ORIGINAL

Impact of feature selection on the prediction of global horizontal irradiation under ouarzazate city climate

Impacto de la selección de características en la predicción de la irradiación horizontal global bajo el clima de la ciudad de Ouarzazate

Benchikh Salma[1] ✉, Jarou Tarik[1], Lamrani Roa[1], Nasri Elmehdi[1]

[1]Advanced systems engineering laboratory, Ibn Tofail University. Kenitra, Morocco.

ABSTRACT

Ensuring accurate forecasts of Global Horizontal Irradiance (GHI) stands as a pivotal aspect in optimizing the efficient utilization of solar energy resources. Machine learning techniques offer promising prospects for predicting global horizontal irradiance. However, within the realm of machine learning, the importance of feature selection cannot be overestimated, as it is crucial in determining performance and reliability of predictive models. To address this, a comprehensive machine learning algorithm has been developed, leveraging advanced feature importance techniques to forecast GHI data with precision. The proposed models draw upon historical data encompassing solar irradiance characteristics and environmental variables within the Ouarzazate region, Morocco, spanning from 1st January 2018, to 31 December 2018, with readings taken at 60-minute intervals. The findings underscore the profound impact of feature selection on enhancing the predictive capabilities of machine learning models for GHI forecasting. By identifying and prioritizing the most informative features, the models exhibit significantly enhanced accuracy metrics, thereby bolstering the reliability, efficiency, and practical applicability of GHI forecasts. This advancement not only holds promise for optimizing solar energy utilization but also contributes to the broader discourse on leveraging machine learning for renewable energy forecasting and sustainability initiatives.

Keywords: Feature Selection; Machine Learning Algorithms; Solar Irradiation; GHI; Prediction.

RESUMEN

Garantizar previsiones precisas de la Irradiancia Horizontal Global (GHI) es un aspecto fundamental para optimizar la utilización eficiente de los recursos energéticos solares. Las técnicas de aprendizaje automático ofrecen perspectivas prometedoras para predecir la irradiancia horizontal global. Sin embargo, en el ámbito del aprendizaje automático, no se puede sobrestimar la importancia de la selección de características, ya que es crucial para determinar el rendimiento y la fiabilidad de los modelos predictivos. Para abordar este problema, se ha desarrollado un algoritmo de aprendizaje automático integral que aprovecha técnicas avanzadas de importancia de las características para predecir los datos de GHI con precisión. Los modelos propuestos se basan en datos históricos que abarcan características de irradiación solar y variables ambientales dentro de la región de Ouarzazate, Marruecos, que abarcan desde el 1 de enero de 2018 hasta el 31 de diciembre de 2018, con lecturas tomadas a intervalos de 60 minutos. Los hallazgos subrayan el profundo impacto de la selección de características en la mejora de las capacidades predictivas de los modelos de aprendizaje automático para el pronóstico de GHI. Al identificar y priorizar las características más informativas, los modelos muestran métricas de precisión significativamente mejoradas, reforzando así la fiabilidad, la eficiencia y la aplicabilidad práctica de los pronósticos de GHI. Este avance no sólo es prometedor para optimizar la utilización de la energía solar, sino que también contribuye al discurso más amplio sobre el aprovechamiento del aprendizaje automático para la previsión de las energías renovables y las iniciativas de sostenibilidad.

INTRODUCTION

For solar energy systems to be designed, operated, and controlled in the most efficient manner possible,[1,2,3] a precise forecasting of solar irradiance is necessary. One of the most crucial variables in solar irradiance prediction models, global horizontal irradiance (GHI), describes the total quantity of solar radiation that impacts a horizontal surface at a specific point.[4,5] But predicting GHI is difficult since it depends on a complex interplay of several meteorological and environmental variables, including, temperature and humidity.[6,7,8] In recent years, a number of forecasting models have been Introduced to target various prediction time horizons, such as extremely short-term (within an hour), short-term (within a day or day ahead), medium-term (1 month), and long-term (1 year).[9]

