Optimizing Solar Irradiance Prediction: Feature Selection for All-Sky Image Processing using a Hybrid Prediction Method

Joylan Nunes Maciel, Gustavo Campoi de Souza, Willian Zalewski, Jorge Javier Gimenez Ledesma, and Oswaldo Hideo Ando Junior

***Abstract*—The forecasting of solar irradiance is crucial for photovoltaic solar energy generation, as production is subject to intermittency due to climatic conditions, such as cloud cover, wind and, temperature. Based on the Hybrid Prediction Method (HPM), this study investigated the influence of a set of all-sky image processing features on the HPM's Artificial Neural Network prediction accuracy. Using correlation-based attribute selection, three predictive models with different input feature sets were evaluated. The results show that, when considering all horizons together and paired, the Medium set of 6 features achieves prediction accuracy statistically similar to the Complete set with 9 features, reducing the computational time (14.4%) and model input dimensionality (33.3%). However, when comparing individual horizons, the Complete set outperforms the Medium set at 5- and 15-minute horizon, while maintain similar accuracy at the 1-minute horizon. The Reduced set, with three features, consistently underperformed. This study provides news insights into the optimization of solar irradiance forecasting using HPM, contributing to advances in photovoltaic energy forecasting.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9222

***Index Terms*— all-sky image, solar irradiance prediction, artificial neural network, Hybrid Prediction Method.**

I. Introduction

THE growing energy demand has been increasingly met by renewable sources like wind and photovoltaics, with significant investments in solar energy installations, particularly in China, Japan, the United States, Brazil, Germany, and India [1], [2]. Photovoltaic solar energy adoption has spurred research into methods for predicting solar irradiance, the key factor in energy production [3]. The primary challenge in solar energy generation is the intermittency caused by atmospheric variations, especially cloud cover [4], [5]. Since 2017, there has been a surge in publications on p photovoltaic energy prediction [5], mainly using Machine Learning (ML) and Deep Learning (DL) techniques [4], [6], [7], [8], [9]. Advances in computational power and DL models have further promoted prediction studies using all-sky images [10], [11], [12], [13], [14], [15].

A research group at the Federal University of Latin American Integration (UNILA) has developed scientific studies on the short-term Prediction of Solar Photovoltaic Energy Generation (PSPEG) [16], [17], [18], [19], [20], [21]. Based on this research, the study by [21] proposed and evaluated a new hybrid approach to predicting solar irradiance, applying Image Processing (IP) [22] and Machine Learning techniques [23]. This approach is referred to as Hybrid Prediction Method (HPM). The HPM was developed from a dataset containing historical all-sky image data (180°), meteorological, and solar irradiance information, created with controlled quality and samples collected every minute over a complete period of 3 years (2014 to 2016) [24].

Based on the HPM, the aim of this research is to investigate the set of features (characteristics) extracted from all-sky images, proposed in the HPM, in order to verify the influence of different combinations of these features in predicting short-term solar irradiance (1, 5 and 15 minutes), using the Artificial Neural Network (ANN) model [23] adopted in the HPM. The hypothesis to be investigated is that an optimized set of these features can provide predictions with similar accuracy to the use of all features proposed in the HPM, thus reducing complexity and computational processing.

A theoretical literature review on concepts and works related to PSPEG, is presented in Section 2. The material and methods area described in Section 3. Section 4 presents and discusses the experimental results. The conclusion, contributions and, limitations are outlined in Section 5.

II. Theoretical Background

The growing demand for clean, so-called renewable energy is driving investment in sources such as photovoltaic solar energy. However, the intermittency of energy generation, caused mainly by variations in cloud cover [25], requires accurate methods of predicting solar irradiance [26]. In this context, Prediction of Solar Photovoltaic Energy Generation (PSPEG) plays a crucial role in optimizing photovoltaic system operations. Its enables adjustments in load distribution to maintain grid stability, facilitate real-time management through the activation of backup resources, promotes early trading in the energy market, and allows for more reliably battery sizing, thereby enhancing both operational and economic performance of photovoltaic systems [27], [28].

The associate editor coordinating the review of this manuscript and approving it for publication was Ruth Aguilar (*Corresponding author: Joylan Nunes Maciel*).

J. N. Maciel, G. C. de Souza, W. Zalewski, and J. J. D. Ledesma are with Federal University of Latin American Integration, Foz do Iguaçu, Brazil (e-mails: joylan.maciel@unila.edu.br, gustavocampoi17@gmail.com, willian.zalewski@unila.edu.br, and jorge.ledesma@unila.edu.br).

O. H. Ando Junior is with Federal Rural University of Pernambuco, Cabo de Santo Agostinho, Brazil (e-mail: oswaldo.ando@ufrpe.br).

The accuracy and precision of the methods have progressed significantly over the years, driven by the adoption of Artificial Intelligence, especially Machine Learning (ML) and Deep Learning (DL) [7], [9], [29], [30], [31]. The term accuracy refers to the closeness to the real values (deviations or bias) that are already known, while precision refers to the standard deviation, or variability in the repetition of predictions [32].

