2022 IEEE Kansas Power and Energy Conference (KPEC) | 978-1-6654-6591-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/KPEC54747.2022.9814765

Solar Radiation Forecasting Using Artificial Neural Networks Considering Feature Selection

Reza Nematirad
Department of Electrical and
Ccomputer Engineering
Kansas State University
Manhattan, Ks, USA
Nematirad@ksu.edu

Anil Pahwa
Department of Electrical and
Computer Engineering
Kansas State University
Manhattan, KS, USA
Pahwa@ksu.edu

***Abstract*— Due to various factors, including worries about greenhouse gas emissions, supporting government policies, and decreased equipment costs, the expansion of solar-based energy generation, notably in the form of photovoltaics, has accelerated significantly in recent years. Solar panels continue to face several challenges regarding their practical integration and reliability. These concerns originate from the variable nature of the solar resource. Solar generation has inherent variability, which poses problems associated with the costs of supplemental generation and grid reliability. Therefore, high accuracy solar forecasting is required. Several machine learning strategies are broadly employed for solar power forecasting. However, analyzing solar radiation characteristics in order to select the features that have a meaningful correlation between inputs and outputs of machine learning algorithms has received less attention. This study uses a multilayer perceptron (MLP) artificial neural network (ANN) with Bayesian optimization to forecasting solar radiation. The Pearson Correlation Coefficients (PCCs) are used to select effective features. The simulation findings reveal that the accuracy assessment metrics are higher when employing feature selection for prediction.**

Keywords—solar forecasting, neural networks, feature selection, Pearson correlation coefficient, multilayer perceptron.

I. Introduction

Growing demand for renewable energy is increasing the economic and technical complications of photovoltaic systems and their integration into the power grid [1]. There are many factors that contribute to these problems [2]-[3], including the fluctuating nature of the solar resources, seasonal differences in production and demand, and the high cost of energy storage. As a result, ancillary generators typically support solar plants during periods of high variability, increasing the investment and operation costs of these plants [4]. Network operators or utilities can successfully schedule short or long-term planning to analyze the effects of ancillary generators on power systems and maintain grid reliability [5]. Furthermore, independent system operators or utilities must be able to precisely forecast solar generation or solar radiation over various time horizons so that high levels of photovoltaic systems can be integrated while maintaining reliability [6]. Therefore, solar forecasting is crucial for increasing solar penetration into the grid. Because the generation of solar panels is directly related to solar irradiance, radiation prediction results can also be used for planning.

Applications of machine learning techniques have grown significantly in all aspects of power systems. The author in [7] used long short-term memory, random forest, and support vector machine to forecast the error between the real-time and day-ahead price in the power market. It is also shown in [8] that reinforcement learning can be successfully implemented for the coordinated management of electric vehicles. In [9], machine learning is used by which agents can learn to forecast electricity prices and strategically bid to maximize their benefits.

Solar radiation forecasting is one of the most important artificial intelligence applications in power systems. Based on the application, different methods for predicting solar radiance and generation are used, such as numerical weather prediction (NWP), statistical machine learning algorithms, and climatology [10]. Many studies show that using machine learning algorithms on weather inputs is an effective and accurate technique [11]. In [12], three machine learning algorithms, including multilayer perceptron (MLP), support vector machine (SVM), and multivariate regression, are used to forecast solar radiation, and MLP outperforms SVM and multivariate regression. Some authors have used auto-regressive moving average (ARMA) to predict solar generation using solar radiation data from Solar Anywhere [13].

The feature selection in input data is significant to forecasting problems. There are two major reasons for feature selection: 1) Feature selection enables machine learning to train quicker, and 2) It improves the model's accuracy by preventing overfitting. [14]. There are several ways to categorize feature selection methods. It can be divided into filters, wrappers, embedding, and hybrid approaches [15].

Authors of [16] applied feature importance and principle component analysis (PCA) to select the most influential features and determined that cloud cover, ultraviolet index, ambient temperature, relative humidity, dew point, wind bearing, sunshine duration, and length of days significantly affect global horizontal irradiance forecasting. While their results look promising, new approaches can be investigated to further refine the feature selection process. Also, the authors of [16] did not provide any comparisons with and without feature selection to evaluate feature selection's effects on neural network accuracy.