Nantian et al.[10] introduce a new feature selection approach for short-term solar irradiance prediction, integrating conditional mutual information (CMI) and Gaussian process regression (GPR). The aim is to assess feature correlations and redundancies using conditional mutual information, and to use GPR with adaptively determined hyper-parameters for prediction. Sequential forward selection (SFS) is used to identify the optimal subset of features and the GPR covariance function. Results demonstrate improved prediction accuracy with reduced feature dimensions, showcasing the efficacy of the proposed method in enhancing solar irradiance prediction performance. Takahiro et al.[11] study explores enhancing solar irradiance forecasting in the Kanto region by integrating the mesoscale ensemble prediction system (MEPS) into support vector regression (SVR)-based predictors. Despite previous challenges in accuracy stemming from numerical weather prediction (NWP), MEPS offers a promising solution. By leveraging multiple network configurations, SVR models were developed using MEPS inputs, resulting in improved accuracy and reduced maximum prediction error for global horizontal irradiance (GHI). The integration of MEPS not only enhances average accuracy but also addresses systematic errors, offering a robust framework for solar irradiance forecasting in the Kanto region. Domingos S et al.[12] introduces HetDS, a heterogeneous ensemble dynamic selection model, designed to forecast solar irradiance with superior accuracy compared to individual models. HetDS selects the most suitable forecasting model from a pool of seven well-established methods, including ARIMA, SVR, MLP, ELM, DBN, RF, and GB. Experimental evaluation using four datasets of hourly solar irradiance measurements in Brazil demonstrates HetDS's superior overall accuracy across five prominent error metrics. By mitigating the risk of selecting an inappropriate model, HetDS enhances system generalization, consistently outperforming standalone methods in nearly all comparisons. Marco et al.[13] study examines the efficacy of exogenous inputs for short-term GHI forecasting, employing various feature selection techniques to identify pertinent variables such as ultraviolet index, cloud cover, and temperature. Five machine learning models, including LSTM, were employed to assess predictive performance, with LSTM exhibiting the highest accuracy (MAD of 24,51 %) for GHI forecasting up to 4 hours ahead. Integration of exogenous inputs notably enhanced forecasting performance beyond 15-minute horizons, reducing errors by over 22 % in 4-hour predictions, yet offering minimal improvements for very short horizons. David Puga-Gil et al.[14] investigated the application of machine learning models, including random forest (RF), support vector machine (SVM), and artificial neural network (ANN), to predict monthly global solar irradiation based on data from six measurement stations in Rias Baixas, Spain. The ANN models exhibited superior performance during both development and testing phases, as well as in extrapolating knowledge to other locations. Achieving mean absolute percentage error (MAPE) values between 3,9 % and 13,8 % for model development and an overall MAPE between 4,1 % and 12,5 % for seven additional locations, ANNs emerged as effective tools for modeling and predicting monthly global solar irradiation in data-rich environments, with potential for extrapolation to nearby areas. Additionally, RF, a widely utilized method for regression and classification tasks, was employed in the study to forecast solar irradiation. Utilizing data from 13 measurement stations in the Rias Baixas area along the Galician coast in Spain, RF models were developed leveraging the available dataset from these stations.

Using meteorological and environmental data,[15] machine learning (ML) algorithms have become a strong tool for estimating solar irradiance.[16] However, the quality and number of input features used in ML-based GHI prediction models might have an impact on their accuracy and reliability.[17] In order to increase the precision and robustness of the prediction models, feature selection is the process of choosing the most pertinent and instructive input features for a specific prediction task.[18] The accuracy and dependability of forecasting solar irradiation can be improved by providing more information on the impact of feature selection on GHI prediction models. This, in turn, may help with resource planning, solar energy system optimization, and grid-scale solar power integration that is sustainable.[19,20] The ability of the proposed method to handle non-linear relationships, perform automatic and interpretable feature selection, and efficiently handle high-dimensional and potentially incomplete datasets makes it an excellent choice for feature selection in global horizontal

irradiation forecasting. While other methods have their own strengths, the comprehensive benefits of the proposed methodology align well with the specific challenges and requirements of accurate global horizontal irradiation (GHI) forecasting.

The total quantity of solar radiation that a surface receives over a specific time period per unit area is known as solar irradiance. The various characteristics and elements of solar irradiation include:

Direct Normal Irradiance (DNI): is the amount of solar energy absorbed by a surface parallel to the trajectory of the sun, which is measured per unit area. In general, maintaining a surface's normal to the incoming radiation can optimize its annual irradiance. This amount is especially important for systems that concentrate solar thermal energy and track the position of the sun.[21]

Direct Horizontal Irradiance (DHI): is the amount of solar radiation that a horizontal surface that is not directly exposed to the sun receives per unit area. It arrives equally from all directions.