The last decade has witnessed a growing volume of research into PSPEG, especially since 2017 [5]. Despite the lack of a universal consensus on the classification of these studies, some characteristics clearly distinguish them [21], [27]: Predicted horizon - refers to the future time period for estimating solar energy generation. Shorter time frames are keys for real-time operations in photovoltaic systems [28]; Source and data type - data for training and testing models come from sources like weather stations, irradiance sensors and images [27]; Prediction algorithm class - includes different types of predictive algorithms, such as statistical, physical, Artificial Intelligence-based and hybrid models [27], and; Error measures - metrics used to evaluate prediction model performance, like the Coefficient of Determination (R^2) and (normalized) Root Mean Square Error ((n)RMSE). Each metric highlights different aspects of accuracy and prediction errors [33].

Most of the recent PSPEG studies have adopted methods that use deep learning [4], [28]. However, a recent study proposed and evaluated a new hybrid approach to predicting solar irradiance for application in photovoltaic power generation. Called the Hybrid Prediction Method (HPM), the method proposed in [21] jointly applies Image Processing and Machine Learning techniques to make short-term predictions of solar irradiance at horizons of 1, 5, 15, 30 and 60 minutes. HPM has been evaluated using different performance metrics. However, this study adopted the following accuracy metrics: the Coefficient of Determination (R^2), and Root Mean Square Error (RMSE) (equations 1 and 2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(o_i - p_i)^2}{\sum_{i=1}^{N}(o_i - \mu)^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(o_i - p_i)^2} \quad (2)$$

In above measures, N represents the total number of predicted points, μ refers to the mean of the observed distribution, p_i to the i_{th} predicted point and o_i represents the i_{th} observed point. Comparing the accuracy of models with different sample sizes is done by normalizing the features to the mean μ of the observed data, where $nRMSE = RMSE/\mu$ [26].The Determination Coefficient (R^2) measures the quality and accuracy of the predictive model's fit to the original data. MBE captures overall positive or negative trends in the errors between predicted and actual data; but positive and negative errors can offset each other, potentially masking the true error magnitude. RMSE is the most widely used metric in the PSPEG literature [34], [35], because it is scale-dependent and sensitive to large individual errors due to the squaring of residuals.

Fig. 1 shows the data and execution flow of the HPM with the two main stages: (i) Explicit Features Image Extraction which applies Image Processing (IP) and (ii) Prediction Processing which applies Machine Learning (ML).

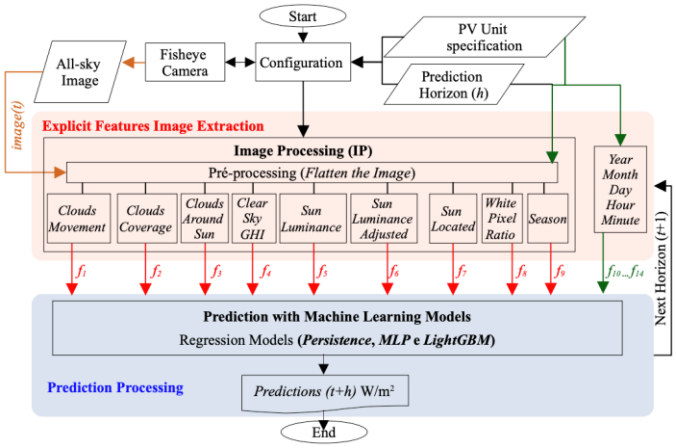


Fig. 1. Data flow diagram and execution of the Hybrid Prediction Method (HPM). Source: Adapted from [21].

In the diagram, the all-sky images are pre-processed in the IP stage, which extracts a set of features (f_1 to f_9) regarding the characteristics of each image. The features and their respective percentage of computing time within the HPM as are follows: *Clouds Movement* (4.0%), *Clouds Coverage* (5.1%), *Clouds Around Sun* (14.5%), *Clear Sky GHI* (41.5%), *Sun Luminance* (10.2%), *Sun Luminance Adjusted* (10.2%), *Sun Located* (9.8%), *White Pixel Ratio* (4.6%), and *Season* (0.1%). All these features are used as inputs for an Artificial Neural Network (ANN) model adopted in the HPM. Therefore, these input features (f_1 ... f_9), shown in Fig. 1, will be evaluated to investigate their influence on the prediction of HPM solar irradiance, specifically at the 1, 5 and 15-minute horizons.

III. Materials and Methods

This section presents the experimental methods used to analyze the set of features provided in the HPM [21]. The methodological procedure, dataset, tools and technologies used are detailed. According to [36], this is an applied study with the aim of generating new knowledge and optimizing a practical method for predicting short-term solar irradiance. The problem is approached in a qualitative-quantitative way, applying statistical analysis and addressing the cost-benefit of using the features in the HPM. The study is exploratory, as it provides greater familiarity with the problem, including the analysis of examples that stimulate understanding. The techniques employed are bibliographical research and experimental statistical analysis, with a case study.

The HPM was developed using a large, standardized, quality-controlled dataset containing historical information on solar irradiance and all-sky images, among others. Collected in the city of Folsom/CA/USA, the data covers the complete period of three (3) years (2014-2016) with samples available every minute [24]. After the Feature Extraction stage, with the Image Processing (IP) techniques shown in Fig. 1, the set of

features (*f*) was computed for days 13 to 15/03/2016, and exemplified in Fig. 2, where Folsom GHI (target) is the target variable to be predicted.

