Machine Learning Based Prediction of Solar Power Generation by Incorporating Denoising Methods

Jyothi Godlaveti [1],Padma Lalitha.M[2],Harshitha Chejerla[3], Chaitanya Palem[4], Chandra Sekhar Vanta[5] & Manvitha Mahalakshmi.N[6]

[1,3 ,4 &5] UG Scholars, Department of EEE, Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh, India
[6]UG Scholar,Department of EEE, Annamacharya Univesity, Rajampet , Andhra Pradesh , India
[2]Professor, Department of EEE, Annamacharya university, Rajampet, Andhra Pradesh , India

Corresponding Author : mpl@aitsrajampet.ac.in

Abstract. To achieve optimal planning of the energy system and stability of this system, it is essential to have precise forecasting of the solar power. Based on long-term solar production and meteorological records obtained at the Annamacharya University in Rajampet, this paper proposes a machine learning-based architecture to photovoltaic power prediction. In the bid to improve the quality of data, Empirical Mode Decomposition (EMD) and Wavelet Transform (WT) are applied to remove noise and extract the meaningful signal components. Various machine learning models are trained on the denoised data including regression, ensemble, and deep learning models. Considering the model performance, R2 MAE, MedAE and RMSE are utilized. The analysis of the results demonstrates that denoising-based models, in particular, ensemble and deep learning approaches, are better than traditional methods in terms of forecasting the accuracy. The proposed framework will promote reliable integration of photovoltaic systems in smart grids and management of solar energy.

Keywords :WT, EMD, solar power forecast, deep learning, machine learning, renewable energy, and smart grid.

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_52

1 Introduction

The rapid development of the renewable energy technologies has transformed solar electricity to form a substantial part of the modern power systems. Solar power prediction is critical to grid operation, planning and cost management with increased utilization of solar energy. Nevertheless, solar power output is influenced by weather conditions such as temperature, humidity, movement of clouds, and the amount of sun energy all of which are changing in a complex manner. This minimizes use of conventional forecasting methods.

The recent research indicated that machine learning and deep learning models have a great effect in enhancing the accuracy of solar power forecasting in combination with the use of preprocessing methods like the wavelet transform and empirical mode decomposition methods. Such techniques assist in deleting noise in weather information and enhancing the predictability. Recent studies have proved that ensemble learning models which include gradient boosting method and random forest give a higher forecasting accuracy than conventional statistical models. Thus, combining machine learning models with signal denoising methods will help to improve the performance of solar power prediction and optimize renewable energy.

2 Literature Survey

Antonanzas et al. [1] established that machine learning and advanced forecasting are more efficient than the standard statistical techniques due to their capability to diabolical nonlinear interactions between the solar power output and environmental variables. Hong et al. [2] proposed the significance of proper energy forecasting and stressed that data-driven methods are more effective when they are trained on dataset gathered in different weather conditions. A thorough review of methods of solar forecasting conducted by Inman et al. [3] showed that machine learning methods have been especially efficient when applied during the dynamic conditions of the atmosphere.

The authors of the article Voyant et al. [4] examined the issue of multi-horizon solar radiation forecasting and demonstrated that time-series models with the use of meteorological inputs can enhance the quality of forecasting. Mohandes et al. [5] established that renewable energy parameters like the speed of wind can be forecasted using neural network methods which implies the possibility of the neural networks in predicting renewable energy sources. Pedro and Coimbra [6] concluded that the short-term weather changes carry a wide-ranging impact on the performance of the time-series forecasting models.

Voyant et al. [7] also demonstrated that the optimized artificial neural network models can be used to improve the accuracy of solar radiation prediction significantly. According to Reikard [8], traditional statistical time-series forecasting techniques are susceptible to atmospheric variability, and therefore they cannot be useful in highly variable weather conditions. Randimbivololona et al. [9] suggested hybrid forecasting methods which are the combination of ARMA with the use of neural networks in order to enhance the accuracy of solar radiation prediction.

