BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 05/06 pp309–317
Volume 5, Number 4, December 2022
DOI: 10.26599/BDMA.2022.9020003

Effect of Feature Selection on the Prediction of Direct Normal Irradiance

Mohamed Khalifa Boutahir*, Yousef Farhaoui, Mourade Azrour, Imad Zeroual, and Ahmad El Allaoui

Abstract: Solar radiation is capable of producing heat, causing chemical reactions, or generating electricity. Thus, the amount of solar radiation at different times of the day must be determined to design and equip all solar systems. Moreover, it is necessary to have a thorough understanding of different solar radiation components, such as Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), and Global Horizontal Irradiance (GHI). Unfortunately, measurements of solar radiation are not easily accessible for the majority of regions on the globe. This paper aims to develop a set of deep learning models through feature importance algorithms to predict the DNI data. The proposed models are based on historical data of meteorological parameters and solar radiation properties in a specific location of the region of Errachidia, Morocco, from January 1, 2017, to December 31, 2019, with an interval of 60 minutes. The findings demonstrated that feature selection approaches play a crucial role in forecasting of solar radiation accurately when compared with the available data.

Key words: machine learning; deep learning; feature importance; renewable energies; solar radiation

1 Introduction

As the primary renewable energy source, a thorough understanding of the various aspects of global solar radiation is critical and extremely beneficial for various applications, including architectural design, agricultural meteorology, weather prediction, climate monitoring, health, and even tourism applications and research[1–4]. Although solar radiation components are one of the most frequently measured meteorological variables, the number of measuring sites remains limited, especially in emerging and underdeveloped countries. Additionally, certain measurements may be incorrect or doubtful due to equipment maintenance and calibration issues[5]. Nevertheless, some excellent alternatives and projects have provided various solar radiation and other meteorological data for various regions worldwide, such as the National Solar Radiation DataBase (NSRDB)[6–8], the Prediction Of Worldwide Energy Resources (POWER)[9], the Copernicus Atmosphere Monitoring Service (CAMS)[10, 11], and the global meteorological database METEONORM[12, 13].

Machine learning and deep learning algorithms have gained significant attention in recent decades for their capacity to analyze arbitrary non-linear connections[14–17]. Numerous methodologies are used to anticipate solar radiation, such us historical solar radiation data, meteorological data, and satellite-derived cloud images. Paoli et al.[18] predicted daily solar radiation time series using artificial neural networks based on pre-processed daily solar radiation time series. The solar radiation estimation model is constructed using a MultiLayer Perceptron (MLP) network, which is the most frequently used architecture in Artificial Neural Networks (ANNs), and the results are compared with those obtained using AutoregRessive (AR), and AutoRegressive Integrated Moving Average (ARIMA). Wang et al.[19] investigated the effect of

• Mohamed Khalifa Boutahir, Yousef Farhaoui, Mourade Azrour, Imad Zeroual, and Ahmad El Allaoui are with the Faculty of Sciences and Techniques, Moulay Ismail University, Errachidia 52000, Morocco. E-mail: moha.boutahir@edu.umi.ac.ma; y.farhaoui@fste.umi.ac.ma; mo.azrour@umi.ac.ma; mr.imadine@gmail.com; hmad666@gmail.com.

* To whom correspondence should be addressed.

Manuscript received: 2021-12-31; accepted: 2022-01-24

weather prediction on solar radiation data using K-Nearest Neighbors (KNN) and Support Vector Machines (SVMs). The results demonstrate that SVMs performs admirably when tiny sample sizes are being used. Munir and Chung[20] suggested a day-ahead solar radiation forecasting method for microgrids that do not rely on past solar irradiance data; instead, it relies on readily accessible meteorological data. They compared the Long Short Term Memory (LSTM) algorithm and the Feed Forward Neural Network (FFNN). Torres-Barrán et al.[21] investigated the solar radiation prediction problem using Gradient Boosted Regression (GBR), eXtreme Gradient Boosting (XGB), and Random Forest Regression (RFR). Both XGB and RFR are measured to outperform SVR. Almaraashi et al.[22] used the ReliefF algorithm, the Monte Carlo Uninformative Variable Elimination (MCUVE) algorithm, the random-frog algorithm, and the Laplacian Score (LS) algorithm as feature selection methods to forecast daily solar radiation levels throughout Saudi Arabia. Fan et al.[23] analyzed the effectiveness of the SVM and four tree-based models for daily horizontal radiation prediction: M5 model Tree (M5Tree), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and gradient Boosting with Categorical features support (CatBoost). Chaibi et al.[24] evaluated the forecasting accuracy of Light Gradient Boosting Machine (LightGBM) to three benchmark models for predicting global solar radiation, including MultiLayer Perceptron (MLP), Multiple Linear Regression (MLR), and Support Vector Regression (SVR).