This paper uses a multilayer perceptron artificial neural network to develop short-term solar radiance forecasting. The Pearson Correlation Coefficients (PCCs) are applied to score the correlation between inputs and outputs. In addition, Bayesian optimization is used to enhance neural networks' performance

978-1-6654-6591-5/22/$31.00 ©2022 IEEE

and accuracy. This work compares the effect of using PCC versus not using PCC on neural network's performance.

II. DATA COLLECTION AND METHODOLOGY

A. Data Collection

The information used to create the models in this study came from the Kaggle database for Hawaii [17] and [18]. This meteorological data is made up of eight elements: radiation, temperature, pressure, humidity, wind speed, wind direction, sunrise time, and sunset time. The data was collected at a five-minute interval from September to December 2017. Furthermore, the length of a day is calculated by subtracting sunset from sunrise time and then converting it to decimal metrics. It should be noted that the day length data is provided daily, while the rest of the information is provided every 5 minutes. For matching all data dimensions, the length of the day is also listed every five minutes.

B. Feature Selection

The feature selection technique is used in developing a learning model to reduce the number of input variables. Reduced input variables help lessen the computation volume and increase the model's accuracy as well. The goal of feature selection is to release redundant features before feeding them into the machine learning algorithm. A supervised and unsupervised approach could be considered when it comes to feature selection methods. The wrapper, filter, and intrinsic processes are used to select features. Many models are developed with various features, and the best model is selected based on a performance metric among all the models created by the wrapper feature selection. Filter feature selection processes employ statistical techniques to correlate features and the outputs. Intrinsic feature selection methods are embedded within some machine learning algorithms, such as random forest and light gradient boosting algorithms [19].

PCC is often used for feature selection in artificial intelligence. According to PCC, a correlation between two variables is numerically measured as a scaler between -1 and +1. The closer the PCC is to the ±1, the stronger the linear correlation between the two variables. The PCC of two completely uncorrelated variables is 0 [20]. The PCC is defined as [21]:

$$R_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Where cov(X,Y) represents the covariance between variables X and Y, and σ_X, σ_Y indicates standard deviations of variables X, Y respectively. Because the outputs of (1) can have both positive and negative values, squaring ensures that all values are between 0 and 1, which is defined as the coefficient determination [22].

C. Multilayer Perceptron

Neural networks are computing systems that try to simulate and model biological neural networks in the human brain. In a human brain, numerous neurons are connected in a complicated fashion to each other. Neural networks can be mathematically constructed in which input data is fed into the neurons. All neurons' inputs are multiplied with their weight and then summed by a bias value. Afterward, these results are entered into activation functions. This study utilizes log-sigmoid, hyperbolic tangent, and Purelin activation functions. Since weight values are random, minimizing the mismatch between activation function outputs and actual values is required. For this purpose, MLP is used. The output of each neuron is as follow [23]:

$$u_k = \sum_{j=1}^{m} X_j . w_j + b_k \tag{2}$$

Where w_j is the weight between neuron k and jth input, X_j indicates the input vector, k is the number of neurons, m is the number of input toward neuron k, and b_k is the bias parameter for kth neuron. For the artificial neural network model to be more accurate and capable of performing nonlinear regressions, vector u_k is passed through a nonlinear activation function as follows:

$$S_k = \varphi(u_k) \tag{3}$$

Where φ is an activation function such as ReLU, Sigmoid, and Tanh. All neurons are linked, and the ANN is formed.

D. Evaluation Metrics

To measure the ANN model's accuracy, we have used Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) when evaluating forecasted results. More details about these metrics are discussed in [24].

III. RESULTS AND DISCUSSINO

In this study, solar irradiance is forecasted by ANN in two different states. First, the PCC correlated the input data, and the data with the more robust correlation with radiation are identified for use in the ANN model. Second, ANN forecasts solar irradiance without feature selection. Also, the data are split into three sets: training, cross-validation, and testing datasets. More specifically, we randomly selected 80% of the data as the training set. The remaining 20% of data are randomly divided into the cross-validation and testing datasets with 10% of data for each set. Table I shows the values of the ANN parameters that are used.