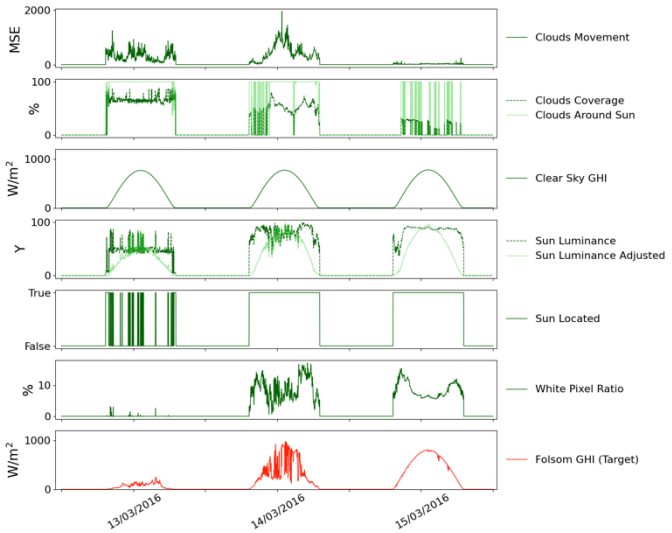


Fig. 2. Example of HPM image processing features and solar irradiance (GHI). Source: [21].

HPM uses all the features shown in Fig. 2 as input to make predictions, using an ML model based on Artificial Neural Networks (ANN) of the MultiLayer Perceptron (MLP) type [21]. Except for the year, month, day, hour and minute, there is the complete set of nine (9) features used in the training, validation and testing.

Several techniques can be applied to attribute selection in regression problems, such as Correlation, Mutual Information, Feature Importance and Information Gain [37]. This study applied Correlation Selection, using Pearson's Correlation Coefficient, which summarizes the degree of relationship between two variables [38]. Based on the correlation matrix in Fig. 3, ordered by the decreasing absolute value of correlations between the features and the Folsom GHI variable, three input sets were defined to evaluated their impact on short-term solar irradiance prediction accuracy (1, 5 and 15 minutes) using the ANN model used in [21]. These sets are categorized as Complete, Medium and Reduced, with their corresponding processing times expressed as percentages, as detailed in Table I.

TABLE I
DETAILS OF THE THREE INPUT SETS AND PROCESSING TIMES

Input Set	Image Processing Features from HPM	Overall Computing Time
Complete 9 variables	Clouds Movement, Clouds Around Sun, Sun Luminance Adjusted, Clear Sky GHI, Clouds Coverage, Sun Luminance Season, Sun Located, White Pixel Ratio	100%
Medium 6 variables	Clouds Movement, Clouds Around Sun, Sun Luminance Adjusted, Clear Sky GHI, Clouds Coverage, Sun Luminance	85.6%
Reduced 3 variables	Clouds Movement, Clouds Around Sun, Sun Luminance Adjusted	28.7%

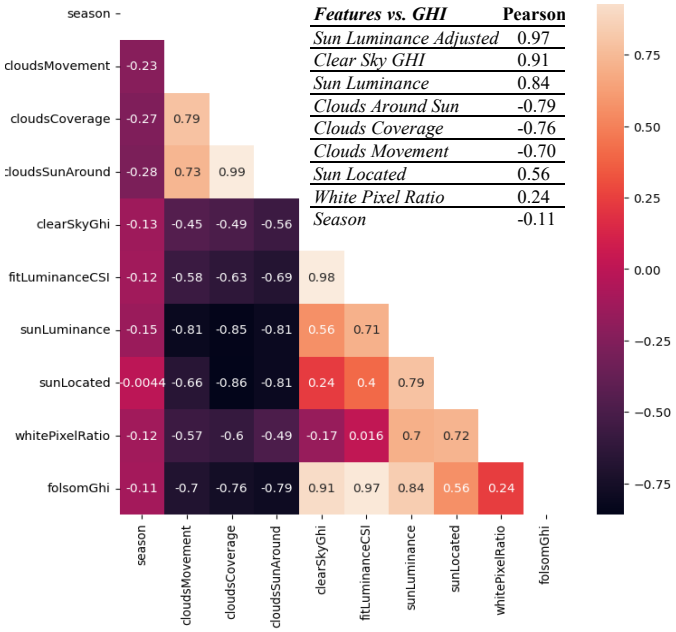


Fig. 3. Pearson correlation matrix of all features.

Based on the principle that the Sun Luminance Adjusted feature originates from the Clear Sky GHI and Sun Luminance features [21], these two features were replaced with the next most correlated features (Clouds Movement and Clouds Around Sun) into the Reduced Set. The final composition of these sets (Table I) was used to construct, train, validate and test the Artificial Neural Network (ANN) predictive model from HPM. Thus, the ANN-MLP configuration in this study is the same as that adopted in the HPM, and is detailed in the Table II.

TABLE II
HYPERPARAMETERS OF THE ANN [21]

Total neurons = 402	Hidden layers = 3
Activation Function = ReLU	Batch Size = 128
Neurons in output layer (y) = 1	Learn Rate = 0.001
Neurons in input layer (x) = 14, 11 or 8	Epochs = 200
Structure = $\{x + hidden(s)\_layer(s) + y\}$ = $\{x + 30 + 210 + 150 + 1\}$	

In addition, Table III shows the *k*-fold cross-validation approach [39] adopted for training, validation with 3 partitions or k=3 [21]. From the total period of three years of information available in the dataset, two years are used to train the models and, one full year (33.3% of the dataset) was used to test the model. This strategy was adopted to enable the model to capture the annual seasonality of solar irradiance more effectively.