Yang et al. [10] provided an overall overview of the solar irradiance and photovoltaic power prediction methods, and proposed that the combination of preprocessing, feature selection, and created using the ensemble method can enhance the reliability of forecasting. Lalitha et al. [11] presented the use of machine learning algorithms in predicting solar power using real data of an Indian solar installation. The empirical mode decomposition was proposed by Huang et al. [12] as a method of studying nonlinear and non-stationary time series, and the

most popular wavelet-based signal processing methods are proposed by Mallat [13] and applied in the renewable energy forecasting field to reduce noise and preprocess the signal.

3 System Architecture

The proposed solar power prediction system will be a system with a module of data collecting, preprocessing, and forecasting. Solar power or the power produced by the solar plant is influenced by meteological factors such as temperature, humidity and sun radiance and it is these factors that are included in the prediction models at the 500 kW installed solar power plant at the Annamacharya University.The solar power is supplied by eight inverters, four inverters of 100 kW capacity and four inverters of 25 kW capacity. The data of generation of one 100 kW inverter was monitored during a period of five years in this work through the monitoring platform [11]. The data includes 93,756 samples that were taken at 15 minutes. The records have the values of DATE_TIME, AC_POWER, and DC_POWER. The model was also trained and tested on the same period of weather data.

A.DataCollection:

Install the Annamacharya University, 500 kW solar power plant, 8 inverters of 100 kW and 4 inverters of 25 kW. In this research, the monitoring platform was utilized to record the data on the generation of one 100 kW inverter within five years [11]. The samples in the dataset amounted to 93,756, which were carried out after every 15 minutes. Every record contains the values DATE_TIME, AC power and DC power. To train and evaluate the model, the weather data in the same period of time was collected as well.

Fig. 1 shows the total generation data utilized for solar power forecasting.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93756 entries, 0 to 93755
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   DATE_TIME   93756 non-null  object
 1   AC_POWER    93756 non-null  float64
 2   DC_POWER    93756 non-null  float64
dtypes: float64(2), object(1)
memory usage: 2.1+ MB
```

	DATE_TIME	AC_POWER	DC_POWER
0	01-01-2020 06:15	0.04165	0.041650
1	01-01-2020 06:30	0.35867	0.364767
2	01-01-2020 06:45	1.77553	1.825467
3	01-01-2020 07:00	3.40627	3.506753
4	01-01-2020 07:15	6.78082	6.969747

Fig. 1: Information about Generation data

Weather Data:

The meteorological data needed to determine the location of the institute (latitude: 14.2252, longitude: 79.1395) were obtained in the NASA POWER web site [12] since the weather measurements are not recorded there. The weather data that contains DATE_TIME (which is simultaneously referred to as T) and temperature, humidity and sun irradiation properties contains 43,848 recordings collected in five years. To align the generation dataset with the dataset of the forecasts, and achieve higher forecasting precision, the data was initially sampled at 1-hour time intervals and later resampled at 15-minute time intervals when linear

interpolation was used. Fig. 2 gives an overview of the weather data used in the proposed solar power prediction model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43848 entries, 0 to 43847
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   DATE_TIME     43848 non-null  object
 1   TEMPERATURE   43848 non-null  float64
 2   HUMIDITY      43848 non-null  float64
 3   IRRADIATION   43848 non-null  float64
dtypes: float64(3), object(1)
memory usage: 1.3+ MB
```

	DATE_TIME	TEMPERATURE	HUMIDITY	IRRADIATION
0	01-01-2020 00:00	20.31	15.19	0.0
1	01-01-2020 01:00	20.25	15.16	0.0
2	01-01-2020 02:00	20.15	15.12	0.0
3	01-01-2020 03:00	20.11	15.09	0.0
4	01-01-2020 04:00	20.12	15.08	0.0

Fig. 2: Information about Weather data

B. Details Pre-processing:

Data are collected and then cleaned and set to be analyzed. The data on weather and solar generation is aggregated at the preparation stage to form a single dataset with the help of a common time so that further processing and model training could be performed.

