International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2025, 15(1), 110-120.

Optimized Solar Energy Forecasting for Sustainable Development Using Machine Learning, Deep Learning, and Chaotic Models

Taraneh Saadati, Burak Barutcu*

Energy Institute, Istanbul Technical University, Maslak, Istanbul, 34469, Turkey. *Email: barutcub@itu.edu.tr

Received: 04 September 2024 **Accepted:** 22 November 2024 **DOI:** https://doi.org/10.32479/ijeep.17766

ABSTRACT

This study applies four forecasting approaches—Ensemble Learning (EL), Deep Learning (DL), Machine Learning (ML), and Chaotic modeling—to predict energy production from the Konya Eregli solar power plant in Turkey. Using Python, it incorporates ambient temperature and solar cell temperature as exogenous variables alongside endogenous energy data. A year's worth of 10-min interval data is trained, with two subsequent months forecasted by each model. The False Nearest Neighbors algorithm optimizes the embedding dimension for the chaotic analysis, and an optimized Echo State Network, achieving an R-squared above 0.97, is used for accurate forecasting. Additional models include Long-Short-Term Memory and Gated Recurrent Unit (DL), eXtreme Gradient Boosting and Random Forest (EL), and Extreme Learning Machine and Feed Forward Neural Network (ML). Each model is optimized using the Tree-structured Parzen Estimator, a Bayesian optimization approach. Evaluation metrics reveal all models performed well with the integration of endogenous and exogenous variables, with LSTM achieving the best results. This research advances solar energy forecasting, supporting Sustainable Development Goals (SDGs) related to affordable and clean energy, climate action, and sustainable communities through improved renewable energy management.

Keywords: Time Series Forecasting, Renewable Energy, Chaotic Analysis, Machine Learning, Deep Learning, Sustainable Development
JEL Classifications: P28, Q01, Q47

1. INTRODUCTION

As solar energy generation becomes more prevalent, precise forecasting of solar power output becomes essential for optimal utilization of resources and efficient grid management (Hernández-Torres et al., 2015). A large portion of the studies aim to forecast the energy using meteorological factors such as solar radiation instead of endogenous inputs like output electricity (Kostylev and Pavlovski, 2011; Bizzarri et al., 2012; Jiménez-Pérez and Mora-López, 2014; Lauret et al., 2015), because of its dependence on nonlinear parameters and more difficulty to be modeled and predicted than natural parameters. Advancing a sustainable future strategy offers significant potential to facilitate the transfer of advanced production technologies between countries, create effective opportunities for managing natural resources, and foster collaboration in addressing environmental challenges (Erdoğan et al., 2021). Solar energy is a rapidly growing and sustainable source of power that plays a crucial role in the transition to renewable and clean energy. The transition to low-carbon energy systems, which is essential for addressing climate change and energy security issues, is significantly propelled by the widespread adoption of renewable energy technologies (Tiba and Belaid, 2021). Forecasting the spread of renewable energy technologies is crucial for developing an effective energy agenda and establishing realistic targets for electricity generation. Time series forecasting techniques, coupled with statistical, ML, DL, and artificial intelligence (AI) approaches, have proven to be powerful tools for accurately and reliably predicting solar power generation patterns (Vrettos and Gehbauer, 2019).

AI approaches combine statistical, machine learning, and deep learning techniques to enhance the accuracy and robustness of solar power time series forecasting. AI-based forecasting systems leverage the strengths of various models, combining their predictive capabilities to achieve more accurate and reliable results. By integrating real-time data, historical records, and meteorological forecasts, AI systems enable dynamic and adaptive forecasting, ensuring optimal management of solar energy resources. Eventually, solar power time series forecasting is a critical aspect of effective grid management and resource planning in the renewable energy sector. By employing statistical, ML, DL, and AI approaches, accurate predictions of solar power output can be achieved, enabling stakeholders to optimize energy utilization, make informed decisions, and pave the way for a cleaner and more sustainable future. Utilization of DL and EL techniques in the field of solar energy forecasting is appearing in the recent studies comparing to Support Vector Machine (SVM) and Artificial Neural Networks (ANN), which are the two most implemented approaches during the past decades (de Freitas Viscondi and Alves-Souza, 2019). Moreover, there are limited number of studies using Chaotic analysis for solar energy forecasting, specifically using the Chaos based neural network model: Echo State Network (ESN).

The focus on solar power forecasting aligns closely with Goal 7: Affordable and Clean Energy, and Goal 13: Climate Action, of the United Nations Sustainable Development Goals (SDGs) (Su et al., 2024; Baiardi, 2023). There are factors indicated by the empirical studies, such as economic growth and energy consumption, that impact climate change on national and global levels (Erdoğan et al., 2022). By advancing the accuracy of solar energy forecasts, this research supports the broader agenda of increasing the share of renewable energy in the global energy mix, thereby promoting sustainable and resilient energy infrastructure while combating climate change—a crucial goal given that air pollution is a key factor influencing sustainable growth (Ozturk et al., 2022).