TABLE III
SPLITTING THE TRAINING, VALIDATION AND TEST DATASET

Partition	2014	2015	2016
fold-1	Test	Train	Train
fold-2	Train	Test	Train
fold-3	Train	Train	Test

The validation data are 20% (4.8 months) of train partition in each fold.

Based on these definitions, the experimental analysis method, shown in Fig. 4, was designed and executed in the Google Colab [40], using the Python programming language [41] [41]and the TensorFlow, Keras, Numpy, Scikit-Learn libraries [42] to build, train, validate and test the ANN models (Table II), using the cross-validation approach [43] (Table III).

Therefore, three predictive models were created for each input set (one for each fold), using steps 1 to 8 in Fig. 4. At the end of the experiments, the average of the three folds was computed for each input set. Next, an analysis and comparison of the prediction accuracies between different input sets was carried out, using the MSE, RMSE and R^2 metrics, between the ANNs with different input sets: Complete, Medium and Reduced. The results and discussion of these comparisons are presented in the next section.

IV. Results and Discussion

After performing the experimental analysis, the results were collected and are summarized for all the sets, horizons and folds shown in Table IV, including the average and standard deviation (SD) for each combination.

Although the study focuses on comparing ANNs with different input sets, additional results are notable. As shown in Fig. 5, the Reduced set models exhibit the highest error variability (standard deviation). Across all evaluated horizons (1, 5 and 15 minutes), the Reduced set consistently produced the largest average prediction error, reinforcing that a greater number of input variables enhances prediction accuracy. This demonstrates that incorporating more image processing features enables the ANN models to better capture intermittencies with improved accuracy and precision.

In general, the RMSE values in the Reduced Set were higher than the other sets in all three horizons, as shown in Fig. 6-a. This indicates that the set's three input features (*Clouds Movement, Clear Sky GHI, Sun Luminance Adjusted*), despite having a correlation of over 70% with the Folsom GHI feature, did not provide sufficient support for the HPM ANN to capture the variability and maintain the prediction performance of the other input sets (Medium and Complete). The boxplot diagram (Fig. 6-b) demonstrates the greater variability and lower accuracy of the Reduced Set predictions.

On the other hand, the accuracies of the Medium and Complete sets were similar over the three horizons evaluated (Fig. 6-a), especially when considering variability (standard deviation). Therefore, the Medium Set with six (6) features, three less than the Complete Set (reduction of 33.3%), made it possible to make predictions with similar accuracy to the use of all the 9 features in the Complete Set, adopted in the HPM [21].

The accuracy of a regression prediction model based on Machine Learning is influenced by the dataset used [23]. In this study, we clarify that the aim was not to develop an optimized model with fine tuning of the ANN's hyperparameters [23], but to evaluate the influence of these features on the performance accuracy of the HPM model [21]. Therefore, the same structure and configuration of the MultiLayer Perceptron model adopted in the HPM [21], was applied in this study.

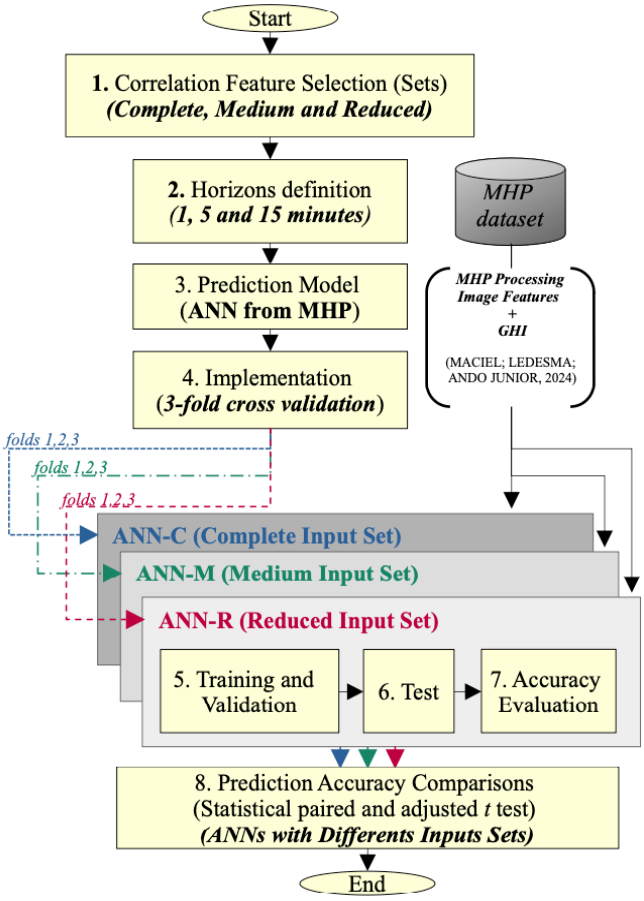


Fig. 4. Experimental method adopted in this study.