The rest of the paper is structured as follows: Section 2 defines the components of solar radiation. Section 3 includes charts and figures that illustrate the dataset, the proposed architecture, and the models used for the predictions. The results are reported in Section 4, and the conclusion is provided in Section 5.

2 Background

To comprehend the solar radiation system, it is necessary to recognize the distinctions between its numerous components and properties.

Direct Normal Irradiance (DNI) is the quantity of solar radiation received per unit area by a surface that is always perpendicular (or normal) to the rays coming from the sun. Generally, a surface's yearly irradiance could be optimized by maintaining its normal to the incoming radiation. This amount is particularly relevant for concentrating solar thermal and sun's location track systems[25].

Diffuse Horizontal Irradiance (DHI) is the amount of radiation received per unit area by a surface that is neither shaded nor shadowed and has been scattered by molecules and particles in the environment. It arrives equally from all directions[25].

Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation that a surface horizontal to the ground receives from above[25].

A particularly useful statistic for photovoltaic systems combines both DNI and DHI, as demonstrated in Fig. 1[25]. θ is the zenith angle, which is the incidence angle of direct radiation on the horizontal plane.

3 Models and Methods

3.1 Data source

In this study, data from NSRDB, which was created by the National Renewable Energy Laboratory (NREL), were used.

The NSRDB is the most frequently used publicly available database that offers solar irradiance data from satellites across the globe. As illustrated in Table 1, NSRDB provides broadband irradiance estimation, including GHI, DHI, and DNI, as well as other auxiliary factors, such as sun zenith angle and cloud type[6].

3.2 Data visualization

Data visualization generates a visually summarized version of the data to demonstrate the relationships between data parameters, to study the changes of the different factors, and/or to demonstrate how these

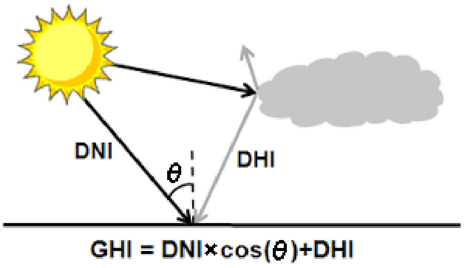


Fig. 1 Equation of calculating GHI using DNI and DHI.

Table 1 Data and parameters.

Data source: NSRDB
Period: January 1, 2017 to December 31, 2019
Time scale: 60 minutes
Parameters: Year, Month, Day, Hour, Minute, Temperature, Clearsky DHI, Clearsky DNI, Clearsky GHI, Cloud type, Dew point, DHI, DNI, Fill flag, GHI, Ozone, Relative humidity, Solar zenith angle, Surface albedo, Pressure, Precipitable water, Wind direction, Wind speed

properties affect the variable under research[26].

The next parts presents data visualization examples using the Matplotlib package in a Google Colab environment using Jupyter Notebook with Python 3.8.

3.2.1 Heat map

The heat map data reveal that the temperature and wind speed are the most critical weather parameters that impact the solar radiation components, particularly our targeted output. Figure 2 presents that a high association exist between the solar radiation components whether they are regular or calculated when the sky is clear.

3.2.2 Line charts

The line chart depicts the average value of the evolution of the solar radiation parameters for the year 2019, grouped by months.

Figure 3 demonstrates that the DNI and GHI increased during the spring season from February to May, reach their maximum values in May (DNI reaching 1055 W/m^2 and GHI reaching 1084 W/m^2). Meanwhile, the DHI, remained quite stable, apart from a slight increase in May.

3.3 Predictions with full data parameters

Multiple machine learning and deep learning algorithms (RF, XGBoost, Decision Trees (DT), and so on) were evaluated by utilizing all the 25 data features in the first experimentation, consequently we got poor accuracy results 72% (as seen in Fig. 4).

3.4 Proposed architecture

As the study's purpose was to estimate the DNI, many models were evaluated to determine which one generates the best results.

As shown in Fig. 5, an experimental strategy architecture was designed to discover factors that enhance the efficacy of solar radiation prediction.