TABLE I
Characterization of ANN parameters

Parameter	Value
Solver	Adam
Alpha	0.0001
Batch_size	Auto
Learning_rate	0.001
Power_t	0.5
Max_iter	4000
Shuffle	True
Random state	42
Hyperparameter optimizer	Bayesian

A. *Feature Selection Results*

Fig. 1 depicts all features and output in order to provide a better understanding of the relationship between inputs and target. Fig. 1 shows visually that the ambient temperature and air pressure are highly correlated with solar radiation. On the other hand, solar radiation appears to have only a very slight correlation with time and day length. To analyze the relationship between the features and outputs mathematically, the results of PCC are exhibited in Table II and Fig. 2. As shown in Fig. 2, ambient temperature, wind direction, humidity, and air pressure strongly correlate with solar radiance, but day length, time, and wind speed do not. These results verify the visual observations from Fig.1. The PCC between solar radiation and ambient temperature is approximately 0.73, which is the highest among all parameters. Thus, ambient temperature should be considered as a powerfully relevant parameter in modeling. The PCC for air humidity, wind direction, and air pressure are 0.23, 0.23, and 0.12, respectively. As a result, these parameters are among the marginally relevant features but not redundant. Finally, the PCC between wind speed, day length, time, and radiance is negligible. Therefore, these data are irrelevant and unnecessary to include. However, in [16], length of day is considered as a selected feature and used for solar forecasting. Based on these results, ambient temperature, air pressure, humidity, and wind direction are considered features for the ANN model.

Table II
PCC BETWEEN SOLAR RADIATION AND METEOROLOGICAL PARAMETERs

Inputs	Pearson correlation coefficient
Time	0.004050
Temperature (^{o}C)	0.734903
Air Pressure(bar)	0.119016
Humidity (%)	0.226026
Wind Direction (Degrees)	0.230354
Wind speed(m/s)	0.073639
Day length(h)	0.002000
Radiance (w/m^2)	1

TABLE III
MAE AND RMSE RESULT OF ANN IN TWO FORMS

Form	Evaluation Metric	value
With PCC	MAE	109.45
	RMSE	177.57
Without PCC	MAE	111.24
	RMSE	182.25

B. *ANN Results*

Fig. 3 depicts solar radiation forecasted with and without PCC, as well as actual radiation. Instead of showing all the points in this figure, a few points are selected to illustrate the results. As shown in this figure, feature selection provides predicted solar radiation values closer to the actual solar radiation value in most cases. However, Sample 2 appears to be an outlier. Table III lists the MAE and RMSE evaluation metrics. It is clear that utilizing feature selection increases forecasting accuracy. RMSE and MAE metrics improve by 2.56 and 1.6 percent by applying feature selection. These results are superior to those reported in [16] for forecasting using the artificial neural network (ANN). The RMSE in the best case of ANN with feature selection using PCA, is 18,633.53 w/m^2, but in this study, it reached 177.57 w/m^2. This implies that PCC did better than PCA for feature selection. For example, PCC detected day length as a redundant feature while PCA used it in the learning machine as an important feature. Further, enhancing neural network accuracy with Bayesian optimization identified the hyperparameters that improved the network accuracy.

IV. CONCLUSION AND RECOMMENRATIONS

In this study, an application of machine learning is demonstrated for planning and management of power systems. An artificial neural network is developed for solar irradiation forecasting. Pearson Correlation Coefficients for feature selection and Bayesian optimization are used to improve model accuracy. The simulation results indicate that the ANN is more accurate when feature selection is applied. Also, the combination of PCC and neural networks by Bayesian optimization significantly outperforms the combination of PCA