Global horizontal irradiance (GHI): is the entire quantity of solar radiation per unit area on a horizontal surface, including both the direct component of solar radiation that comes directly from the sun and the diffuse component of solar radiation that is diffused by the atmosphere. GHI is usually measured in units of watts per square meter (W/m^2).[21,22]
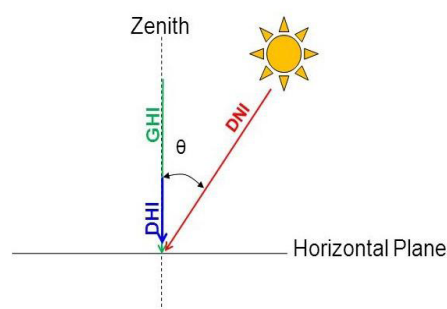

Figure 1. Equation of global horizontal irradiation using DNI and DHI

METHOD

Data Description

Simulations are based on data from The National Solar Radiation Database (NRSDB) developed by National Renewable Energy Laboratory (NREL), The NSRDB is widely used open-access database offering data on solar irradiance through satellites all across the world. The period covered by the database is from 1st January, 2018 to 31 December, 2018, with 1-hour intervals. Table 1 shows all the database variables.

Table 1. Database variables	
Data Source	**NSRDB**
Period	1st January 2018 to 31st December 2019
Time Scale	60 minutes
Variables	Year, month, day, hour, minute, temperature, clearsky DHI, clearsky DNI, clearsky GHI, precipitable water, cloud type, dewpoint, DHI, DNI, fill flag,GHI, ozone, relative humidity, solar zenith angle, surface albedo, pressure, wind direction, wind speed

Proposed Methodology

In machine learning methods, feature selection is frequently employed to select the set of variables most effective in reflecting the source data, minimizing data size and model complexity and increasing prediction precision. It is common practice to test several feature selection techniques and select the variable set that has the highest forecasting performance. As an alternative, an ensemble feature selection can be used by integrating the benefits of various feature selection techniques.

The five basic steps of the suggested method for predicting solar radiation are shown in figure 3. First, data on solar radiation and other meteorological data are collected, considerable database.[23] The data is then cleaned by pre-processing, which includes eliminating anomalies and inputting missing values. The following phase involves selecting the most important variables using an ensemble feature selection method, then, several machine learning techniques are employed after the data has been divided into training, validation, and test sets. Finally, different statistical metrics are employed to rate the forecasting accuracy of system. Experiments were carried out with Python 3.8. The results of each model are presented in terms of accuracy, in order to assess the effectiveness of each feature importance technique. The results of the experiment were then examined.[24]
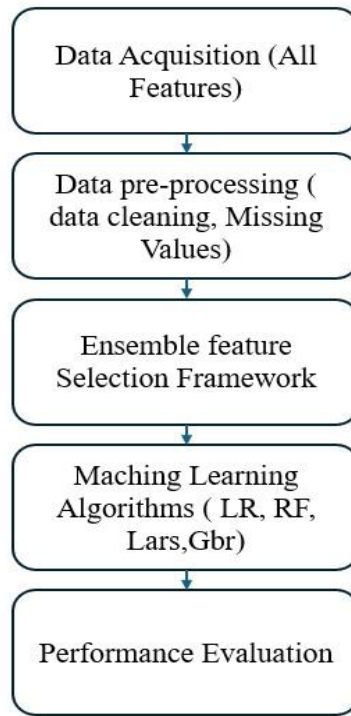
Figure 2. Proposed methodoly flowchart

