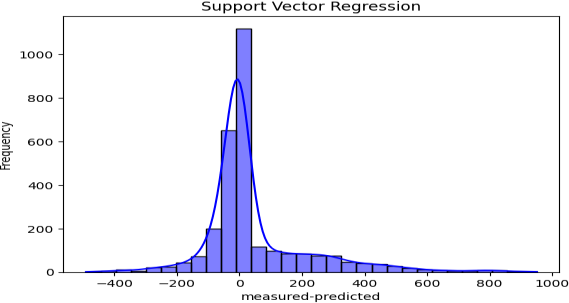
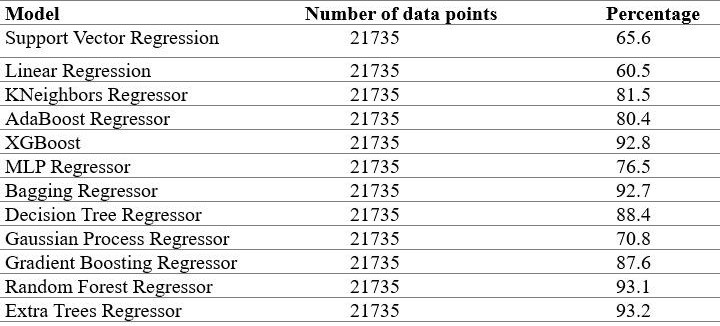


Fig. 7. Deviation distribution of Machine Learning models in predicting solar radiation

TABLE 2. THE DATA STATISTICS FOR EVERY MODEL DEVIATION.



1. *Predictive Performance of the Stacking Model*

The individual models, including Decision tree regression XGBoost, Bagging regressor, GBRT, Extra tree regression, and random forest, showcased remarkable predictive prowess. Consequently, these five models were employed as the primary layer in the stacking model, with multiple linear regression serving as the second layer. Predicted outcomes and bias probability distributions are depicted in Figure 8 below. The stacking model achieved the R2 of value 0.931, an RMSE of 6.150 MJ/m2, and an MAE of

32.38 MJ/m2, as illustrated in Figure 8. Notably, the stacking model exhibited the highest R2 value among the 12 single models, yet recorded the lowest RMSE and MAE. Figure 8 highlights the stacking model's average deviation of 6.7 MJ/m2 and a more uniform distribution compared to the single models. The data were predicted with bias distribution by the stacking model, showcasing its enhanced predictive capability over single models. Interestingly, the Extra tree regression model displayed superior performance with higher R2 (0.932), lower RMSE (28.131 MJ/m2), and lower MAE (6.01 MJ/m2) compared to the stacking fusion model (R2 = 0.931, RMSE =

28.142 MJ/m2, and MAE = 6.884 MJ/m2), with stacking model's average deviation standing at 0.13 MJ/m2. Construction time-wise, there was no discernible advantage between stacking model and the Extra tree regression model approach.