The visual similarity of the results (Fig. 6-a) prompted the application of a statistical test to validate the overall performance comparison of models with different input sets, as well as pairwise comparisons at equal prediction horizons. The Null Hypothesis ($H0$) stated that the average prediction accuracy (RMSE) between model pairs is statistically equal ($\alpha = 0.05$). The Shapiro-Wilk test [38] confirmed that the data follow a normal distribution. Thus, to validate *H0*, the Nadeau and Bengio corrected paired t-test was selected, as it is commonly used in Machine Learning model comparisons and accounts for the dependence of training data generated by cross-validation with 3-folds [43]. The analysis was conducted for each input set separately, across individual forecast horizons (1, 5 and 15 minutes) and for all horizons combined.

Considering all the horizons together, the models with different input sets were compared. The results indicates no significant difference between the Complete and Medium sets (Table V). However, the Reduced Set models show a statistically significant difference compared to the others (p-value < 0.05), regardless of the prediction horizon. Thus, the overall performance of the Reduced Set is inferior to the other models.

TABLE IV
PREDICTION ACCURACY OF HPM-ANN MODELS (3 INPUT SET)

Input	Horizon	*Fold*	RMSE	nRMSE	R^2
Complete Set (ANN-C)	1	1	74.00	0.054	0.938
		2	75.12	0.059	0.934
		3	74.66	0.051	0.937
	5	1	74.30	0.054	0.937
		2	73.69	0.058	0.937
		3	76.20	0.052	0.934
	15	1	85.21	0.062	0.918
		2	79.21	0.063	0.927
		3	84.77	0.058	0.919
Medium Set (ANN-M)	1	1	77.28	0.057	0.932
		2	74.48	0.059	0.935
		3	81.26	0.055	0.925
	5	1	80.26	0.059	0.927
		2	77.99	0.062	0.929
		3	81.18	0.055	0.925
	15	1	90.89	0.067	0.906
		2	80.25	0.063	0.925
		3	84.73	0.058	0.919
Reduced Set (ANN-R)	1	1	88.93	0.065	0.910
		2	88.50	0.070	0.909
		3	91.24	0.062	0.906
	5	1	93.11	0.068	0.902
		2	90.51	0.071	0.904
		3	96.64	0.066	0.894
	15	1	112.55	0.082	0.856
		2	93.04	0.073	0.899
		3	101.15	0.069	0.884

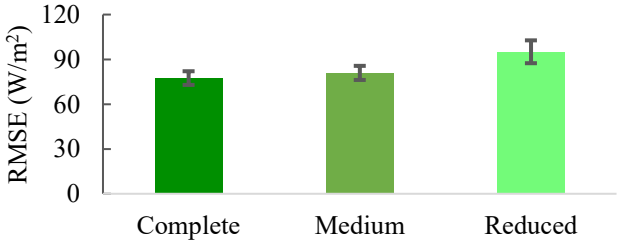


Fig. 5. Average accuracy overall per evaluated input set.

A detailed analysis for each prediction horizon shows that for the 1- and 5-minutes horizons, all comparisons between the Complete, Medium and Reduced sets show significant differences, except between Complete and Medium for the 1-minute horizon. For the 15-minute, significant differences are observed between Complete and Medium, and between Complete and Reduced, but not between Medium and Reduced (Table V). The lack of significant difference between Medium and Reduced at 15 minutes may be due to increased variability in results with lower input dimensionality, which can compromise model stability. These findings align with existing literature [43], which emphasizes that a larger number of input variable tends to capture relevant patterns more effectively, leading to lower prediction errors. However, reducing the variables can result in the loss of critical information [44].

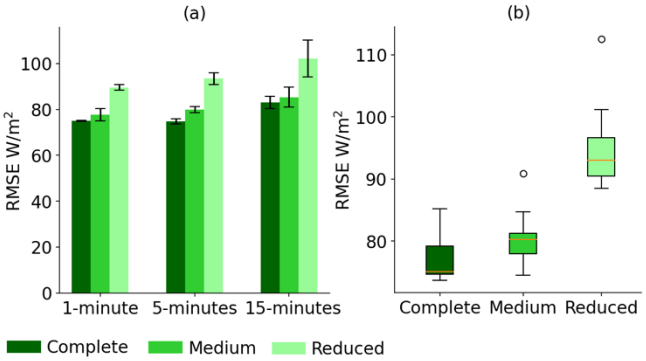


Fig. 6. Overall prediction accuracy per input set.

TABLE V
STATISTICAL COMPARISONS (RMSE)

All the horizons together (overall model's performance)				*t*-statistic	*p*-value	Signifi-cance
	Complete	vs.	Medium	-1.914	0.196	
	Complete	vs.	Reduced	-6.167	0.025	*
	Medium	vs.	Reduced	-6.218	0.025	*
Equal horizons (pairwise comparisons)						
1 min.	Complete	***vs.***	Medium	-1.472	0.279	
	Complete	***vs.***	Reduced	-16.156	0.004	*
	Medium	***vs.***	Reduced	-10.153	0.010	*
5 min.	Complete	vs.	Medium	-10.571	0.009	*
	Complete	vs.	Reduced	-17.830	0.003	*
	Medium	vs.	Reduced	-7.259	0.018	*
15 min.	Complete	vs.	Medium	-5.528	0.031	*
	Complete	vs.	Reduced	-6.937	0.020	*
	Medium	***vs.***	Reduced	-2.475	0.122	

*Denotes the existence of a significant difference (*p-value* < 0.05).