Heat map

A heat map is a graphical representation of data where the values of a matrix are represented by color gradients. Darker colors generally indicate higher values, while lighter colors reflect lower values. the heat map analysis likely revealed insights into the relationship between GHI, Solar Zenith Angle, and solar radiation components. The DNI and the angle of the solar zenith are the two most important meteorological factors influencing the components of solar radiation, and in particular the anticipated performance, according to the heat map data. When the sky is clear, figure 2 shows that there is a strong correlation between the calculated and regular components of solar radiation.
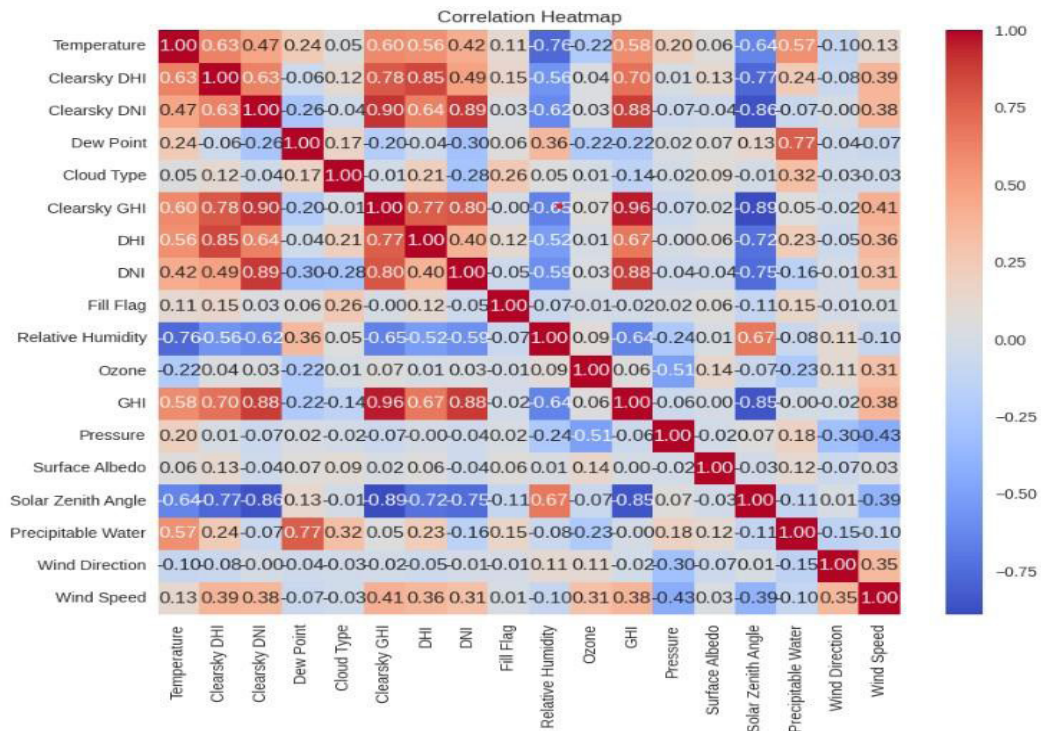

Figure 3. Heat map data and coorelation between variables

Machine learning for feature selection

Feature importance scores are crucial in a predictive modeling approach, as they provide information about the data, information about the model, and a foundation for selecting characteristics and performing dimension reduction. These steps can significantly enhance the effectiveness and efficiency of a predictive model for the given problem. Feature selection is a critical step in the development of accurate models for predicting global horizontal irradiance (GHI). By identifying and using the most relevant variables, the performance of machine learning models can be improved, over-fitting reduced and computational efficiency enhanced. By employing the suggested methodology, the effectiveness of multiple ML systems in predicting solar radiation is assessed. The algorithms evaluated include Random Forest (RF), Linear Regression (LR), Lasso Regression (LARS), and Gradient Boosting Regression (GBR).

Random Forest Regressor is a machine-learning algorithm from the family of models based on decision trees. It is a type of ensemble model which combines several decision trees to produce a more accurate and robust prediction model.[25] The RF regressor is particularly useful for regression problems, where the aim is to forecast a continuous output variable. It can handle both numerical and categorical input features, and is relatively insensitive to outliers and missing values in the data. Additionally, it gives a measure of feature importance, that can be applied for feature selection and to gain insights into the underlying relationships between the input and output variables.

Lasso Least Angle Regression (LARS) is a machine learning approach used for regularization and feature selection. In doing so, it offers effective solutions for high-dimensional datasets by taking into account both the correlation and prediction power of variables concurrently. LARS favors sparse solutions by combining components of least squares estimation and forward stepwise selection. When working with datasets where there are more features than samples, it is very helpful as it reduces overfitting and provides insights into pertinent variables.

Linear Regressor (Lr), is a machine learning algorithm used to forecast a continuous output based on one or more input features[26] that have a linear relationship with the output. It is one of the simplest and most widely used algorithms for regression tasks. The algorithm defines the relationship between the input characteristics and the output variable using a linear equation, which can be represented as $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$, where y is the predicted output variable, $x_1, x_2, ..., x_n$ are the input features, and $b_0, b_1, b_2, ..., b_n$ are the coefficients that the algorithm learns during training.

Gradient boosting regression (GBR), is a machine learning method combines several weak learners, usually decision trees, one after the other to create prediction models. By repeatedly fitting new models to the residuals of the earlier predictions, iteratively minimizing errors. GBR is well-liked for a variety of regression tasks, such as forecasting, classification, and ranking, due to its strong predictive accuracy and resilience against over-fitting. It is a valuable tool in the field of machine learning because of its adaptability, interpretability, and capacity to handle complex correlations in data.