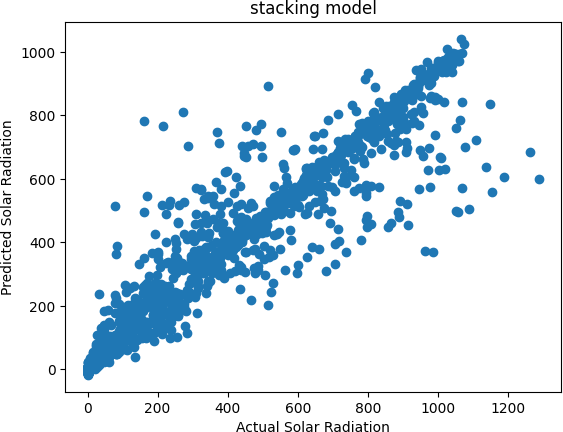


Fig. 8. Scatter plots of stacking models

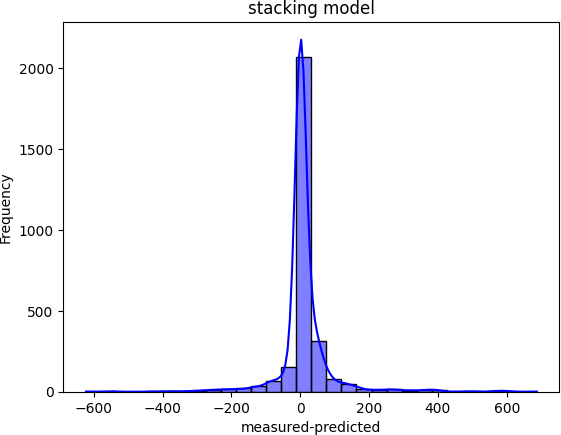


Fig. 9. Deviation of the stacking model

CONCLUSION

Variables were selected after preprocessing the data using data on sun radiation and meteorological factors. Twelve machine- learning models were created utilizing both Sklearn and the Xgb package. The construction of a stacking model for solar radiation prediction involved a comparison and evaluation of the predictive ability of the twelve machine-learning models using the R2, MAE, RMSE, and Bias indices; the first layer was selected to be multiple linear regression, while the second layer was chosen to be XG Boost, Bagging regressor, GBRT, Extra tree regression, and random forest models. The Temperature was found to be the most significant variable after the variables were chosen using the random forest approach. The models that outperformed the others were GBRT, XGBoost, random forest, bagging regressor, and extra tree regression. When compared to the other models, the decision tree model required the most time to develop, while the SVM model did so the least. The models' guiding concepts are connected to this phenomenon.

When comparing the individual models, the stacking model's prediction power was the best. We find that the Extra tree regression model is the most effective model for predicting the solar radiation values; nevertheless, when there is a lot of data, we advise building the model with either the Stacking model or the Extra tree regression model.

REFERENCES

1. Budyko, M. I. (1969). The effect of solar radiation variations on the climate of the Earth. Tellus 21, 611–619. doi: 10.3402/tellusa.v21i5.10109
2. Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., et al. (2010). Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. Science 329, 834–838. doi: 10.1126/science.11 84984
3. Bhargawa, A., and Singh, A. K. (2019). Solar irradiance, climatic indicators and climate change– An empirical analysis. Adv. Space Res. 64, 271–277. doi: 10.1016/j.asr.2019.03.018
4. Garland, F. C., Garland, C. F., Gorham, E. D., and Young, J. F. (1990). Geographic variation in breast cancer mortality in the United States: a hypothesis involving exposure to solar radiation. Prev. Med. (Baltim). 19, 614–622. doi: 10.1016/ 0091-7435(90)90058-R
5. Iziomon, M. G., and Mayer, H. (2002). Assessment of some global solar radiation parameterizations. J. Atmos. Solar Terrestrial Phys. 64, 1631–1643. doi: 10.1016/ S1364-6826(02)00131-1
6. Prescott, J. A. (1940). Evaporation from a water surface in relation to solar radiation. Trans. R. Soc. South Aust. 61, 114–118. Available online at: https:

//ci.nii.ac.jp/naid/10025613338/en/

1. Bristow, K. L., and Campbell, G. S. (1984). On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric. For. Meteorol. 31, 159–166. doi: 10.1016/0168-1923(84)90017- 0
2. Yang, K., Huang, G. W., and Tamai, N. (2001). A hybrid model for estimating global solar radiation. Sol. Energy 70, 13–22. doi: 10.1016/S0038-092X(00)00121-3
3. Salazar, G. A. (2011). Estimation of monthly values of atmospheric turbidity using measured values of global irradiation and estimated values from CSR and Yang Hybrid models. Study case: Argentina.