In the specific case of the HPM, depending on the prediction horizon, the smaller number of input features did not influence the prediction accuracy performance. For example, it can be inferred that using six (6) features from the Medium Set produces similar accuracy to using nine (9) features from the Complete Set, depending on the prediction horizon. Fig. 7 shows examples of ANN predictions using the Complete Set, where it can be seen that the models manage to capture the periods of irradiance variability (*Folsom GHI*), but are limited in capturing the variation intensities of solar irradiance. This suggests the need for further studies.

According to the results, this study can contribute to improving and optimizing the processing and execution of the HPM proposed in [21], since the reduction in the number of input variables in the Medium Set (6 features), for 1-minute horizon, allows the prediction accuracy to remain similar to the use of all the features adopted in the Complete Set (9 features). This reduction in the number of input attributes decreases the complexity of the predictive model and the computational pre-processing time of the features by 14.4% (Table I).

It is important to note that the ANN models were not subjected to hyperparameter fine-tunning, as the primary focus was on evaluating the impact of different input feature sets on the prediction accuracy of the HPM model.

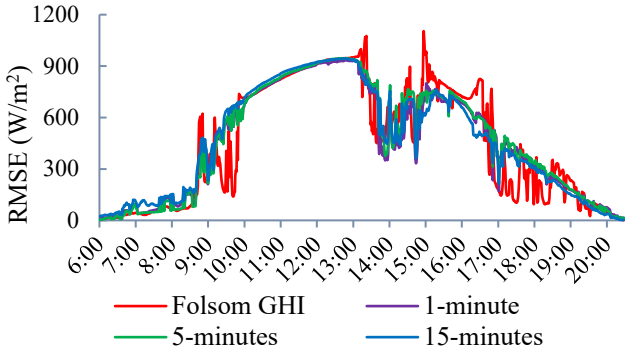


Fig. 7. Prediction examples with the all-input set (28/06/2015).

V. Conclusion

This study conducted an original analysis of the all-sky image processing features set used as inputs in the recent Hybrid Prediction Method (HPM) proposed in [21]. By applying correlation-based attribute selection, we defined and evaluated three sets of input features (Complete, Medium, and Reduced) within the HPM framework using an Artificial Neural Network (ANN) model. The results indicate that the Reduced Set, with only 3 features, significantly underperformed the other input sets.

The ANN models with the Medium set showed no statistically significant difference (p-value < 0.05) in predictive accuracy compared to the Complete Set for the 1-minute horizons. This suggests that a 33.3% reduction in input variables, and 14.4% in computing processing time, can maintain predictive performance for 1-minute horizon. However, the significant difference observed at the 5 and 15-minute horizon suggests that reducing variables may lead to the loss of crucial information [44], increasing prediction error and variability.

These findings are important for the HPM and demonstrate that fewer input variables reduce computational requirements while maintaining similar predictive performance, making the method simpler and interpretable. Further research is required to optimize the HPM, including fine-tuning hyperparameters and refining image processing techniques. In conclusion, this study supports the hypothesis that the Medium set of 6 all-sky image processing features can provide solar irradiance prediction accuracy statistically similar to the Complete set of 9 features, depending on the prediction horizon. To support scientific reproducibility, the supplementary materials are available in: <https://github.com/joylan/id9222ieeela>.

Acknowledgments

The authors would like to thank the Federal University of Latin American Integration (UNILA). O.H.A.J. would like to thank was partially supported by the FACEPE agency (Fundação de Amparo a Pesquisa de Pernambuco) throughout the project with references APQ-0616-9.25/21 and APQ-0642-9.25/22. O.H.A.J. was funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant numbers 407531/2018- 1, 303293/2020-9, 405385/2022-6, 405350/2022-8 and 406662/2022-3, as well as the Program in Energy Systems Engineering (PPGESE) Academic Unit of Cabo de Santo Agostinho (UACSA), Federal Rural University of Pernambuco (UFRPE).

References

[1] ABSOLAR, (2023, Jan. 11). *Energia Solar Fotovoltaica no Brasil - Infográfico*" Associação Brasileira de Energia Solar Fotovoltaica. Accessed on: Jul. 19, 2022. [Online]. Available: https://www.absolar.org.br/mercado/infografico

[2] IEA, "Solar PV," Paris, 2022. Accessed on: Mar. 23, 2023. [Online]. Available: https://www.iea.org/reports/solar-pv

[3] F. Antonanzas-Torres, R. Urraca, J. Polo, O. Perpiñán-Lamigueiro, and R. Escobar, "Clear sky solar irradiance models: A review of seventy models," *Renewable and Sustainable Energy Reviews*, vol. 107, pp. 374–387, Jun. 2019, doi: 10.1016/j.rser.2019.02.032.

[4] C. Ying, W. Wang, J. Yu, Q. Li, D. Yu, and J. Liu, "Deep learning for renewable energy forecasting: A taxonomy, and systematic literature review," *J Clean Prod*, vol. 384, p. 135414, Jan. 2023, doi: 10.1016/j.jclepro.2022.135414.

[5] J. N. Maciel, J. J. G. Ledesma, and O. H. Ando Junior, "Forecasting Solar Power Output Generation: A Systematic Review with the Proknow-C," *IEEE Latin America Transactions*, vol. 19, no. 4, pp. 612–624, Apr. 2021, doi: 10.1109/TLA.2021.9448544.