This study addresses the complex, nonlinear nature of solar power generation, which is often overlooked in traditional forecasting methods that predominantly rely on meteorological or market data. By leveraging the False Nearest Neighbors (FNN) method, this study optimizes the embedding dimension of the time series data. This optimization is critical for accurately capturing the dynamics of solar energy production, a step often neglected in other studies. The significant reduction in FNN ratio highlights the improved representation of system dynamics, demonstrating the effectiveness of this method. Our results show that the exogenous variables enhance the accuracy of the ESN model by providing additional relevant information, thus reducing the FNN ratio and improving the overall predictive performance.

Utilizing data from the Konya Eregli Solar Power Plant, our study provides empirical validation of the proposed methodology in a real-world setting. This practical application underscores the model's robustness and relevance, offering valuable insights into solar energy forecasting.

This research supports the achievement of several SDGs by enhancing the efficiency and reliability of solar energy forecasting. Improved predictions can lead to better management of renewable energy resources, contributing to affordable and clean energy (SDG 7), climate action (SDG 13), and sustainable community development (SDG 11).

2. LITERATURE REVIEW

The literature contains numerous examples of ML algorithms applied for forecasting purposes across various fields (Keskin et al., 2024). Accurately predicting photovoltaic power is crucial for managing the electricity network and ensuring balance between energy generation and consumption. Clearly, accurate predictions allow for efficient use of energy resources and optimal operation of production units (David et al., 2016). In the direct forecasting model, historical data, including PV electricity output and related meteorological data, are used to predict photovoltaic power production. Within this context, the work of (Das et al., 2018) offers a thorough and structured review, extensively covering how the Interrelation between input and output data affects results and highlights the advantages of optimizing these forecasting models.

Considering the type of input data used in the solar energy production forecasting field, and also the most implemented forecasting methods, authors in (de Freitas Viscondi and Alves-Souza, 2019) provide a comprehensive Systematic Literature Review (SLR). The results of this SLR indicate that among the various algorithms, the ANN and SVM methods are the two most commonly used approaches; also, out of the 38 analyzed papers, only 7 used electricity-related input data, and 10 reported data quality assessments in their studies. To further evaluate the mentioned study gap, a number of different papers within this scope are investigated with regards to their considered input data for forecasting models, and availability of data pre-process in their studies. Table 1. presents the summarized results.

According to the results in Table 1, More than half of the papers analyzed in this section of the study have based their research on meteorological data or other external input data, like electricity market-related data, which is not directly related to electricity itself. As a result, these studies mainly emphasize the meteorological factors involved in energy conversion instead of the electrical infrastructure. Also, more than half of the investigated papers have not report data quality assessments, which its significance is evident.

3. MATERIALS AND METHODS

3.1. Machine Learning Models

3.1.1. Feed forward neural network

Feedforward neural networks (FFNN) (Bebis and Georgiopoulos, 1994) are a simple form of artificial neural network in which data moves only from the input layer to the output layer, passing through one or more hidden layers, in a single direction and with no loops or feedback. Each layer is made up of connected neurons (nodes), that process and transmit data. In this network, inputs

Table 1: Summary of investigated studies considering their input data and data quality assessments

Article	Electricity generation Data	Reporting data quality assessments
(Kostylev and Pavlovski, 2011)	No	No
(Bizzarri et al., 2012)	No	No
(Jiménez-Pérez and Mora-López, 2014)	No	No
(Lauret et al., 2015)	No	Yes
(David et al., 2016)	No	No
(Huang et al., 2010)	Yes	No
(Ding et al., 2011)	Yes	No
(Pedro and Coimbra, 2012)	Yes	Yes
(Kühnert et al., 2013)	No	No
(Yona et al., 2013)	No	No
(Hanna et al., 2014)	Yes	No
(Dambreville et al., 2014)	No	Yes
(Aggarwal and Saini, 2014)	No	Yes
(Gandelli et al., 2014)	Yes	Yes
(Russo et al., 2014)	Yes	No
(Dolara et al., 2015)	Yes	No
(Li et al., 2016)	No	No
(Gala et al., 2016)	No	No
(Li et al., 2016)	Yes	No
(Lou et al., 2016)	No	Yes
(Aybar-Ruiz et al., 2016)	No	No
(Li et al., 2016)	No	Yes
(Jiang et al., 2017)	No	No
(Assouline et al., 2017)	No	Yes
(Doorga et al., 2019)	No	No
(Basurto et al., 2019)	No	Yes
(Amiri et al., 2021)	No	Yes

enter the input layer, and as the data moves through the hidden layers, the network performs computations and transformations, with each neuron processing the weighted sum of its inputs through an activation function. The final processed data is then produced at the output layer.