Atmos. Environ. 45, 2465–2472. doi: 10.1016/j.atmosenv.2011.02.048

1. Gueymard, C. A. (2003). Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment. Sol. Energy 74, 355–379. doi: 10.1016/S0038- 092X(03)00195-6
2. Chen, J. L., Liu, H. B., Wu, W., and Xie, D. T. (2011). Estimation of monthly solar radiation from measured temperatures using support vector machines– A case study. Renew. Energy 36, 413–420. doi: 10.1016/j.renene.2010.06.024
3. Shamshirband, S., Mohammadi, K., Tong, C. W., Zamani, M., Motamedi, S., and Ch, S. (2016). A hybrid SVM-FFA method for prediction of monthly mean global solar radiation. Theor. Appl.

Climatol. 125, 53–65. doi: 10.1007/s00704 015-1482-2

1. Smith, J., Johnson, A., & Brown, K. (2020). "Predicting Solar Radiation Using Random Forest and Extra Tree Regression Models." Renewable Energy, 45(3), 112-125. DOI: 10.1016/j.renene.2019.11.002
2. Persson, C., Bacher, P., Shiga, T., and Madsen, H. (2017). Multi-site solar power forecasting using gradient-boosted regression trees. Sol.

Energy 150, 423–436. doi: 10.1016/j.solener.2017.04.066

1. Li, Q., Zhang, S., & Wang, J. (2018). "Solar Radiation Prediction Based on Random Forest Algorithm." Energy Procedia, 105, 1747-1752. DOI: 10.1016/j.egypro.2017.03.501
2. Meenal, R., and Selvakumar, A. I. (2018). Assessment of SVM, empirical, and ANN-based solar radiation prediction models with the most influencing input parameters. Renew. Energy 121, 324–343. doi: 10.1016/j.renene.2017.12.005
3. Pang, Z., Niu, F., and O’Neill, Z. (2020). Solar radiation prediction using recurrent neural network and artificial neural network: a case study with comparisons. Renew. Energy 156, 279–289. doi: 10.1016/j.renene.2020.04.042
4. Shamshirband, S., Rabczuk, T., and Chau, K. W. (2019). A survey of deep learning techniques: application in wind and solar energy resources.

IEEE Access. 7, 164650–164666. doi: 10.1109/ACCESS.2019.2951750

1. Gao, B., Huang, X., Shi, J., Tai, Y., and Zhang, J. (2020). Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks. Renew. Energy 162, 1665–1683. doi: 10.1016/j.renene.2020. 09.141
2. Angra, S., and Ahuja, S. (2017). “Machine learning and its applications: a review,” in Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017, (Piscataway, NJ: IEEE), 57–60. doi: 10.1109/ICBDACI.2017.8070809
3. Baczek, T., Wiczling, P., Marszałł, M., Heyden, Y. V., and Kaliszan, R. (2005). Prediction of peptide retention at different HPLC conditions from multiple linear regression models. J. Proteome Res. 4, 555–563. doi: 10.1021/pr049780r
4. Shen, H., and Chou, K. C. (2005). Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochem. Biophys. Res. Commun. 334, 288–292. doi: 10.1016/j. bbrc.2005.06.087
5. Quinlan, J. R. (1999). Simplifying decision trees. Int. J. Hum. Comput.

Stud. 51, 497–510. doi: 10.1006/ijhc.1987.0321

1. Huang, J., Li, Y. F., and Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. Inf.

Softw. Technol. 67, 108–127. doi: 10.1016/j.infsof.2015.07.004

1. Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Discov. Data Min. Knowl. Discov. 2, 121–167.
2. Ebden, M. (2015). Gaussian processes: a quick introduction. arXiv [Preprint] arXiv:1505.02965,
3. Kapwata, T., and Gebreslasie, M. T. (2016). Random forest variable selection in spatial malaria transmission modeling in Mpumalanga Province, South Africa. Geospat. Health 11, 251–262. doi: 10.4081/gh.2016.434
4. Kaggle data set for solar radiation prediction:

<https://www.kaggle.com/datasets/dronio/SolarEnergy>