Our architecture comprises two main steps as follows. First, the feature importance for our dataset was computed. A variety of models and tree-based classifiers,

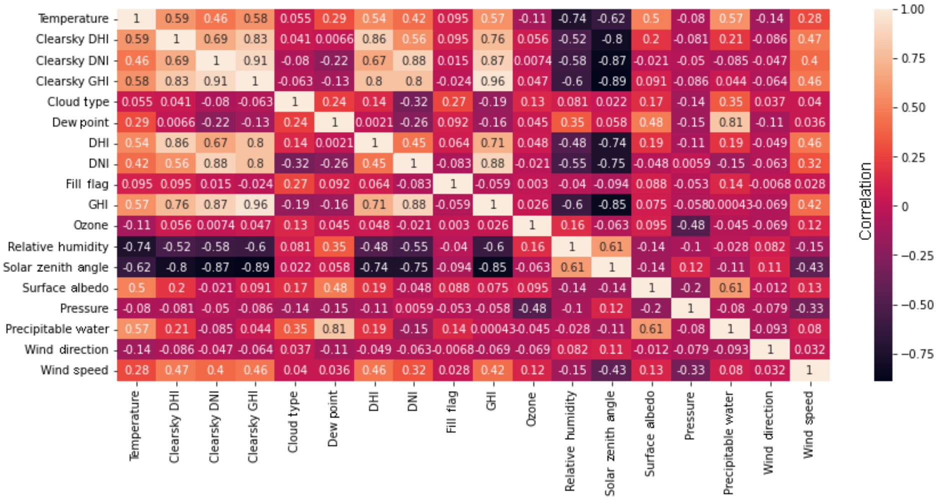


Fig. 2 Heat map data and the correlation between different parameters.

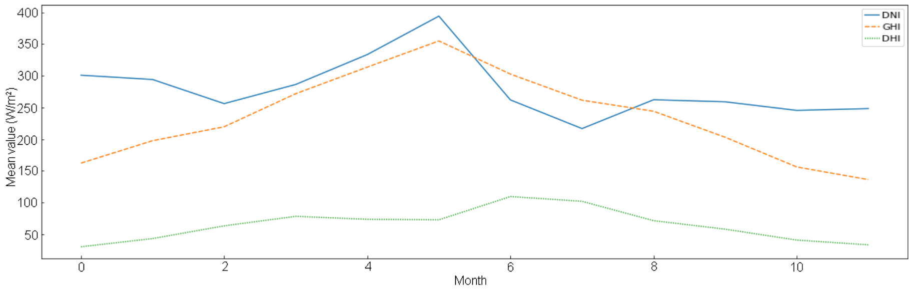
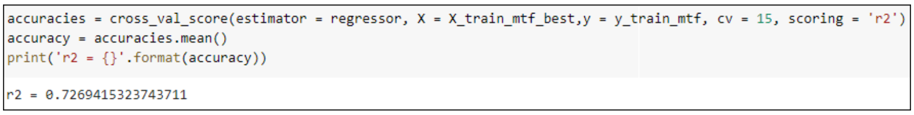


Fig. 3 Development of DNI, GHI, and DHI in 2019 for Errachidia region.



```
accuracies = cross_val_score(estimator = regressor, X = X_train_mtf_best,y = y_train_mtf, cv = 15, scoring = 'r2')
accuracy = accuracies.mean()
print('r2 = {}'.format(accuracy))

r2 = 0.7269415323743711
```

Fig. 4 Accuracy results using the full data features.

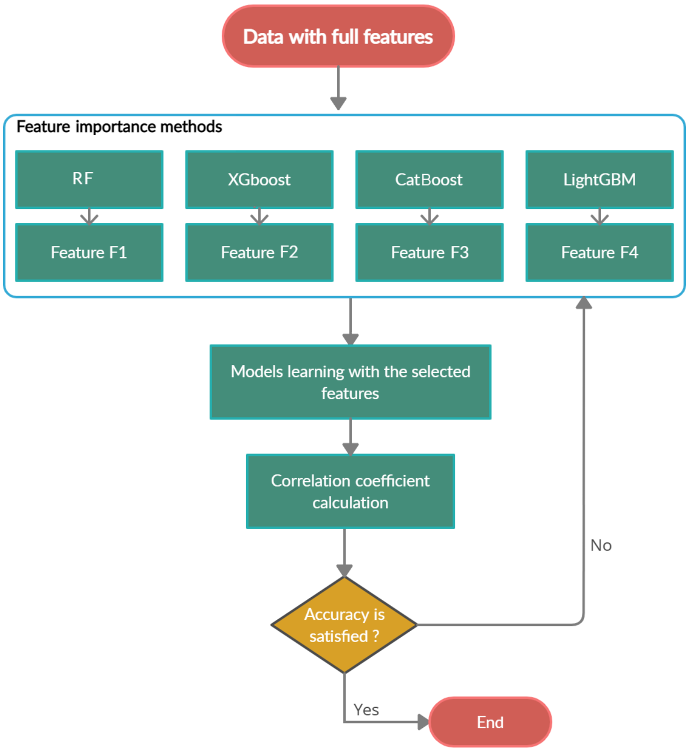


Fig. 5 Proposed architecture for feature selection.