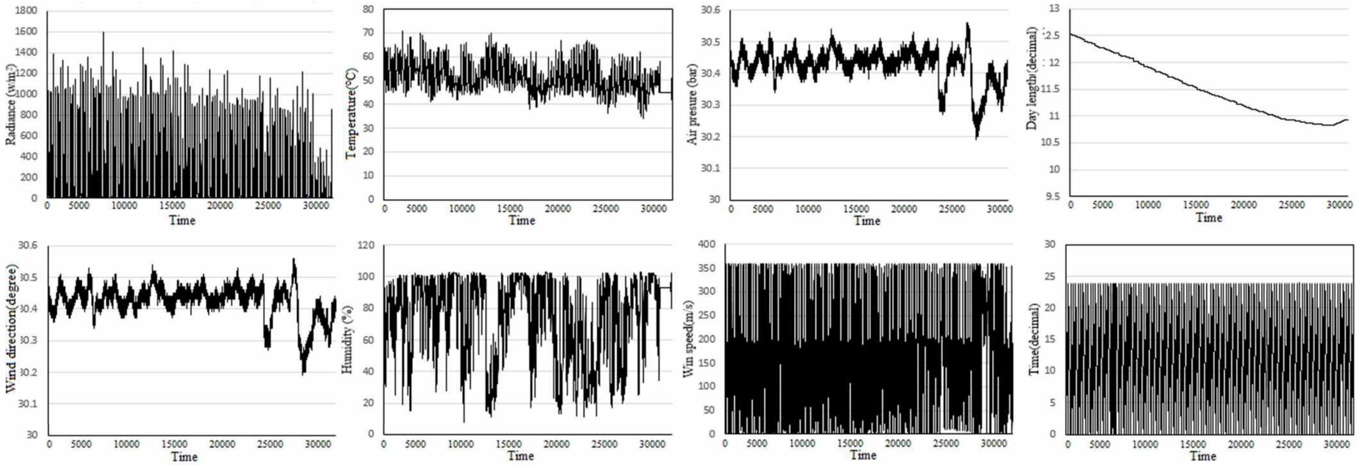

Fig. 1. Radiation, temperature, air pressure, day length, wind direction, humidity, wind speed, and time.

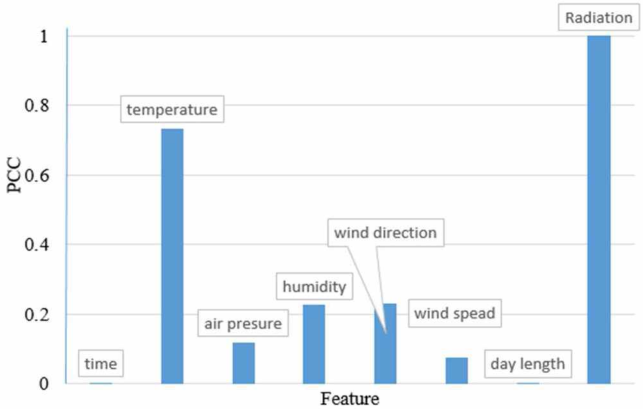

Fig. 2. PCC between solar radiance and Meteorological parameters. *For example, R-time means PCC between solar radiance and time.

feature selection and neural networks. The results, however, are inferior to the best result reported in [16] with PCA and XGBoost (an algorithm based on sequential ensemble of decision trees). In the future research, different feature selections methods such as wrapper and filter methods can be combined with machine learning algorithms to improve solar irradiation forecasting. Ensembles of machine learning algorithms such as tree decisions can be used to forecast solar radiation. Furthermore, other optimization techniques to tune and improve neural network coefficients and weights can be implemented.

REFERENCES

[1] Dindar, Amin, Shiva Ourang, and Erfan Gholami Ghadikola. "Development of a Communication-Assisted Adaptive Overcurrent Protection Scheme in Smart Distribution Networks in Presence of Wind and Solar Generation."

[2] A. F. Soofi, S. D. Manshadi, G. Liu and R. Dai, "A SOCP Relaxation for Cycle Constraints in the Optimal Power Flow Problem," in *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1663-1673, March 2021, doi: 10.1109/TSG.2020.3023890.

[3] Sirat, Ali Parsa. "Loss minimization through the allocation of dgs considering the stochastic nature of units." *arXiv preprint arXiv:1911.06748* (2019).

[4] Lew D, Piwko R. Western wind and solar integration study. Tech. rep.. National Renewable Energy Laboratories; 2010. Technical Report No. NREL/SR- 50e47781

[5] Bayani, Reza, Mohammed Bushlaibi, and Saeed D. Manshadi. "Short-term Operational Planning Problem of the Multiple-Energy Carrier Hybrid AC/DC Microgrids." arXiv preprint arXiv:2012.12788 (2020).