Fig. 3 indicates the integrated dataset which incorporates solar generation and meteorological data. This is because the data set has two target variables (AC power and DC power) and their correlation is near unity; therefore, considering this, one of the target variables is eliminated to make the prediction model simpler.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93756 entries, 0 to 93755
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   DATE_TIME    93756 non-null  object
 1   AC_POWER     93756 non-null  float64
 2   DC_POWER     93756 non-null  float64
 3   TEMPERATURE  93756 non-null  float64
 4   HUMIDITY     93756 non-null  float64
 5   IRRADIATION  93756 non-null  float64
dtypes: float64(5), object(1)
memory usage: 4.3+ MB
```

	DATE_TIME	AC_POWER	DC_POWER	TEMPERATURE	HUMIDITY	IRRADIATION
0	01-01-2020 06:15	0.04165	0.041650	21.4225	15.7650	65.9425
1	01-01-2020 06:30	0.35867	0.364767	21.8850	16.0000	100.9850
2	01-01-2020 06:45	1.77553	1.825467	22.3475	16.2350	136.0275
3	01-01-2020 07:00	3.40627	3.506753	22.8100	16.4700	171.0700
4	01-01-2020 07:15	6.78082	6.969747	23.2275	16.5825	211.0950

Fig. 3: Information about merged dataset