including random forest, XGBoost, CatBoost, and LightGBM, were employed. Second, the effect of the number of features was investigated. In this phase, each ensemble classifier takes different sets of features (obtained in the preceding step from the different classifiers) and integrates them to the model in order to obtain the highest accuracy. If the accuracy is not satisfactory, the first step repeated until the best features are obtained for training our models.

Experiments were conducted on Google Colab using Jupyter Notebook and Python 3.8. The training set was 80%, whereas 20% for the testing set. To analyze the performance of each feature importance method, the results of each model have been expressed in terms of accuracy. Finally, the experimental outcomes were examined.

3.5 Regressors and feature importance

“Feature importance” refers to strategies that assign a value to input parameters on the basis of their predictive power for a target variable[27].

Feature importance scores are critical components of a predictive modelling project because they provide insights into the data, insights into the model, and the foundation for dimensionality reduction and feature selection, which can improve the efficiency and effectiveness of a predictive model on the problem[27].

Such scores are valuable and may be used on various situations in a predictive modelling issue, including the following:

- Understanding the data completely.
- Improving the comprehension of a model.
- Reducing the number of input features.

In this paper, various feature importance algorithms were explored below.

3.5.1 Random forest regressor

This estimator is a built-in component of the RF algorithm that includes a mechanism for calculating significance that may be used in conjunction with the “Gini importance” technique to determine the essential features of data columns. It measures the degree to which each characteristic minimizes the split’s impurity (the feature with the highest decrease is selected)[28].

3.5.2 XGBoost regressor

This regressor is a component of the XGBoost algorithm; the advantage of employing the gradient boosting approach is that it is reasonably easy to extract relevance scores for each attribute after the boosted trees are generated. The more frequent attribute is employed in decision trees to make critical judgements, the greater its relative importance. The significance of the features is then averaged across all the decision trees in the model[27].

3.5.3 CatBoost regressor

This regressor is an integral feature of the CatBoost algorithm. To determine the feature's relevance, CatBoost simply subtracts the metric (loss function) acquired when the model is used in a regular situation (with the feature included) from the metric obtained when the model is used without the feature. The greater the difference, the more important the feature[29].

3.5.4 LightGBM regressor

LightGBM is a robust version of the boosting method that is similar to XGBoost but differs in a few key aspects, such us in the way the tree or base learners are created. Compared with other ensemble strategies, LGBM develops trees leaf by leaf, which minimizes loss throughout the sequential boosting phase[30].

4 Results and Discussions

To evaluate our suggested architecture, the experiment was run in a Google Colab environment using 50 iterations for each method. The average, maximum, minimum, and standard deviations of training and testing errors were computed using various error metrics, including root mean squared error and coefficient of determination R2.

The first step of the architecture determines the most pertinent features. Figures 6–9 illustrate the results. The RF algorithm's results are illustrated in Fig. 6. As demonstrated in Fig. 6, the three important parameters that significantly impact our objective feature are: GHI, DHI, and Clearsky DNI, as well as the solar zenith angle. Figure 7 depicts the findings of the XGBoost regressor,

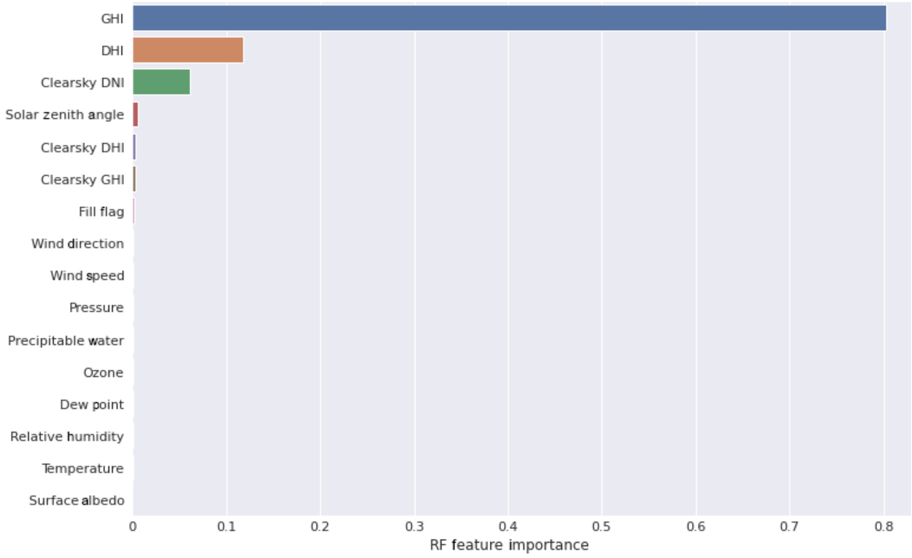


Fig. 6 RF feature importance.