RESULTS AND DISCUSSION

This section displays the results of the suggested methodology for solar forecasting using several ML algorithms.[27] An investigation is additionally conducted into the utility of doing an ensemble feature selection. In this part, four distinct machine learning (ML) methods for estimating sun radiation are compared; Random Forest Regressor, Linear Regressor, Lasso Least Angle Regression, and Gradient Boosting Regression. Algorithm performance is evaluated by means of the coefficient of determination (R^2). It ranges from 0 to 1 and measures the variance in the predictions. A coefficient of 1 implies that the model recognizes observed data precisely, whereas a value of 0 shows that the model's predictions fail badly when applied to unobserved data. The initial phase of the architecture involves identifying the most relevant features. Figures 4–7 showcase the outcomes of this process. Specifically, figure 5 presents the results obtained from the LR algorithm. It reveals that three key parameters—GHI, Clearsky Global Horizontal Irradiance (GNI) Direct Normal Irradiance (DHI), along with the DHI—significantly influence our target feature.

By examining the difference between actual and anticipated values of GHI, the effectiveness of forecasting models can be assessed, highlighting their predictive accuracy. In addition, it highlights the profound influence and interconnection of key variables such as DNI, DHI, zenith angle of the sun and temperature on global horizontal irradiance. These variables, as shown by correlation maps and feature importance techniques, play an essential role in shaping solar radiation patterns. Their collective impact highlights the complex dynamics inherent in solar energy forecasting, underscoring the need for comprehensive modeling approaches that take into account both meteorological conditions and solar radiation characteristics. By recognizing these complex relationships, forecasters can refine their models to deliver more accurate and reliable forecasts, improving the effectiveness and efficiency of solar energy utilization.

The results of the proposed architecture are presented, and show that, of all the algorithms evaluated, the RF regressor is the best approach in terms of feature importance, achieving maximum accuracy (99,95 %) when 11 features are used.
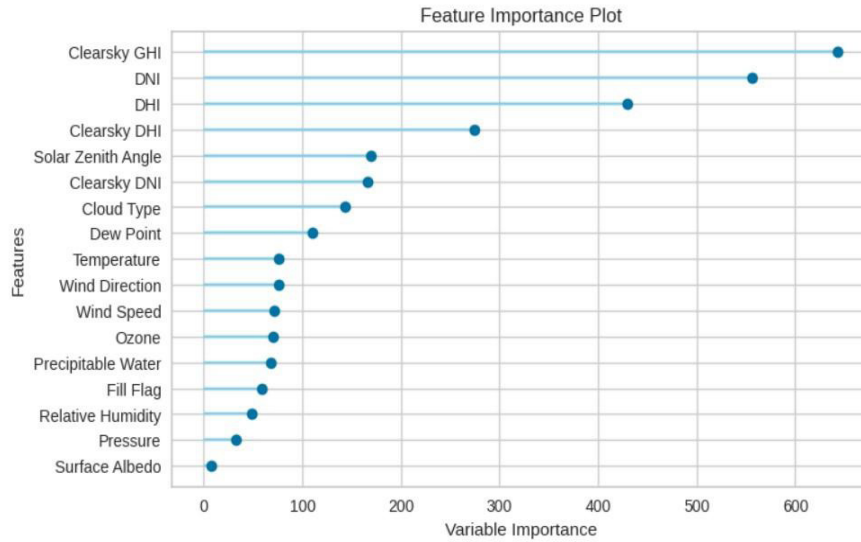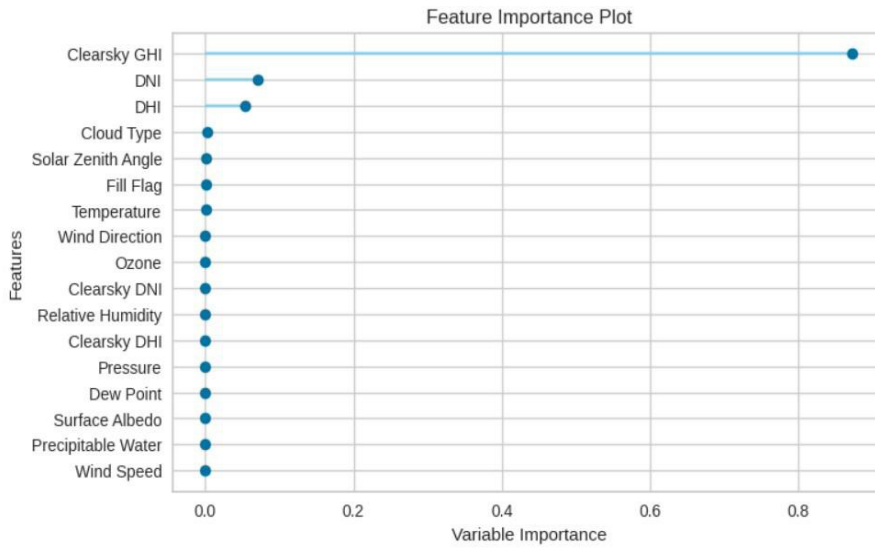
Figure 4. Linear regressor feature importance
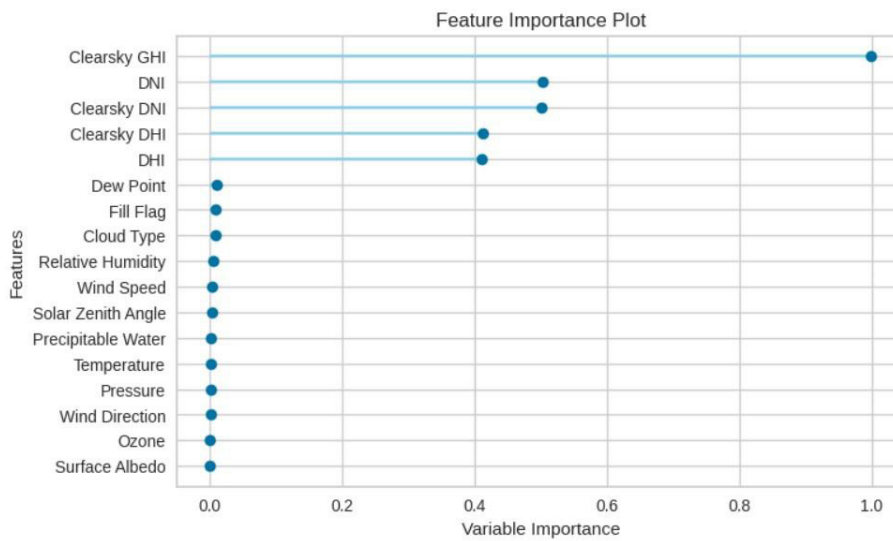