[6] D. S. Kumar, G. M. Yagli, M. Kashyap, and D. Srinivasan, "Solar irradiance resource and forecasting: a comprehensive review," *IET Renewable Power Generation*, vol. 14, no. 10, pp. 1641–1656, 2020, doi: 10.1049/iet-rpg.2019.1227.

[7] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Convers Manag*, vol. 198, p. 111799, Oct. 2019, doi: 10.1016/j.enconman.2019.111799.

[8] E. D. Obando, S. X. Carvajal, and J. Pineda Agudelo, "Solar Radiation Prediction Using Machine Learning Techniques: A Review," *IEEE Latin America Transactions*, vol. 17, no. 04, pp. 684–697, Apr. 2019, doi: 10.1109/TLA.2019.8891934.

[9] P. Kumari and D. Toshniwal, "Deep learning models for solar irradiance forecasting: A comprehensive review," *J Clean Prod*, vol. 318, no. May, p. 128566, Oct. 2021, doi: 10.1016/j.jclepro.2021.128566.

[10] F. Wang *et al.*, "A minutely solar irradiance forecasting method based on real-time sky image-irradiance mapping model," *Energy Convers Manag*, vol. 220, p. 113075, Sep. 2020, doi: 10.1016/j.enconman.2020.113075.

[11] S. Dev, F. M. Savoy, Y. H. Lee, and S. Winkler, "Estimating Solar Irradiance Using Sky Imagers," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.04981

[12] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, and J. Shields and W. Byron, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," *Solar Energy*, University of California, 9500 Gilman Drive, San Diego, CA, United States: Elsevier, Nov. 2011, pp. 2881–2893. doi: 10.1016/j.solener.2011.08.025.

[13] Z. Zhen *et al.*, "Ultra-short-term irradiance forecasting model based on ground-based cloud image and deep learning algorithm," *IET Renewable Power Generation*, vol. 16, no. 12, pp. 2604–2616, Sep. 2022, doi: 10.1049/rpg2.12280.

[14] H. Yang, L. Wang, C. Huang, and X. Luo, "3d-cnn-based sky image feature extraction for short-term global horizontal irradiance forecasting," *Water (Switzerland)*, vol. 13, no. 13, Jul. 2021, doi: 10.3390/w13131773.

[15] R. A. Rajagukguk, R. Kamil, and H. J. Lee, "A Deep Learning Model to Forecast Solar Irradiance Using a Sky Camera," *Applied Sciences*, vol. 11, no. 11, p. 5049, May 2021, doi: 10.3390/app11115049.

[16] H. Victor *et al.*, "Construção de um Banco de Dados para a Predição do Potencial de Geração de Energia Solar Fotovoltaica," in *IX Encontro Anual de Iniciação Científica - EICTI*, Foz do Iguaçu-PR, 2020, p. 2020.

[17] V. M. Serrano Ardila, J. N. Maciel, J. J. G. Ledesma, and O. H. Ando Junior, "Fuzzy Time Series Methods Applied to (In)Direct Short-Term Photovoltaic Power Forecasting," *Energies (Basel)*, vol. 15, no. 3, p. 845, Jan. 2022, doi: 10.3390/en15030845.

[18] V. H. Wentz, J. N. Maciel, J. J. Gimenez Ledesma, and O. H. Ando Junior, "Solar Irradiance Forecasting to Short-Term PV Power: Accuracy Comparison of ANN and LSTM Models," *Energies (Basel)*, vol. 15, no. 7, p. 2457, Mar. 2022, doi: 10.3390/en15072457.

[19] J. N. Maciel, "Método Híbrido de Predição da Irradiância Solar com Processamento de Imagens e Inteligência Artificial Aplicável a Geração de Energia Solar Fotovoltaica," Ph.D thesis, ILATIT, Universidade Federal da Integração Latino-Americana (UNILA), Foz do Iguaçu, PR, 2022.

[20] J. N. Maciel, V. H. Wentz, J. J. G. Ledesma, and O. H. Ando Junior, "Analysis of Artificial Neural Networks for Forecasting Photovoltaic Energy Generation with Solar Irradiance," *Brazilian Archives of Biology and Technology*, vol. 64, no. spe, 2021, doi: 10.1590/1678-4324-75years-2021210131.

[21] J. N. Maciel, J. J. G. Ledesma, and O. H. Ando Junior, "Hybrid prediction method of solar irradiance applied to short-term photovoltaic energy generation," *Renewable and Sustainable Energy Reviews*, vol. 192, p. 114185, Mar. 2024, doi: 10.1016/j.rser.2023.114185.

[22] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 2nd ed., Upper Saddle River, NJ, US: Prentice-Hall, 2002.

[23] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd edition. United Kingdom: Pearson Education, 2016.

[24] H. T. C. Pedro, D. P. Larson, and C. F. M. Coimbra, "A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 3, p. 036102, May 2019, doi: 10.1063/1.5094494.

[25] K. Lappalainen and S. Valkealahti, "Photovoltaic mismatch losses caused by moving clouds," *Solar Energy*, vol. 158, no. October, pp. 455–461, 2017, doi: 10.1016/j.solener.2017.10.001.

[26] R. Blaga, A. Sabadus, N. Stefu, C. Dughir, M. Paulescu, and V. Badescu, "A current perspective on the accuracy of incoming solar energy forecasting," *Prog Energy Combust Sci*, vol. 70, pp. 119–144, 2019, doi: 10.1016/j.pecs.2018.10.003.