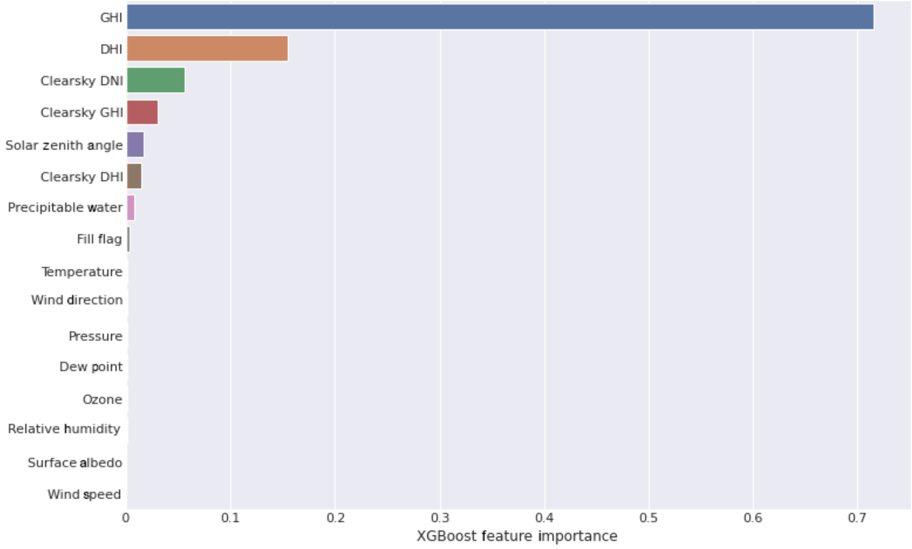


Fig. 7 XGBoost feature importance.

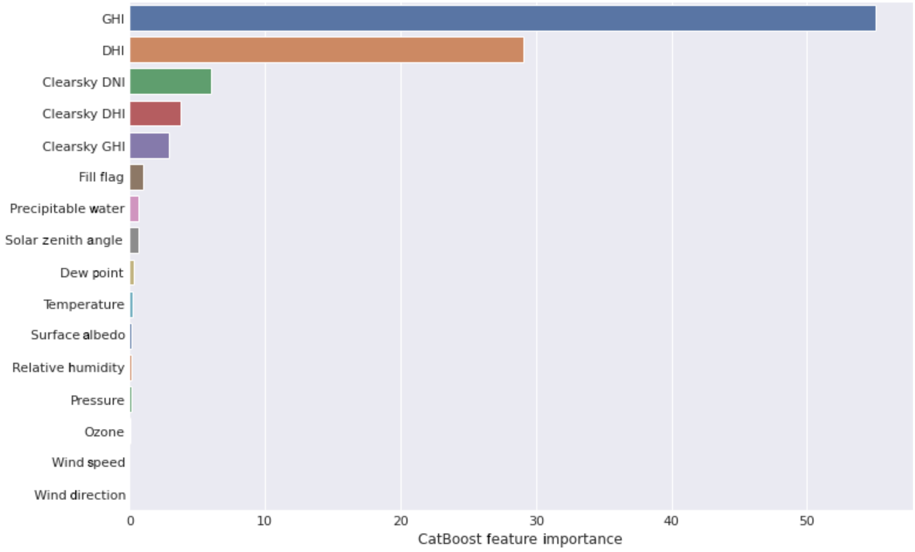


Fig. 8 CatBoost feature importance.