```
Correlation between AC_POWER and DC_POWER = 1.00
<Axes: xlabel='AC_POWER', ylabel='DC_POWER'>
```

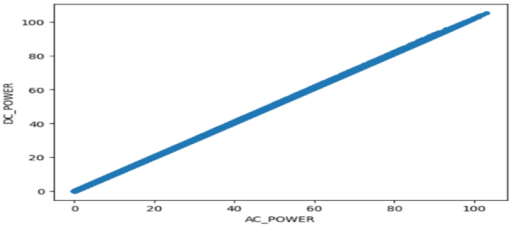

Fig. 4: Linear Relationship of AC and DC power.

The relationship between DC and AC power has been observed in Fig. 4 in which the correlation is nearly one indicating that the two are directly proportional. To simplify the model, DC power is removed in the dataset. To improve the quality of data and ensure reliability in the analysis, the rows with zero values are also eliminated.

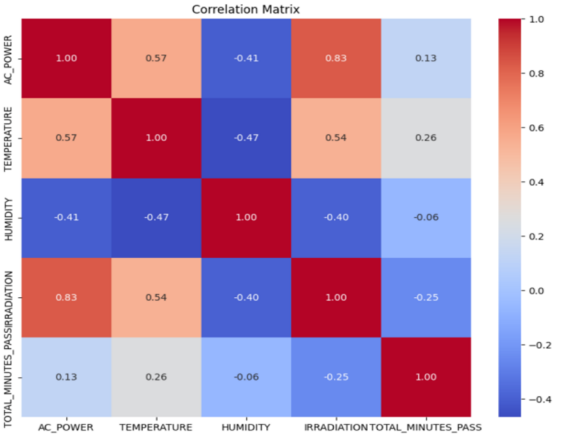

Fig5: correlation matrix

The correlation chart in Fig.:5 shows the effect of temperature, humidity and the sun radiation on the AC power. Box-plot analysis is employed to identify and remove the outliers in the combined data to enhance greater accuracy of the model and reduce biased predictions.

C. Feature Engineering:

Feature engineering is performed on the cleaned dataset in order to identify meaningful input features. Pattern analysis, production of the TOTAL_MINUTE_PASS feature based on date and time, partitioning of input features and target variables, partitioning the data into training and testing data sets, exploration of the associations among the variables and feature scaling are all part of this process.

Timestamp Feature: Totally, Minutes Passed: This is a time-based feature that is developed since machine learning models cannot utilize raw timestamps directly. This facilitates easier understanding of the model on the variation of solar power with time.

Split entails splitting the data into two groups (Training and Testing):The weather and solar power data are combined and divided into two variables y, which is AC power, and input features (X), which comprise temperature, humidity, and irradiance. To train the model and test it, the data is then split into training and testing sets.

The model is trained on the training set (80 per cent) of the dataset and the testing set (20 per cent) is utilized to evaluate the performance of the model on data that has never been encountered.

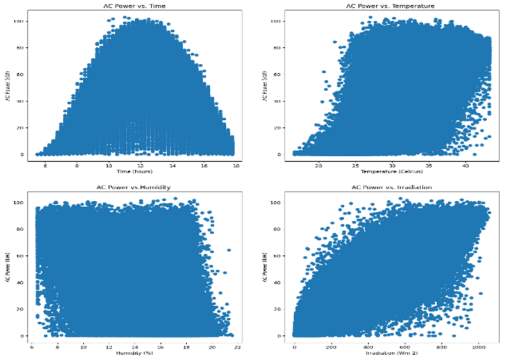

Fig. 6: AC power against input variables.

The correlation between the AC power and the selected input variables are represented by scatter plots in Fig. 6.

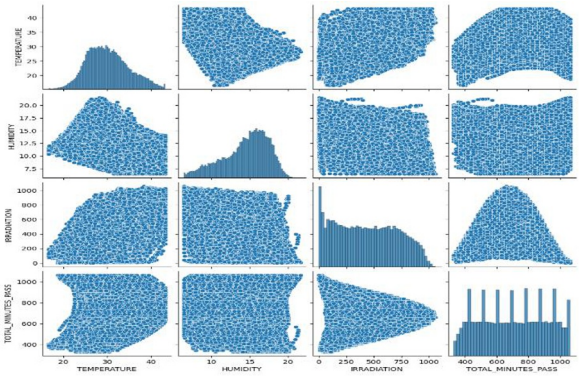

Fig 7:Graphs for all input variables

Fig. 7 presents a pair plot, or a scatterplot matrix, which is an effective visual assistance that demonstrates the relationships between multiple variables.Feature Scaling: In this case, the target variable is scaled with the help of the np.log1p(x) method.

D. Machine Learning Model :

The proposed method of solar power forecast is highly dependent on machine learning. Accurate solar power prediction models are trained on processed, past, and feature-filtered data.

Linear regression is an approximation model that predicts numerical results with a linear correlation determined by input data and AC power.