3.1.2. Extreme learning machine

Extreme learning machines (ELM), proposed by (Huang et al., 2004), are feedforward neural networks with single hidden layer. They are known for their fast-learning speed, strong generalization ability, and capability to approximate any function. The algorithm for extreme learning machines works by first randomly assigning biases and weights to the hidden layer. Then, it computes the hidden layer output matrix by applying these weights, biases, and activation functions to the input data. Next, it calculates the output weight matrix by multiplying the Moore-Penrose inverse of the output matrix from the hidden layer with the training data matrix. Finally, this output weight matrix is used to make predictions on new data.

3.2. Deep Learning Models

3.2.1. Long-short-term-memory

The Long Short-Term Memory (LSTM) Neural Network (Hochreiter and Schmidhuber, 1997) is regarded as one of the top models for time series forecasting. According to Korstanje's book (Korstanje, 2021), the LSTM cell significantly boosts long-term memory by enabling the learning of additional parameters. This capability makes it the most powerful Recurrent Neural Network (RNN) for forecasting, particularly effective with data showing long-term trends. Currently, LSTMs are recognized as state-of-the-art models for forecasting (Korstanje, 2021).

3.2.2. Gated recurrent unit

The Gated Recurrent Unit (GRU) (Cho et al., 2014) is a modern Recurrent Neural Network (RNN) similar to the LSTM. It addresses the vanishing gradient problem seen in standard RNNs by utilizing update and reset units. These units decide what information needs to be passed to the output. They can be trained to retain information from several previous time steps without extending the temporal range too much or to discard irrelevant information for better predictions. With careful training, GRUs can perform exceptionally well, even in complex situations.

3.3. Ensemble Learning Models

3.3.1. Random forest

Random forests (RF) (Breiman, 2001) consist of several decision trees, where each tree is influenced by the values of a randomly chosen set of features, applying an identical distribution across all trees. With the increase in the number of trees in the forest, the generalization error approaches a limiting value. This error is affected by the effectiveness of the individual trees and their intercorrelation. Randomly choosing features for each node's split results in error rates similar to those of Adaboost (Schapire, 2013) but are more resilient to noise.

3.3.2. Extreme gradient boosting

XGBoost, or eXtreme Gradient Boosting (Chen and Guestrin, 2016), is an enhanced version of the gradient boosting algorithm. It is highly regarded for its effectiveness and is commonly used in machine learning competitions. XGBoost is known for its high predictive accuracy and is nearly 10 times faster than other gradient boosting methods. It includes several enhancements to prevent overfitting and improve overall performance. Suitable for both regression and classification tasks, XGBoost is designed to efficiently manage large and complex datasets.

3.4. Chaotic Model- Echo State Network

Chaos-based neural networks combine principles from chaos theory and neural networks to create models capable of complex behavior and pattern recognition. Echo State Networks (Jaeger, 2001) are the most popular way to use the reservoir computing (RC) method (Shahi et al., 2022). RC is mainly based on the ideas of recurrent neural networks, but it has a simpler way of training. In RC, only the weights of the output layer are trained, whereas the values of other parameter are assigned randomly and not trained further. Even with this simplified approach, ESNs have been successful in predicting future steps in modeling chaotic dynamic systems and nonlinear time series without needing a lot of computing power (Bianchi et al., 2017).

3.5. False Nearest Neighbors

False nearest neighbors (Kennel et al., 1992) is a technique used to identify the suitable embedding dimension for a time series. It helps in identifying whether the reconstructed state space using a certain embedding dimension is sufficiently capturing the characteristics of the system that the time series represents. If the embedding dimension is too low, nearby points in the reconstructed space might not actually correspond to neighboring points in the original space, leading to false results. FNN analysis helps in avoiding this by finding the minimum embedding dimension required to accurately represent the dynamics of the system. So, while the false nearest neighbors' approach is primarily used for determining the appropriate embedding dimension, the process inherently involves embedding the time series data into a higher-dimensional space. Therefore, it can be considered as a method for embedding time series data.

3.6. Bayesian Optimization-Tree-Structured Parzen Estimator

Bayesian optimization approach provides promising result in optimizing the objectives with high computational cost, and "black boxes" functions which do not have a clear explanation, as well as the functions that can only be evaluated by noisy mechanism (Garnett, 2023). It is especially useful for scenarios in which the function evaluations are costly, for example hyperparameter tuning in machine learning models, engineering designs optimization, or even experimental sciences where each measurement requires substantial resources. The Tree-structured Parzen Estimator (TPE) is a popular Bayesian optimization (BO) technique known for its exceptional performance across numerous applications (Watanabe, 2023). TPE has been crucial for hyperparameter optimization in deep learning models, contributing to victories in Kaggle competitions (Watanabe and Hutter, 2022). Additionally, (Watanabe et al., 2022) was the winner of the AutoML 2022 competition on "Multiobjective Hyperparameter Optimization for Transformers" with implementation of TPE.