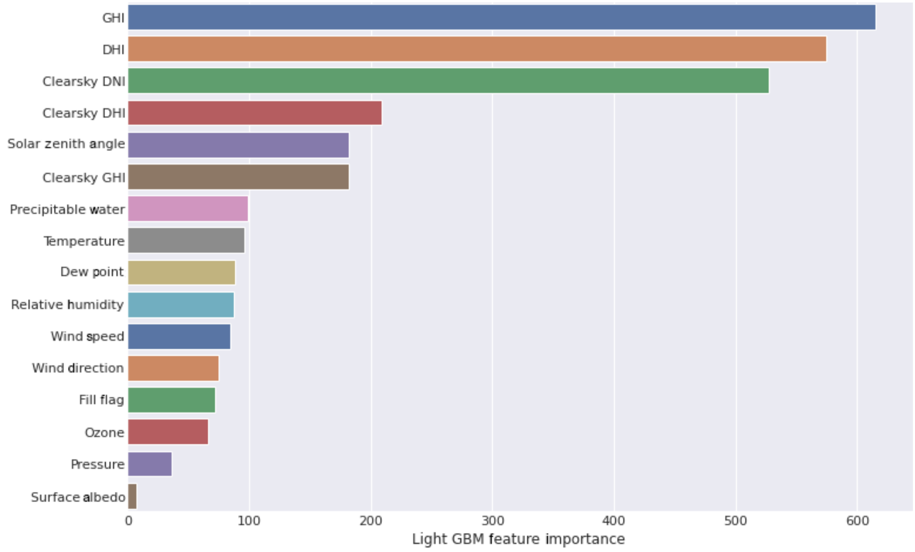


Fig. 9 LightGBM feature importance.

emphasizing the relevance of GHI, DHI, Clearsky DNI, and Clearsky GHI, as well as the solar zenith angle and precipitable water. the CatBoost regressor yields results that are comparable to those of XGBoost, except for the existence of the fill flag, as seen in Fig. 8. Finally, the results obtained using the LightGBM regressor are more convergent than those obtained using other regressors, as illustrated in Fig. 9. The results provide a significant score for several features that are almost completely ignored by other regressors, such as temperature, dew point, and humidity.

For the second step of the architecture, various lists of characteristics obtained in the first step from the various regressors were examined. Then, utilizing those attributes, several models were built for computing the predicted DNI. As seen in Fig. 10, the discrepancy between the observed and anticipated DNI demonstrates that the models produce excellent prediction results.

As previously observed in the correlation map and from the various feature importance methods, a significant correlation and direct influence exist between GHI, DHI and solar zenith angle as a solar radiation characteristic, and temperature as a meteorological with the desired DNI output.

Table 2 summarizes the results of our proposed architecture and demonstrates that the RF regressor achieves the highest accuracy (99.57%) when 11 features are used in an execution time of 3 minutes and 23 seconds, making it the best feature importance method among the evaluated algorithms.

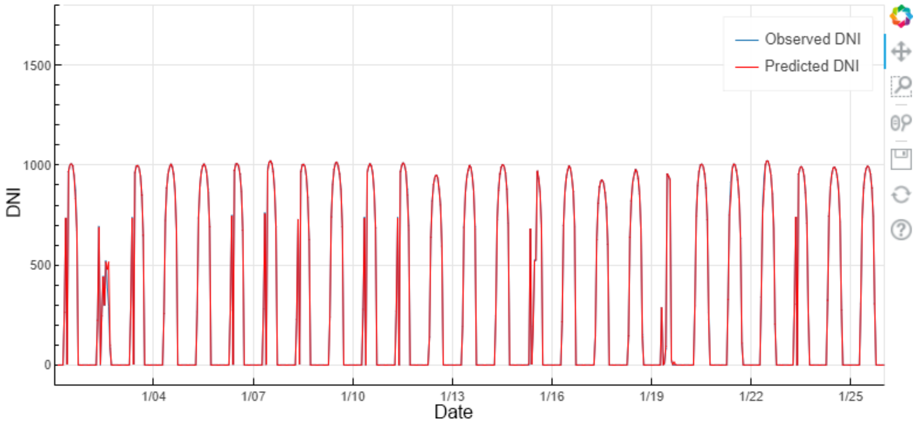


Fig. 10 Difference between observed and predicted DNI in 2019.

Table 2 Results of our proposed architecture.

Model used	Feature importance used	Number of features	Accuracy (%)	Execution time
RF	RF regressor	11	99.57	3 minutes and 23 seconds
XGBoost	XGBoost regressor	13	97.87	6 minutes and 5 seconds
CatBoost	CatBoost regressor	10	98.99	4 minutes and 38 seconds
LightGBM	LightGBM regressor	11	99.08	7 minutes and 15 seconds

5 Conculsion and Future Works

This study compared the performance of four ensemble approaches: RF, XGBoost, CatBoost, and LightGBM. The approaches were utilized to identify the essential variables for predicting DNI. The results reveal that RF achieves the highest accuracy after 50 iterations while 11 features were used. Feature selection techniques can be utilized to exploit the increased correlations between various characteristics related to the desired output without losing prediction quality. Future research may use these findings to identify the amount to which particular data parameters contribute to improve accuracy.