Figure 5. Random forest regressor feature importance

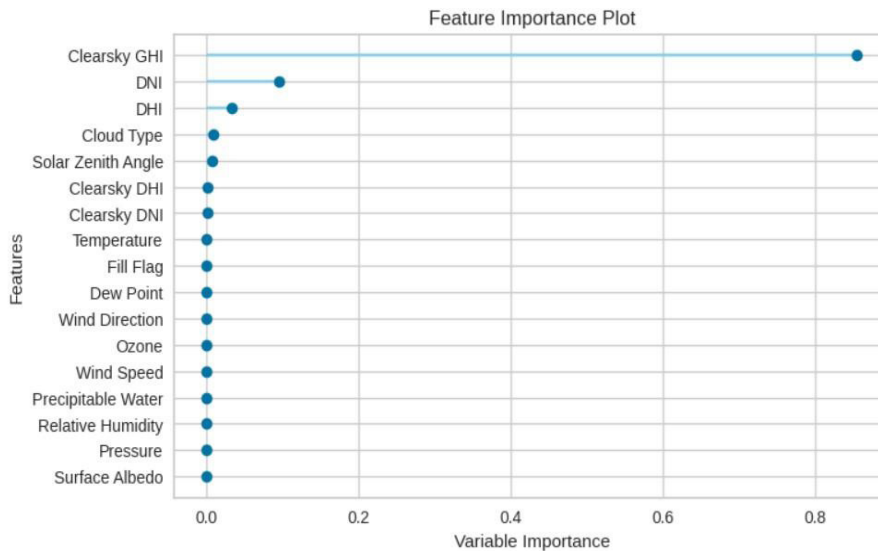**Figure 6.** Lasso Least Angle Regression feature importance