Random Forest Regression has several decision trees, which reduce overfitting and enhance accuracy in prediction.

The process of hyperparameter optimization of the Random Forest model is done with the help of grid search CV, optimizing parameters such as the depths of the trees and the number of trees. R 2 score is used to cross-validate the model selection.

XGBoost is effective in most weather conditions and it is also effective in finding nonlinear associations in structured data.

Deep Neural Networks (DNN) consist of multiple hidden layers that replicate complex nonlinear patterns through the use of activation functions and pronounced connections.

E. Model Evaluation and Selection.

EMD Empirical Mode Decomposition EMD is a signal processing technique that separates complex signals into Intrinsic Mode Functions (IMFs). It is effective with nonlinear and time-varying data, e.g. solar power signaling and weather. EMD is capable of improving the quality of forecasting and quality of data collected by reducing noise and highlighting meaningful trends [12].

Figure 8 shows the intrinsic mode functions (IMFs 917) obtained as a result of Empirical Mode Decomposition of the AC power signal. These IMFs contain the information of the signal that is being transmitted.

IMFs 9 and 10 are less noisy in short term fluctuations as compared to the former IMFs. IMF 11 and 12 have a major dynamic variability of the power signal and oscillate with smoother movements. The slow changes in the power usage are indicated by low-frequency oscillations exhibited by IMFs 13 and 14. IMFs 15 and 16 display long-term trends and are characterized by really slow changes. IMF 17 represents the overall trend of the signal.
The IMFs of lower order (IMF 18) were not used in the process of reconstruction since they are primarily composed of high-frequency noise. IMFs 9-17 were selected in order to

generate the denoised signal, and they were added up. Wavelet: The Wavelet Transform (WT) can be used to separate wavelet-varying data and noise. It is capable of capturing time and frequency information due to the analysis of signals at different scales. This improves the quality of data and accuracy of forecasting, reduces noise, and contributes to identifying important features [13].

The original AC power signal that has various noise-associated variations is shown in the first graph. In the second graph, it is shown that most of the noise is removed, and the main power pattern becomes more apparent, which is the EMD denoised signal.

The third graph shows the wavelet denoised signal which is much smoother and has a definite periodic pattern.

Wavelet denoising gives a less noisy signal than EMD with both giving less noise in general.

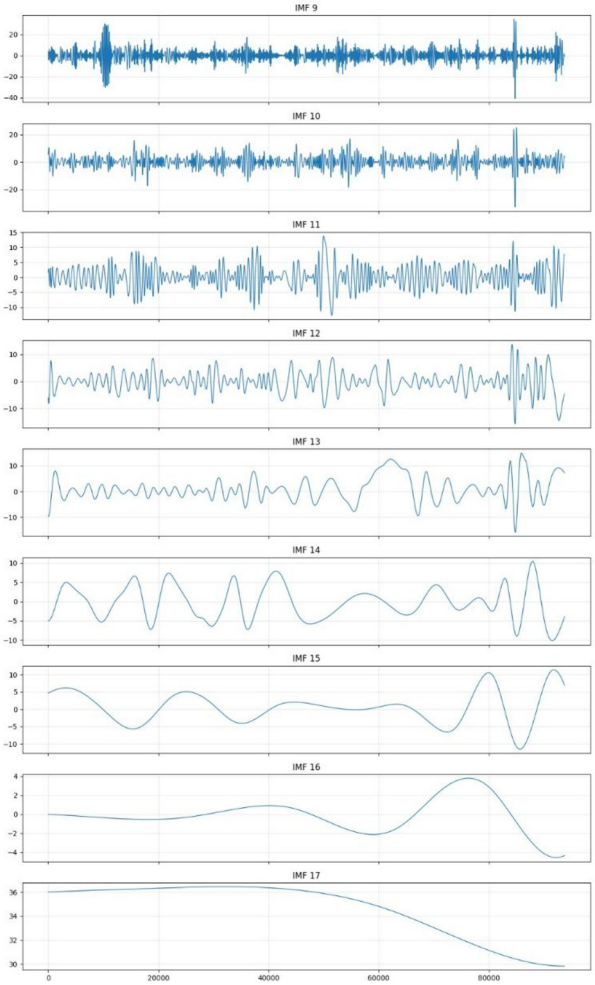

Fig.8: Intrinsic mode functions (IMFs 917) of signal.

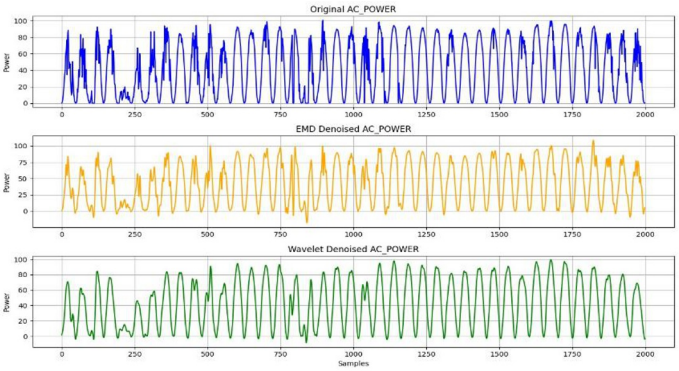

Fig.9: Comparison of Original and Denoised AC Power Signals.

IV. RESULT AND DISCUSSION