References

[1] M. Oliver and T. Jackson, Energy and economic evaluation of building-integrated photovoltaics, *Energy*, vol. 26, no. 4, pp. 431–439, 2001.

[2] H. Jiang, N. Lu, J. Qin, W. J. Tang, and L. Yao, A deep learning algorithm to estimate hourly global solar radiation from geostationary satellite data, *Renew. Sustain. Energy Rev.*, vol. 114, p. 109327, 2019.

[3] L. Chen, G. J. Yan, T. X. Wang, H. Z. Ren, J. Calbó, J. Zhao, and R. Mckenzie, Estimation of surface shortwave radiation components under all sky conditions: Modeling and sensitivity analysis, *Remote Sen. Environ.*, vol. 123, pp. 457–469, 2012.

[4] P. D. Fu and P. M. Rich, A geometric solar radiation model with applications in agriculture and forestry, *Comput. Electron. Agric.*, vol. 37, nos. 1–3, pp. 25–35, 2002.

[5] K. Kaba, M. Sarıgül, M. Avci, and H. M. Kandırmaz, Estimation of daily global solar radiation using deep learning model, *Energy*, vol. 162, pp. 126–135, 2018.

[6] D. Z. Yang, Validation of the 5-min irradiance from the national solar radiation database (NSRDB), *J. Renew. Sustain. Energy*, vol. 13, no. 1, p. 016101, 2021.

[7] W. Q. Zhang, W. Kleiber, A. R. Florita, B. M. Hodge, and B. Mather, Modeling and simulation of high-frequency solar irradiance, *IEEE J. Photovolt.*, vol. 9, no. 1, pp. 124–131, 2019.

[8] C. A. Gueymard, A. Habte, and M. Sengupta, Reducing uncertainties in large-scale solar resource data: The impact of aerosols, *IEEE J. Photovolt.*, vol. 8, no. 6, pp. 1732–1737, 2018.

[9] O. N. Mensour, S. Bouaddi, B. Abnay, B. Hlimi, and A. Ihlal, Mapping and estimation of monthly global solar irradiation in different zones in Souss-Massa area, Morocco, using artificial neural networks, *Int.J. Photoenergy*, vol. 2017, p. 8547437, 2017.

[10] B. Benamrou, M. Ouardouz, I. Allaouzi, and M. B. Ahmed, A proposed model to forecast hourly global solar irradiation based on satellite derived data, deep learning and machine learning approaches, *J. Ecol. Eng.*, vol. 21, no. 4, pp. 26–28, 2020.

[11] H. Ettayyebi and K. El Himdi, Artificial neural networks for forecasting the 24 hours ahead of global solar irradiance, *AIP Conf. Proc.*, vol. 2056, no. 1, p. 020010, 2018.

[12] M. A. Jallal, A. El Yassini, S. Chabaa, A. Zeroual, and S. Ibnyaich, A deep learning algorithm for solar radiation time series forecasting: A case study of El Kelaa des Sraghna city, *Rev. d'Intell. Artif.*, vol. 34, no. 5, pp. 563–569, 2020.

[13] W. Bendali, I. Saber, B. Bourachdi, M. Boussetta, and Y. Mourad, Deep learning using genetic algorithm optimization for short term solar irradiance forecasting,

in *Proc. 4th Int. Conf. on Intelligent Computing in Data Sciences* (*ICDS*), Fez, Morocco, 2020, pp. 1–8.

[14] Y. Zhou, Y. F. Liu, D. J. Wang, X. J. Liu, and Y. Y. Wang, A review on global solar radiation prediction with machine learning models in a comprehensive perspective, *Energy Convers. Manag*., vol. 235, p. 113960, 2021.

[15] E. D. Obando, S. X. Carvajal, and J. P. Agudelo, Solar radiation prediction using machine learning techniques: A review, *IEEE Latin America Transactions*, vol. 17, no. 4, pp. 684–697, 2019.

[16] Y. Feng, W. P. Hao, H. R. Li, N. B. Cui, D. Z. Gong, and L. L. Gao, Machine learning models to quantify and map daily global solar radiation and photovoltaic power, *Renew. Sustain. Energy Rev*., vol. 118, p. 109393, 2020.

[17] O. Bamisile, A. Oluwasanmi, C. Ejiyi, N. Yimen, S. Obiora, and Q. Huang, Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions, *Int. J. Energy Res*., doi: 10.1002/er.6529.