Figure 7. Gradient Boosting Regression feature importance

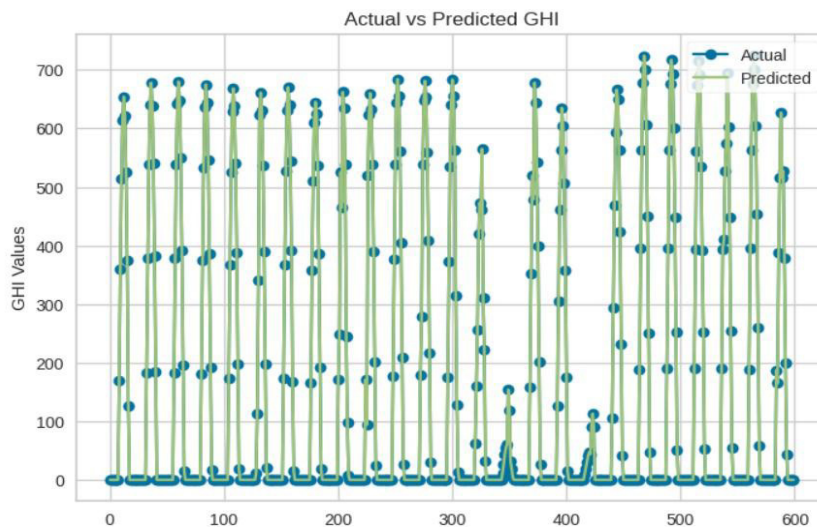| Table 2. Proposed approach results			
Model	**Feature importance used**	**Nombre of features**	**Accuracy (%)**
LR	Linear Regressor	13	98,88
RF	RF Regressor	11	99,95
LLAR	Lasso Least Angle Regression	9	98,87
GBR	Gradient Boosting Regression	10	99,92

Figure 8. Random forest regressor feature importance

CONCLUSIONS

This comprehensive study meticulously assessed the effectiveness of four prominent ensemble methods in forecasting Global Horizontal Irradiance (GHI): Random Forest (RF), Linear Regressor (LR), Lasso Least Angle Regression (LLAR), and Gradient Boosting Regression (GBR). Through rigorous experimentation, we unveiled the remarkable potential of feature selection techniques in optimizing prediction accuracy while maintaining the integrity of the forecast. By strategically identifying and incorporating essential variables, these methods harness the intricate interplay between diverse data attributes crucial for precise GHI prediction. Our findings not only highlight the practical utility of feature selection methodologies but also underscore their pivotal role in advancing renewable energy forecasting applications. By elucidating the intricate relationships between input variables and GHI outcomes, we offer valuable insights for future research endeavors. Subsequent investigations can delve deeper into the nuanced contributions of specific data features towards refining prediction accuracy, thus catalyzing further advancements in renewable energy forecasting models. In essence, this study not only

provides a roadmap for enhancing the performance of GHI prediction models but also contributes to a deeper understanding of the underlying dynamics governing solar irradiance forecasting. Armed with these insights, scientists and practitioners can navigate the complex landscape of renewable energy forecasting with greater precision and efficacy, steering us closer towards a sustainable energy future.

BIBLIOGRAPHIC REFERENCES

1. Boutahir, M.K., Farhaoui, Y., Azrour, M., Zeroual, I., El Allaoui, A.: Effect of feature selection on the prediction of direct normal irradiance. Big Data Mining and Analytics 5 (2022) https://doi.org/10.26599/bdma.2022.9020003

2. BENCHIKH, S., JAROU, T., BOUTAHIR, M.K., NASRI, E., LAMRANI, R.: Improving photovoltaic system performance with artificial neural network control. Data and Metadata 2 (2023) https://doi.org/10.56294/dm2023144

3. Boutahir, M.K., Hessane, A., Lasri, I., Benchikh, S., Farhaoui, Y., Azrour, M.: Dynamic threshold fine-tuning in anomaly severity classification for enhanced solar power optimization. Data and Metadata 2 (2023) https://doi.org/10.56294/dm202394

4. Ono, K., Kunii, M., Honda, Y.: The regional model-based mesoscale ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 147 (2020) https://doi.org/10.1002/qj.3928

5. Liu, C., Li, M., Yu, Y., Wu, Z., Gong, H., Cheng, F.: A review of multitemporal and multispatial scales photovoltaic forecasting methods. IEEE Access 10 (2022) https://doi.org/10.1109/access.2022.3162206

6. Zhou, Y., Liu, Y., Wang, D., Liu, X., Wang, Y.: A review on global solar radiation prediction with machine learning models in a comprehensive perspective. Energy Conversion and Management 235 (2021) https://doi.org/10.1002/qj.3928

7. Rodríguez, F., Azc´arate, I., Vadillo, J., Galarza, A.: Forecasting intra-hour solar photovoltaic energy by assembling wavelet based time-frequency analysis with deep learning neural networks. International Journal of Electrical Power amp; Energy Systems 137 (2022) https://doi.org/10.1016/j.ijepes.2021.107777

8. Solano, E.S., Dehghanian, P., Affonso, C.M.: Solar radiation forecasting using machine learning and ensemble feature selection. Energies 15 (2022) https://doi.org/10.3390/en15197049