The weather and solar data of Annamacharya University is used to evaluate machine learning models to predict solar power generation. The time-series data are denoised with the help of Empirical Mode Decomposition (EMD) and Wavelet Transform prior to training. Deep Neural Network (DNN), Random Forest (default and adjusted), XGBoost, and Linear Regression model are trained and tested on the denoised dataset. The measures of model performance are R^2, MAE, MedAE, and RMSE.

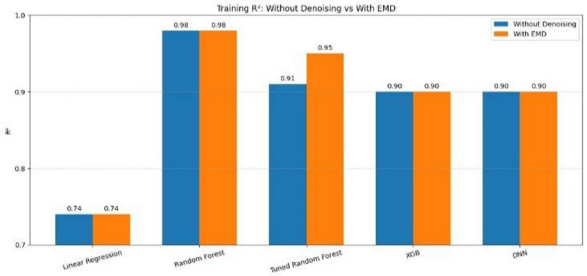

Fig.10: Training R^2 Comparison Between Raw and EMD-Denoised Data

Fig. indicates the training R^2 of machine learning models on the data used in the 15-minute dataset with and without the EMD denoising. The R^2 values of all the models increase after EMD indicating improved learning due to reduction in noise. Random Forest and Tuned Random Forest have the highest R^2 values and therefore they are said to be strong training performance.

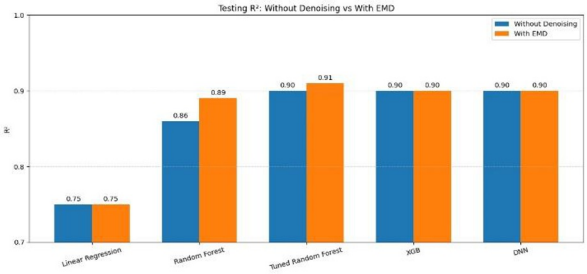

Fig 11: Testing R^2 comparision Between Raw and EMD-Denoised Data

The comparison of machine learning model performance to predict the 15-minute dataset using and without EMD denoising is depicted in Fig.All the models run better with EMD despite the fact that the testing R^2 of the models are slightly less than the training outcomes.

Having the largest R^2, Tuned Random Forest has better generalization and reduced overfitting.

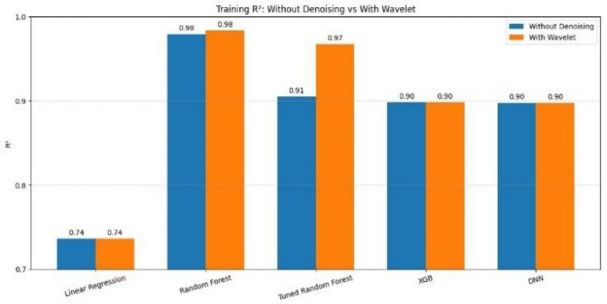

Fig.12: Training R^2 Comparison Between Raw and Wavelet-Denoised Data.

Fig. 12 presents the comparison of machine learning models with and without Wavelet denoising training R^2 of the 15-minute data. The Wavelet Transform is the most effective in training all models with the highest R^2 values being Random Forest and Tuned Random Forest.

Fig. 13 presents the comparison of machine learning models with and without Wavelet denoising training R^2 of the 15-minute data. The Wavelet Transform is the most effective in training all models with the highest R^2 values being Random Forest and Tuned Random Forest

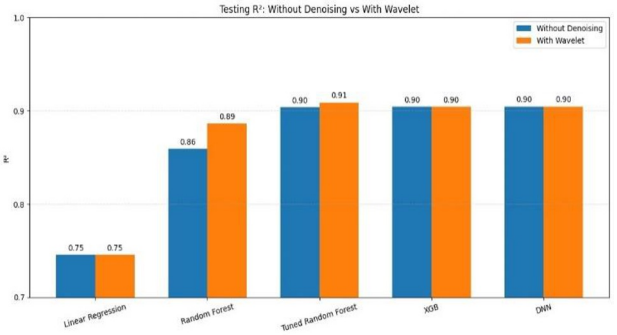

Fig.13: TestingR^2 Comparison Between Raw and Wavelet-Denoised Data.

Table 1: The performance metrics of various models for train dataset

Model	Train	R^2 Score	MAE	MedAE	RMSE
LR	Without denoising	0.7450	15.3754	10.1980	22.4275
	With EMD	0.7504	15.9123	9.798820	27.8181
	With Wavelet	0.7504	16.8608	10.0034	27.9400
RFR	Without denoising	0.9869	3.0291	1.8919	4.4409

	With EMD	0.9833	1.8079	1.0434	3.0512
	With Wavelet	0.9832	2.1574	1.18695	3.59811
Tuned RFR	Without denoising	0.9656	4.6395	2.7854	6.8988
	With EMD	0.9494	3.0956	1.68854	5.0116
	With Wavelet	0.96738	2.6716	1.5156	4.2485
XGBoost	Without denoising	0.8963	8.1119	5.6484	11.4591
	With EMD	0.8981	4.9882	2.86613	7.99062
	With Wavelet	0.8981	4.9882	2.86613	7.99062
DNN	Without denoising	0.8926	8.2405	5.7453	11.6740
	With EMD	0.8972	4.6829	2.43806	7.8497
	With Wavelet	0.8969	5.8725	4.2089	8.6426

Table 2: The performance metrics of various models for test dataset