[27] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016, doi: 10.1016/j.solener.2016.06.069.

[28] R. Asghar, F. R. Fulginei, M. Quercio, and A. Mahrouch, "Artificial Neural Networks for Photovoltaic Power Forecasting: A Review of Five Promising Models," *IEEE Access*, vol. 12, pp. 90461–90485, 2024, doi: 10.1109/ACCESS.2024.3420693.

[29] F. Lin, Y. Zhang, and J. Wang, "Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods," *Int J Forecast*, vol. 39, no. 1, pp. 244–265, Jan. 2023, doi: 10.1016/j.ijforecast.2021.11.002.

[30] G. M. Tina, C. Ventura, S. Ferlito, and S. De Vito, "A State-of-Art-Review on Machine-Learning Based Methods for PV," *Applied Sciences*, vol. 11, no. 16, p. 7550, Aug. 2021, doi: 10.3390/app11167550.

[31] R. A. Rajagukguk, R. A. A. Ramadhan, and H.-J. Lee, "A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power," *Energies (Basel)*, vol. 13, no. 24, p. 6623, Dec. 2020, doi: 10.3390/en13246623.

[32] A. L. Martinez and M. C. R. Dumer, "Adoption of IFRS and the Properties of Analysts' Forecasts: The Brazilian Case," *SSRN Electronic Journal*, Oct. 2012, doi: 10.2139/ssrn.2153173.

[33] C. A. Gueymard, "A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects," *Renewable and Sustainable Energy Reviews*, vol. 39, pp. 1024–1034, 2014, doi: 10.1016/j.rser.2014.07.117.

[34] B. Juncklaus Martins *et al.*, "Systematic review of nowcasting approaches for solar energy production based upon ground-based cloud imaging," *Solar Energy Advances*, vol. 2, p. 100019, 2022, doi: 10.1016/j.seja.2022.100019.

[35] U. K. Das *et al.*, "Forecasting of photovoltaic power generation and model optimization: A review," *Renewable and Sustainable Energy Reviews*, vol. 81, no. August 2017, pp. 912–928, 2018, doi: 10.1016/j.rser.2017.08.017.

[36] A. C. Gil, *Como elaborar projetos de pesquisa*, 6ª ed. São Paulo, Brazil: Atlas, 2017.

[37] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*, IEEE, Dec. 2014, pp. 1–6. doi: 10.1109/ICCIC.2014.7238499.

[38] P. A. Morettin and W. de O. Bussab, *Estatística Básica*, 9 ed. São Paulo, Brazil: Saraivauni, 2017.

[39] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[40] E. Bisong, "Google Colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. 59–64. doi: 10.1007/978-1-4842-4470-8_7.

[41] A. C. Müller, M. A. C, and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016.

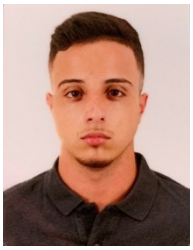
[42] E. Mining, *Python Machine Learning: Understand Python Libraries (Keras, NumPy, Scikit-Lear, TensorFlow) for Implementing Machine Learning Models in Order to Build Intelligent Systems*. Amazon Digital Services LLC - KDP Print US, 2019. [Online]. Available: https://books.google.com.br/books?id=qqQdzAEACAAJ

[43] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," in *Machine Learning*, Lisa Hellerstein, Ed., Netherlands: The MIT Press, 2003, pp. 239–281. doi: https://doi.org/10.1023/A:1024068626366.

[44] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," in *Journal of Machine Learning Research*, *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003 [Online]. Available: http://dblp.uni-trier.de/db/journals/jmlr/jmlr3.html#GuyonE03

Joylan Nunes Maciel received a Bachelor's degree in Computer Science from Western Paraná State University (2005), and a Ph.D. from the Federal University of Latin American Integration (UNILA) in 2022, where he currently works as Associate Professor. Researcher at the Laboratory of Applied Computing

(LACA). His main research interests include computer science, mobile computer, software development, and Artificial Intelligence.



Gustavo Campoi de Souza is a materials engineering student at the UNILA and Laboratory of Applied Computing (LACA). His main research areas are software programming and Artificial Intelligence.

Willian Zalewski received the title bachelor (2006) of Computer Science from the Western Paraná State University, and Ph.D at Federal University of Paraná (2008). Researcher at the Laboratory of Applied Computing (LACA). His main research is in the areas of computer science, Artificial Intelligence, Data Mining, Machine Learning, and Deep Learning.

Jorge Javier Giménez Ledesma has a degree in Electromechanical Engineering from the University Nuestra Señora de la Asunción (2009), a master's degree (2012) and Ph.D (2017) in Electrical Engineering from the Federal University of Juiz de Fora. Adjunct Professor at the UNILA. He works on the following topics: Analysis of the protection system in computer distribution and programming systems.

Oswaldo Hideo Ando Junior, Graduated in Electrical Engineering (2006) with a MBA (2007) from the ULBRA with Master's Degree in Electrical Engineering (2009) and Ph.D. in Engineering (2014) from UFRGS. His research interests included Power Quality, Energy Harvesting, Energy Informatics, Renewable Energy, Smart Grid, Distribution Energy Resource and Storage Energy System. Leader of the Research Group on Energy and Sustainability (GPENSE) and researcher at the Laboratory of Intelligent Electric Grids -LREI.