3.7. Python Libraries

The following Python libraries are used for tasks ranging from data loading and preparation to model building and assessment, and finally to hyperparameter optimization and visualization of results: TensorFlow: For building and training the deep learning and neural network (NN) models, Pandas: For analysis and data manipulation. NumPy: For numerical operations. Matplotlib and Seaborn: For plotting and visualization. scikit-learn: For machine learning utilities, building RF algorithm, and implementation of false nearest neighbors' approach. XGBoost: For building the XGBoost model. math: For mathematical functions. pyESN: For creating and working with Echo State Networks. Hyperopt: For hyperparameter optimization. Keras: From the TensorFlow package for additional model building and optimization (some functions are explicitly from Keras within Tensorflow).

3.8. Evaluation Metrics

The performance of the trained models is examined on unseen (test) data by analyzing the difference between real and forecasted values using five different metrics: Mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and R-squared (R^2). n represents the number of data points or observations, y_i denotes the actual value of the i-th observation and $\hat{y}_i$ refers to the predicted value of the i-th observation.

3.8.1. Mean absolute error

This metric is determined by taking the average of the absolute differences between the actual and predicted values. MAE is calculated by using the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

3.8.2. Mean squared error

MSE is an evaluation metric that calculates the average of the squared differences between the actual and predicted values. It is determined by the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

3.8.3. Root mean squared error

RMSE is an evaluation criteria calculating the square root of the mean of the squared deviations from the actual and predicted values. It is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

3.8.4. Mean absolute percentage error

The Mean absolute percentage error (MAPE) is an evaluation indicator calculating the mean of the absolute percentage errors between the actual and predicted values. It makes the error scale-independent and easier to interpret in relative terms. It is calculated by the below formula:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

3.8.5. R-squared

R-squared (R^2) is an evaluation metric that measures the effectiveness of fit for a regression model. It indicates the extent to which the model's predictions fit the actual data, with values ranging from 0 to 1, where 1 indicates perfect fit. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

$\hat{y}_i$ is the predicted value of the i-th observation and $\overline{y}$ is the mean of the actual values.

4. EMPIRICAL STUDY

4.1. Data

Data is collected from Konya Eregli Solar Power Plant (SPP), a 9.20 MWp SPP constructed as a project by "Tegnatia Enerji" company, in the Ereğli district of the Konya city in Turkey, which has successfully been grid-connected as of April 2016.

Data consist of two parts; first part is the endogenous data, the generated energy time series of eight inverters of the SPP, with 10 min time interval from June 06, 2021 to December 08, 2022, which is the target variable for every forecasting model included in this study. Second part is the exogenous meteorological data including the 10 min time interval ambient temperature and the solar cell temperature of the Konya Ereğli SPP for 2021 and 2022.

In the forecasting models of this study, the only variable being forecasted (target variable) is the energy production, however, there are different combinations of endogenous and exogenous input variables for the forecasting models. So, the time series forecasting problem that is studied within this paper is a multivariate input and univariate output time series problem, with combination of weather and temporal features as exogenous input variables. Using a multivariate approach in this context enables us to leverage the information from all the related input time series to better understand the dynamics and patterns within the data (Brownlee, 2018).

4.2. Data Cleaning and Preprocessing

The collected energy production data is preprocessed using Python programing. These steps include removing the missing values, modifying the scale of the data and removing outliers using Local Outlier Factor (LOF) (Aggarwal and Aggarwal, 2017). Clean raw data is shown in Figure 1. All the data analysis within this study is carried out in Python 3.10.13.

4.3. Exploratory Data Analysis

Analyzing the time series properties is essential for both researchers and policymakers, as energy production and consumption are closely tied to economic performance (Öztürk and Aslan, 2015). The Exploratory Data Analysis consists of methods and analyses which search for the key characteristics of the time series data. These analyses are investigating autocorrelation, seasonality and stationarity. The obtained features will lead us to perform needed tests and modifications in order to enter the modelling part. Some of the important features of the data can be found by plotting them and because of that, data visualization can be very helpful. Figure 1 shows the energy production values of all the eight inverters.

4.3.1. Visualization

To simplify the computations, only one inverter is selected among the eight inverters data for this case study, which is the energy production data of inverter 4 shown in Figure 2.