Model	**Test**	**R^2 Score**	**MAE**	**MedAE**	**RMSE**
LR	Without denoising	0.7396	15.4589	10.3619	22.5334
	With EMD	0.7557	15.5757	9.3929	27.8672
	With Wavelet	0.75765	16.4035	9.6441	27.2669
RFR	Without denoising	0.9070	7.2346	4.3720	10.8379
	With EMD	0.8913	4.76399	2.8414	7.4382
	With Wavelet	0.886267	5.7598	3.20049	9.15791

Tuned RFR	Without denoising	0.9100	7.1125	4.3301	10.6084
	With EMD	0.91421	4.14198	2.31018	6.70699
	With Wavelet	0.90903	4.15210	2.26422	6.77325
XGBoost	Without denoising	0.8918	8.1690	5.6443	11.5871
	With EMD	0.90453	4.87985	2.82827	7.77112
	With Wavelet	0.90453	4.87985	2.82827	7.77112
DNN	Without denoising	0.8891	8.2438	5.6649	11.7289
	With EMD	0.9042	4.5738	2.3910	7.5972
	With Wavelet	0.9037	5.7601	4.1198	8.4404

The results acquired are in agreement with the results of the earlier studies on solar power forecasting. Previous studies have indicated that ensemble learning is more accurate at prediction than traditional regression models like Random Forest and XGBoost. Similar patterns were observed in this research work, where the Tuned Random Forest model recorded the highest R^2 value as well as minimum prediction errors compared to all other models considered, as shown in Table 1 and Table 2. Moreover, it was observed that when denoising models were applied (Empirical Mode Decomposition and Wavelet Transform), the quality of the dataset was enhanced and model performance was improved. These findings validate the claim that the use of preprocessing techniques and machine learning models significantly increases the reliability and accuracy of solar power forecasting.

V. CONCLUSION

This paper introduced a machine learning-based proposing forecasting sun power development on using meteorological and solar generation data taken in the solar power plant at Annamacharya University, Rajampet. Various machine learning models were tested on the basis of performance measures like R^2, MAE, MedAE, and RMSE.

To enhance the model prediction, the EMD signal denoising and Wavelet Transform signal denoising techniques were used to process the solar power data, and then the models were trained. The findings indicated that denoising was very effective in improving data quality, as well as, in improving the prediction performance of machine learning models. The Tuned Random Forest model was the best among all models as it had a higher R^2 and lower values of error prediction.

The suggested solution proves that it is possible to achieve a substantial increase in the precision of solar power prediction with the help of combing denoising methods and machine learning algorithms. This model can help to improve the management of renewable energy and effective application of solar energy into the new smart grid systems. Further

development of work in the field of advanced deep learning models and real-time weather data can be done in the future to enhance the accuracy of predictions.