[6] Venkataraman S, Jordan G, Piwko R, Freeman L, Helman U, Loutan C, et al. Integration of renewable resources: operational requirements and generation fleet capability at 20% rps. Tech. rep.. California: ISO; 2010.

[7] Nizharadze, Nika, Arash Farokhi Soofi, and Saeed D. Manshadi. "Learning the Gap in the Day-Ahead and Real-Time Locational Marginal Prices in the Electricity Market." *arXiv preprint arXiv:2012.12792* (2020).

[8] R. Bayani, S. D. Manshadi, G. Liu, Y. Wang and R. Dai, "Autonomous charging of electric vehicle fleets to enhance renewable generation dispatchability," in CSEE Journal of Power and Energy Systems, doi: 10.17775/CSEEJPES.2020.04000.

[9] Mohtavipour, Seyed Saeid, and Mehdi Jabbari Zideh. "An iterative method for detection of the collusive strategy in prisoner's dilemma game of electricity market." IEEJ Transactions on Electrical and Electronic Engineering 14.2 (2019): 252-260.

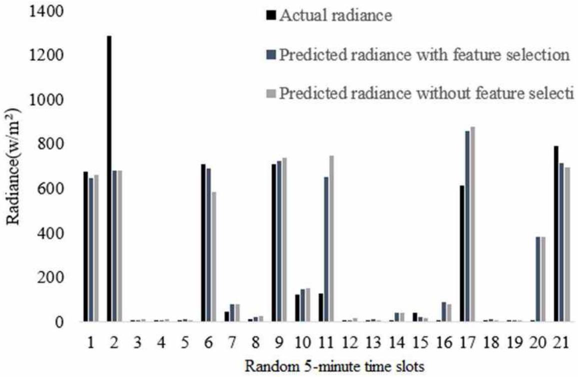

Fig. 3. The actual and predicted radiation using the ANN without and with feature selection.

[10] Tuohy, Aidan, et al. "Solar forecasting: methods, challenges, and performance." IEEE Power and Energy Magazine 13.6 (2015): 50-59.

[11] Saad, Basma, et al. "Assessing the Impact of Weather Forecast Models Combination on the AMS Solar Energy Prediction." 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS). IEEE, 2020.

[12] Sabzehgar, Reza, Diba Zia Amirhosseini, and Mohammad Rasouli. "Solar power forecast for a residential smart microgrid based on numerical weather predictions using artificial intelligence methods." *Journal of Building Engineering* 32 (2020): 101629.

[13] Huang, Rui, et al. "Solar generation prediction using the ARMA model in a laboratory-level micro-grid." 2012 IEEE third international conference on smart grid communications (SmartGridComm). IEEE, 2012.

[14] Jović, Alan, Karla Brkić, and Nikola Bogunović. "A review of feature selection methods with applications." 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO). Ieee, 2015.

[15] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method", Expert Systems with Applications, vol. 41, issue 14, pp. 6371–6385, 2014.

[16] Munawar, Usman, and Zhanle Wang. "A framework of using machine learning approaches for short-term solar power forecasting." Journal of Electrical Engineering & Technology 15.2 (2020): 561-569.

[17] Hackathon N (2017) Solar radiation prediction.

[18] https://www.timeanddate.com/sun/usa/honolulu

[19] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[20] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," Procedia Computer Science, vol. 23, pp. 5–14, 2013.

[21] Blessie, E. Chandra, and E. Karthikeyan. "Sigmis: A feature selection algorithm using correlation based method." *Journal of Algorithms & Computational Technology* 6.3 (2012): 385-394.

[22] P. Socha, V. Miškovský, H. Kubátová and M. Novotný, "Optimization of Pearson correlation coefficient calculation for DPA and comparison of different approaches," 2017 IEEE 20th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2017, pp. 184-189, doi: 10.1109/DDECS.2017.7934563.

[23] Castangia, Marco, et al. "A compound of feature selection techniques to improve solar radiation forecasting." *Expert Systems with Applications* 178 (2021): 114979.

[24] Brownlee, Jason. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery, 2018.