[18] C. Paoli, C. Voyant, M. Muselli, and M. L. Nivet, Forecasting of preprocessed daily solar radiation time series using neural networks, *Sol. Energy*, vol. 84, no. 12, pp. 2146–2160, 2010.

[19] F. Wang, Z. Zhen, B. Wang, and Z. Q. Mi, Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting, *Appl. Sci*., vol. 8, no. 1, p. 28, 2018.

[20] H. Munir and I. Y. Chung, Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach, *Energies*, vol. 12, no. 10, p. 1856, 2019.

[21] A. Torres-Barrán, Á. Alonso, and J. R. Dorronsoro, Regression tree ensembles for wind energy and solar radiation prediction, *Neurocomputing*, vol. 326–327, pp. 151–160, 2019.

[22] M. Almaraashi, Investigating the impact of feature selection on the prediction of solar radiation in different locations in Saudi Arabia, *Appl. Soft Comput*., vol. 66, pp. 250–263, 2018.

[23] J. L. Fan, X. K. Wang, F. C. Zhang, X. Ma, and L. F. Wu, Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data, *J . Clean. Prod*., vol. 248, p. 119264, 2020.

[24] M. Chaibi, E. M. Benghoulam, L. Tarik, M. Berrada, and A. El Hmaidi, An interpretable machine learning model for daily global solar radiation prediction, *Energies*, vol. 14, no. 21, p. 7367, 2021.

[25] S. Vashishtha, Differentiate between the DNI, DHI and GHI? First Green Consulting, https://cutt.ly/xA8nsUd, 2012.

[26] K. Brush, Data visualization, tech target search business analytics, https://searchbusinessanalytics.techtarget.com/definition/data-visualization, 2020.

[27] J. Brownlee, Feature importance and feature selection with XGBoost in python, Machine Learning Mastery, https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/, 2016.

[28] P. Płoński, Random forest feature importance computed in 3 ways with python, MLJAR, https:// mljar.com/blog/feature-importance-in-random-forest/, 2020.

[29] S. Thiesen, CatBoost regression in 6 minutes: A brief hands-on introduction to CatBoost regression analysis in Python, toward data science, https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329, 2021.

[30] Evaluation of Feature Selection Methods, https://simility.com/wp-content/uploads/2020/07/WP-Feature-Selection.pdf, 2020.

Mohamed Khalifa Boutahir is a PhD candidate at the Engineering Science and Technology Laboratory, IDMS Team, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes. In 2019, he received the MEng degree in intelligent processing systems from Faculty of Sciences of Rabat, Morocco. His research interests include classification algorithms, segmentation, data mining, cloud computing, and big data.

Yousef Farhaoui received the PhD degree in computer security from Ibn Zohr University of Science, Morocco in 2012. He is now a professor at the Faculty of Sciences and Techniques, Moulay Ismail University. His research interests include e-learning, computer security, big data analytics, and business intelligence. He is a member of various international associations. He has authored 4 books and many book chapters with reputed publishers, such as Springer and IGI. He is a reviewer for IEEE, IET, Springer, Inderscience, and Elsevier journals. He is also the guest editor of many journals with Wiley, Springer, Inderscience, etc. He has been the general chair, session chair, and panelist in several conferences.

Mourade Azrour received the PhD degree from Faculty of Sciences and Technologies, Moulay Ismail University, Errachidia, Morocco in 2019, and the MEng degree in computer and distributed systems from Faculty of Sciences, Ibn Zouhr University, Agadir, Morocco in 2014. He currently works as a computer science professor at the Department of Computer Science, Faculty of Sciences and Technologies, Moulay Ismail University. His research interests include authentication protocol, computer security, IoT, and smart systems. He is a scientific committee member of numerous international conferences. He is also a reviewer of various scientific journals, such as *International Journal of Cloud Computing* and *International Journal of Cyber-Security and Digital Forensics* (*IJCSDF*).

Imad Zeroual received the PhD degree in computer science from Mohamed First University, Morocco in 2018. He is currently an assistant professor at the Department of Computer Science, Faculty of Sciences and Techniques, Moulay Ismail University. His research interests are in the fields of artificial intelligence and data science. He focuses in natural language processing, machine learning, information retrieval/extraction, and language teaching and learning.

Ahmad El Allaoui is now an assistant professor at the Department of Computer Science, Faculty of Sciences and Techniques, Moulay Ismail University. He is an IDMS Team member. He focuses in semantic image segmentation, medical imaging, classification algorithms, segmentation, image processing, evolutionary algorithms, and genetic algorithms.