Acknowledgments. The authors are so grateful to Annamacharya Institute of Technology and Sciences (AITS), Rajampet, to have been given the space and other necessities to execute this piece of research work.

The authors also acknowledge the Department of Electrical and Electronics Engineering to have guided and encouraged them to complete this study. The project guide Prof. M. Padma Lalitha is also given special credit as she made valuable suggestions on the research and gave constant guidance on the research work.

Disclosure of Interests. The authors confirm that they have no competing interests with the publication of the paper. The study activity in this paper was academic and research in nature and no financial or commercial interests were involved in the results and discussion of the findings**.**

REFERENCES

[1] J. Antonanzas, N. Osorio, R. Escobar, A. Urraca, R. Martinez-de-Pison, F. Antonanzas-Torres. Renewable and Sustainable Energy Reviews - "Review of photovoltaic power forecasting" - Project: Photovoltaic Forecasting.

[2] T. Hong, P. Pinson, S. Fan. IEEE Transactions on Smart Grid - "Global energy forecasting competition 2012" - Project: GEFCom. [3] R. Inman, H. T. C. Pedro, C. F. M. Coimbra. Progress in Energy and Combustion Science - "Solar forecasting methods.

[3] R. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," Progress in Energy and Combustion Science, vol. 39, no. 6, pp. 535–576, 2013.

[4] N. Voyant, C. Paoli, M. Muselli, and M. L. Nivet, "Multi-horizon solar radiation forecasting for Mediterranean locations using time series models," Renewable and Sustainable Energy Reviews, vol. 28, pp. 44–52, 2013.

[5] S. Mohandes, M. Rehman, and T. Halawani, "A neural networks approach for wind speed prediction," Renewable Energy, vol. 13, no. 3, pp. 345–354, 1998.

[6] H. T. C. Pedro and C. F. M. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," Solar Energy, vol. 86, no. 7, pp. 2017–2028, 2012.

[7] N. Voyant, M. Muselli, C. Paoli, and M. L. Nivet, "Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation," Energy, vol. 36, no. 1, pp. 348–359, 2011.

[8] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," Solar Energy, vol. 83, no. 3, pp. 342–349, 2009.

[9] R. Randimbivololona, J. Bonnet, and N. Rabeharisoa, "Hybrid forecasting of solar radiation using ARMA and neural networks," Renewable Energy, vol. 34, no. 3, pp. 799–805, 2009.

[10] D. Yang, J. Kleissl, C. A. Gueymard, H. T. C. Pedro, and C. F. M. Coimbra, "History and trends in solar irradiance and PV power forecasting," Solar Energy, vol. 147, pp. 434–453, 2017.

[11] M. Padma Lalitha et al, "Prediction of Solar Power Using Machine Learning – Annamacharya University Case Study," in Proc. 13th IEEE Int. Conf. on Smart Grid (iSmartGrid 2025), Glasgow, U.K., May 27–29, 2025, pp. 207–213. (Annamacharya Institute of Technology and Sciences, Rajampet, India).

[12] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 454, no. 1971, pp. 903–995, 1998.

[13] S. Mallat, A Wavelet Tour of Signal Processing: The Sparse Way, 2nd ed. San Diego, CA, USA: Academic Press, 1999