4.3.2. Autocorrelation

The distance between each two peak values in the autocorrelation plot, presented in Figure 3 represents the seasonal length in our

Figure 1: Energy production time series of all inverters in the collected data

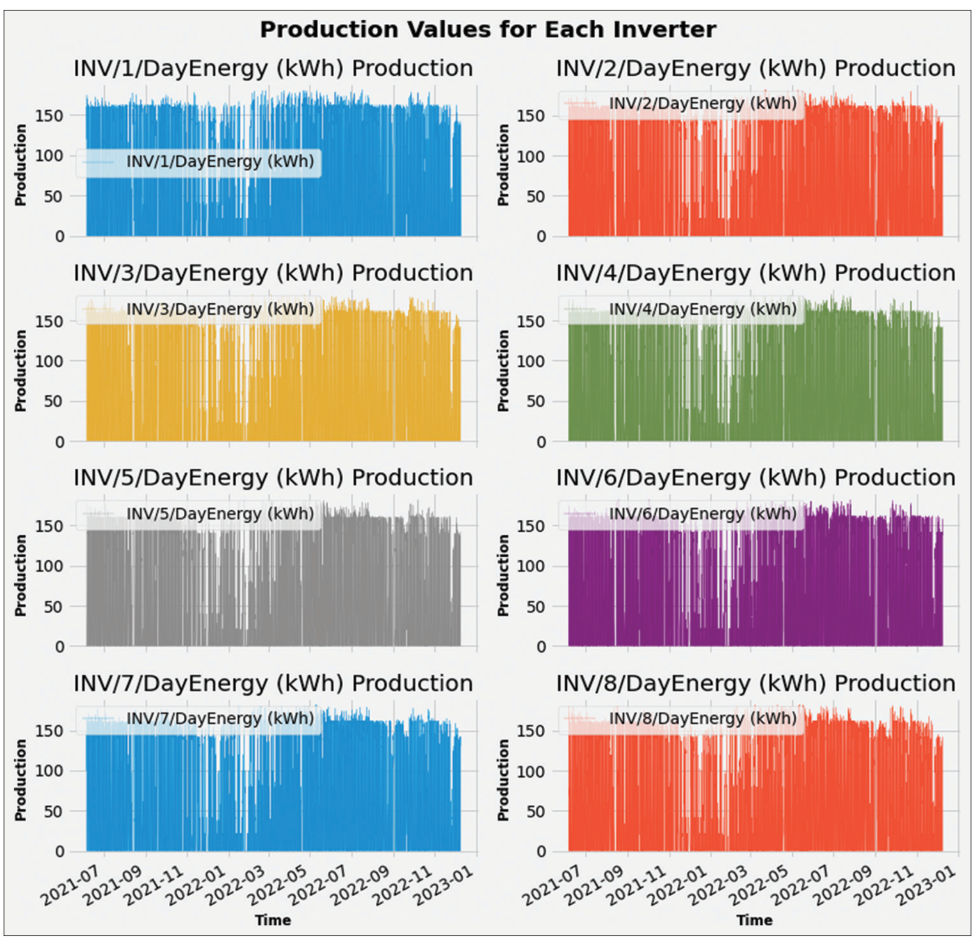


Figure 2: Energy production data in inverter 4. (a) shows the entire data (70973) datapoints and (b) shows the production data for (1000) datapoints

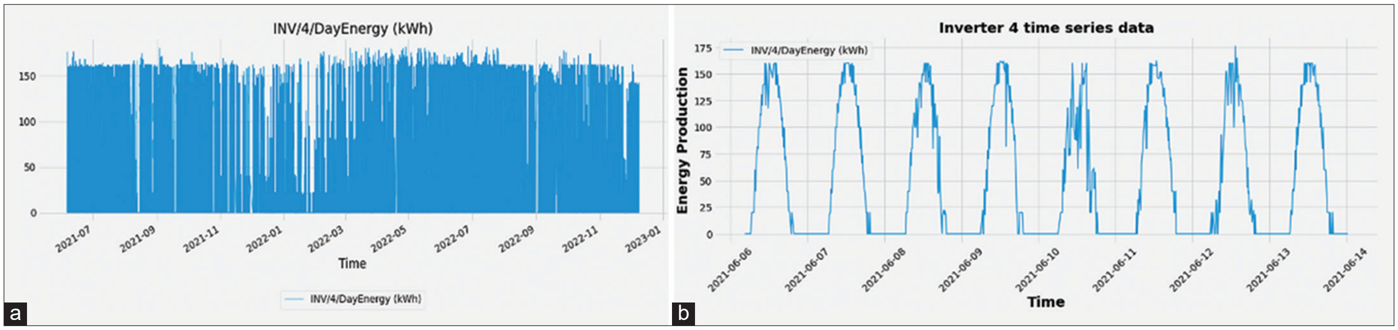
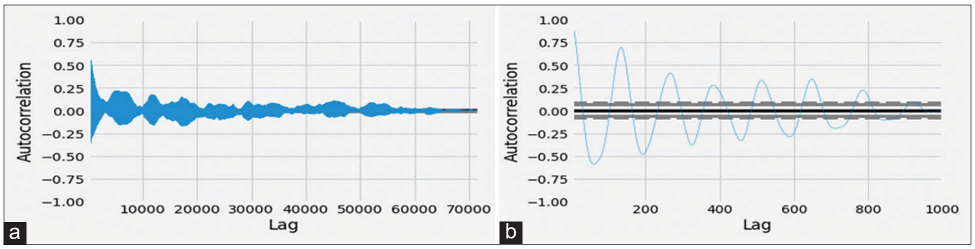