9. Trang, T.-T., Ma, T., Do, T.-N.: Lorap: Local deep neural network for solar radiation prediction. Communications in Computer and Information Science (2023) https://doi.org/10.1007/978-981-99-8296-726

10. Huang, N., Li, R., Lin, L., Yu, Z., Cai, G.: Low redundancy feature selection of short term solar irradiance prediction using conditional mutual information and gauss process regression. Sustainability 10 (2018) https://doi.org/10.3390/SU10082889

11. Takamatsu, T., Ohtake, H., Oozeki, T., Nakaegawa, T., Honda, Y., Kazumori, M.: Regional solar irradiance forecast for kanto region by support vector regression using forecast of meso-ensemble prediction system. Energies 14 (2021) https://doi.org/10.3390/en14113245

12. O. Santos, D.S., al.: Solar irradiance forecasting using dynamic ensemble selection. Applied Sciences 12(7) (2022) https://doi.org/10.3390/app12073510

13. Castangia, M., Aliberti, A., Bottaccioli, L., Macii, E., Patti, E.: A compound of feature selection techniques to improve solar radiation forecasting. Expert Systems With Applications 178 (2021) https://doi.org/10.1016/J.ESWA.2021.114979

14. Puga-Gil, D., Astray, G., Barreiro, E.W., G´alvez, J.F., ., J.C.M.: Global solar irradiation modelling and prediction using machine learning models for their potential use in renewable energy applications. Mathematics 10 (2022) https://doi.org/10.3390/math10244746

15. Bamisile, O., Oluwasanmi, A., Ejiyi, C., Yimen, N., Obiora, S., Huang, Q.: Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions. Int. J. Energy Res 46 (2021) https://doi.org/10.1002/er.65295

16. Pombo, D.V., Bindner, H.W., Spataru, S.V., Sørensen, P.E., Bacher, P.I.: Increasing the accuracy of hourly multi-output solar power forecast with physics-informed machine learning. Sensors 2 22 (2022) https://doi.org/10.3390/s22030749

17. Cha, J., Kim, M.K., Lee, S., Kim, K.S.: Investigation of applicability of impact factors to estimate solar irradiance: Comparative analysis using machine learning algorithms. Applied Sciences 11 (2021) https://doi.org/10.3390/app11188533

18. Gb´emou, S., Eynard, J., Thil, S., Guillot, E., Grieu, S.: A comparative study of machine learning-based methods for global horizontal irradiance forecasting. Energies, 14 (2021) https://doi.org/10.3390/en14113192

19. Tovar, M., Robles, M., Rashid, F.: Pv power prediction, using cnn-lstm hybrid neural network model. case of study: Temixco-morelos, m´exico. Energies, 13 (2020) https://doi.org/10.3390/en13246512

20. Wang, Y., Feng, B., Hua, Q.S., Sun, L.: Short-term solar power forecasting: A combined long short-term memory and gaussian process regression method. Sustainability, 13 (2021) https://doi.org/10.3390/su13073665

21. Konstantinou, M., Peratikou, S., Charalambides, A.G.: Solar photovoltaic forecasting of power output using lstm networks. Atmosphere, 12 (2021) https://doi.org/10.3390/atmos12010124

22. Perez-Astudillo, Bachour, D.: Dni, ghi and dhi ground measurements in doha, qatar. Energy Procedia, 49 (2014) https://doi.org/10.1016/j.egypro.2014.03.254

23. Salter, S.H.: Cloud Albedo Enhancement for Solar Radiation Management. Springer, Boston. https://doi.org/10.1007/springerreference308778

24. Takamatsu, T., Ohtake, H., Oozeki, T.: Support vector quantile regression for the post-processing of meso-scale ensemble prediction system data in the kanto region: Solar power forecast reducing overestimation. Energies 15 (2022) https://doi.org/10.3390/en15041330

25. Saha, P., al.: Novel multimodal emotion detection method using electroencephalogram and electrocardiogram signals," biomedical signal processing and control 92 (2024) https://doi.org/10.1016/j.bspc.2024.106002

26. Shemirani, A.B., Lawaf, M.P.: Prediction of tensile strength of concrete using the machine learning methods. Asian Journal of Civil Engineering 92 (2023) https://doi.org/10.1007/s42107-023-00837-5

27. Arai, K. (ed.): Intelligent Computing. Springer. https://doi.org/10.34133/icomputing

FINANCING

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.
Data curation: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.
Formal analysis: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.
Research: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.
Methodology: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.
Drafting - original draft: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.
Writing - proofreading and editing: Benchikh Salma, Jarou Tarik, Lamrani Roa, Nasri Elmehdi.