Figure 3: Autocorrelation plot for inverter 4. (a) For the entire (70973) datapoints and (b) for (1000) datapoints



data. It was also clearly shown in the production plots, that the data has a daily seasonality. By calculation of two consecutive peak values' distance, the obtained time is around 86,400 s, that is equal to 24 h indicating the seasonal length of our data. It can be seen in Figure 3 that the amplitude of the plot is reducing as the time passes. This refers to the change of the day length during the year, although the seasonal length will be constant in the whole year, because each day has 24 h in a year, but the length of the sunlight hours changes and days become shorter when getting closer to the winter and the changes between the sunrises and sunsets in the following days cause the shape of the presented autocorrelation plot.

4.3.3. Stationarity test

In the scope of Exploratory Data Analysis, the last criterion that needs to be evaluated is Stationarity. Augmented Dickey Fuller (ADF) test (Cheung and Lai, 1995) is implemented in Python for this purpose, by using the Staststmodel library (Seabold and Perktold, 2010). With a highly negative test statistic and a very low P-value in the ADF test result, the null hypothesis of a unit root is confidently rejected. Therefore, the data is likely stationary, indicating that differencing may not be necessary for this time series.

5. RESULTS AND DISCUSSIONS

In this section we present the results of our study's hyperparameter optimization results, using Tree-structured Parzen Estimator are given in section 5.1. Section 5.2 shows the forecasting results from the deep learning, machine learning, and ensemble learning approaches before and after the optimization. Section 5.3 discusses the chaotic analysis, embedding data results and forecasting results of the ESN model.

5.1. Hyperparameter Tuning

All the forecasting models in this study are optimized using Tree-structured Parzen Estimator. The optimized values of the hyperparameters for each model are reported in Table 2.

5.2. Deep Learning, Machine Learning and Ensemble Learning Models

Three groups of forecasting approaches: Deep learning, machine learning, and ensemble learning models are implemented on the endogenous energy production data of Konya Eregli SPP, integrated with ambient temperature and solar cell temperature exogenous variables, to forecast the energy production. A year of data with a 10-min time interval is used for training and two consecutive months are forecasted by each model. It is noteworthy that all the models performed in this section are evaluated across different number of time lags and different temporal exogenous variables extracted from timestamps, to select the configuration with the highest accuracy in forecasting. These temporal features are as follows: "day of year," "hour," "day of week," "quarter," "month" and "year." The forecasting results showed that integration of temporal features and weather exogenous variables alongside with 4 time-lags of the endogenous variable (energy production historical data) creates the best configuration for the ensemble learning approaches (RF and XGBoost), providing the most accurate results. However, for the NN based approaches (ML and DL employed algorithms: FFNN, ELM, LSTM and GRU), integration of only weather exogenous variables with 5 time-lags endogenous inputs leads to the highest performance. The employed features structure for ML, DL and EL groups of models is shown in Figure 4.

The optimum configuration for both groups of the methods is demonstrated in Table 2. It is noteworthy to mention that for each model only the value of most important hyperparameters which

Figure 4: Input features structure. (a) ML and DL groups of models. (b) EL models

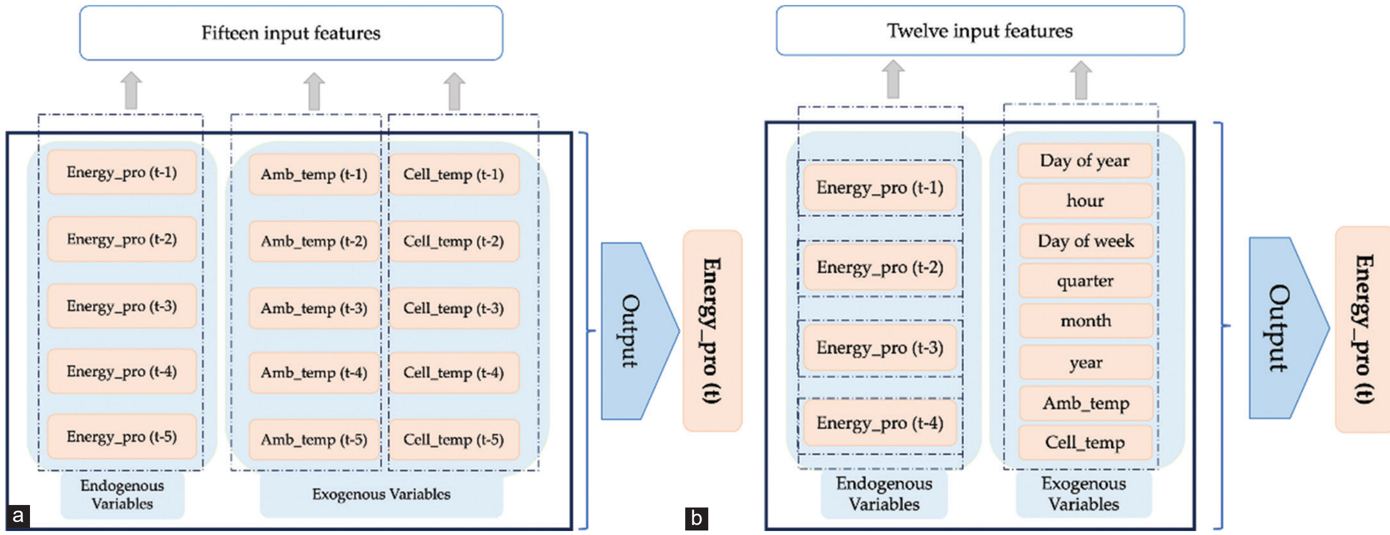


Table 2: Optimized values of the hyperparameters

Model	Hyper-Parameter	Optimized value
LSTM	LSTM units number Learning rate	128
	Batch size	0.0005
	Dropout rate	64
		0.1004
GRU	GRU units number	128
	Learning rate	0.0002
	Batch size	32
	Dropout rate	0.2478
FFNN	Hidden layers number	2
	Neurons number	32
	Learning rate	0.0014
	Batch size	64
	Dropout rate	0.0223
ELM	Hidden Size	190
	Weight lower bound	−0.72
	Weight upper bound	0.86
	Regularization	L2
	L2 lambda	0.02
RF	n_estimators	298
	max_features	10
	max_depth:	14
	min_samples_split	13
	min_samples_leaf	4
XGBoost	colsample_bytree	0.982
	gamma	0.26
	Learning rate	0.2129
	max_depth	5
	min_child_weight	3.5568
	n_estimators	398
	subsample	0.7375
ESN	Number of reservoir units	778
	Sparsity of the reservoir connections	0.021
	Spectral radius of the reservoir weights	0.357
	Magnitude of noise to be added to the reservoir activations	0.0531

affect the model's performance more strongly are reported in Table 2 for the sake of shortness, while there are more numbers of hyperparameters in each model that are optimized in this study with TPE algorithm.

Table 3 shows the forecasting result of each model, before and after optimization across all the calculated evaluation metrics, with most accurate forecasts highlighted in bold. It reveals that the forecasting accuracy by means of R^2, RMSE and MSE is better in the LSTM model, however, the MAPE and MAE metrics indicates a better performance in the GRU model. In the ML group models, FFNN significantly performs better than ELM across all the metrics, and in the EL group, RF and XGBoost models' performance is relatively the same. The comparison of the models across evaluation metrics can be better observed in Figure 5.

For future work, a potential direction would be to apply the proposed model to forecast solar energy production by integrating an IoT-based device connected to the cloud. This approach would facilitate the prediction of critical parameters affecting solar energy production and offer meaningful perspectives for optimizing energy generation. For instance, an alert system for predictive maintenance could be created by deploying the developed model on an Edge Device to monitor and analyze the real-time data and detect anomalies in solar power plants.

5.3. Chaotic Approaches

This section presents the results of embedding the time series energy production data using the FNN algorithm, as well as optimizing the embedding dimension by FNN itself in part 5.3.1. Then the optimized ESN model, with tuned hyperparameters that are reported in Table 2, is implemented both on the original time series data and on the embedded data, and the findings are summarized in Table 4 in section 5.3.2.

5.3.1. Optimization of the embedding dimension using false nearest neighbors

The FNN function defined to embed the data in this study, performs the false nearest neighbors analysis on the input time series data via the following steps:

- Embedding the time series data
- Calculating the distances between each embedded point and its nearest neighbor in the embedded space

Figure 5: Comparison of models' performance

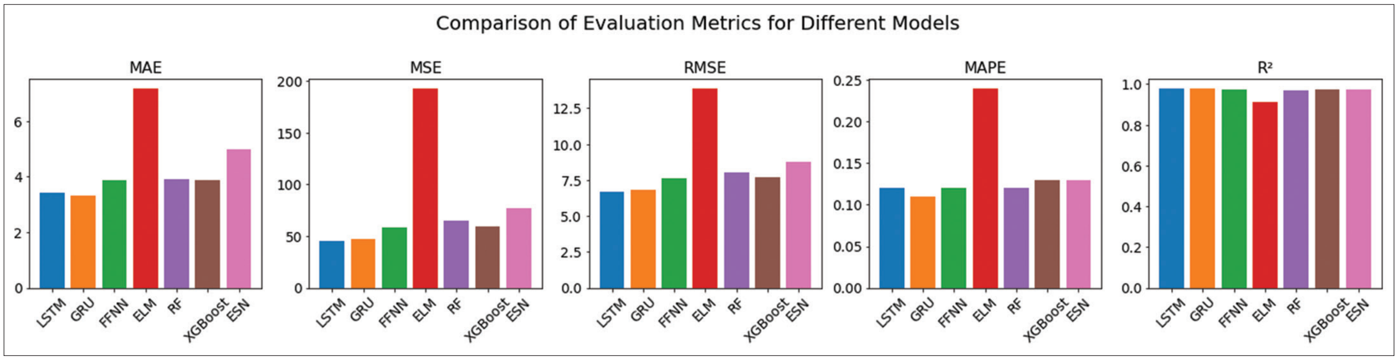


Table 3: Forecasting results of ML and DL models

Metrics/methods	R-squared	MAPE	RMSE	MSE	MAE
Long-short-term memory					
Before optimization	0.976	0.13	7.04	49.57	3.81
After optimization	0.979	0.12	6.70	44.91	3.42
Gated recurrent unit					
Before optimization	0.976	0.12	7.13	50.87	3.68
After optimization	0.978	0.11	6.84	46.76	3.31
Random forest					
Before optimization	0.971	0.12	8.06	65.00	3.91
After optimization	0.971	0.12	8.06	64.96	3.89
Extreme gradient boosting					
Before optimization	0.961	0.14	9.16	83.84	4.37
After optimization	0.972	0.13	7.71	59.40	3.88
Extreme learning machine					
Before optimization	0.76	0.44	22.76	518.00	13.37
After optimization	0.91	0.24	13.88	192.73	7.17
Feed forward neural network					
Before optimization	0.972	0.12	7.76	60.16	3.75
After optimization	0.973	0.12	7.64	58.39	3.86

Table 4: Forecasting results of ESN model

Metrics/methods	R-squared	MAPE	RMSE	MSE	MAE
Echo state network – original data					
Before optimization	0.961	0.17	9.26	85.71	5.13
After optimization	0.964	0.16	8.85	78.27	4.94
Echo state network – embedded data					
Before optimization	0.972	0.13	8.80	77.48	5.02
After optimization	0.972	0.13	8.79	77.26	4.99

- Perturbing each point in the embedded space by adding random noise
- False nearest neighbors calculation
- Calculation of FNN Ratio, denoting the share of false nearest neighbors in the total set of points
- Returning the embedded data.

The FNN ratio that is calculated by this function provides insights into whether the chosen embedding parameters effectively capture the underlying dynamics of the time series. Also, it indicates the proportion of points in our embedded phase space that are falsely identified as nearest neighbors after perturbation. In this section we systematically vary the embedding dimension (m) and calculate the FNN ratio for each dimension. We define the maximum dimension to be 100 and run the optimization algorithm for this range. The optimal embedding dimension is often chosen as the smallest value of m for which the FNN ratio stabilizes or reaches a minimum, which is found to be 70 in our analysis. The FNN ratio after optimization of the embedding dimension is changed from 0. 277 to 0.007 for our case study dataset, which is the merged data of target variable and exogenous variables.

Our obtained FNN ratio is considered as a very low ratio (bellow 0.05) and indicates that the reconstructed phase space with the optimum embedding dimension of 70, more accurately captures the underlying dynamics of the system generating the time series data. This also suggests that there are relatively very few false nearest neighbours, meaning that nearby points in the embedded space remain close after perturbation. This enhanced representation increases confidence in using the embedded data for further analysis or modelling tasks such as prediction, as it better preserves the structure and relationships present in the original time series, which can be clearly observed in the presented prediction results in section 5.3.2.

We can visualize the embedded phase space to explore the structure of the data and identify regions where false nearest neighbors occur. Since the embedded phase space is typically high-dimensional, we need to use dimensionality reduction techniques such as principal component analysis (PCA) (Jolliffe, 2005) to project the data onto a lower-dimensional space for visualization. PCA is a dimensionality reduction technique that converts high-dimensional data into a lower-dimensional space, aiming to preserve as much variance in the data as possible. The principal components are the orthogonal axes in the new lower-dimensional space that capture the directions of maximum variance in the original data.

- We use PCA to reduce the dimensionality of the embedded phase space to two dimensions.
- The embedded data is then transformed using PCA.
- Finally, we plot the transformed data in the two-dimensional PCA space.

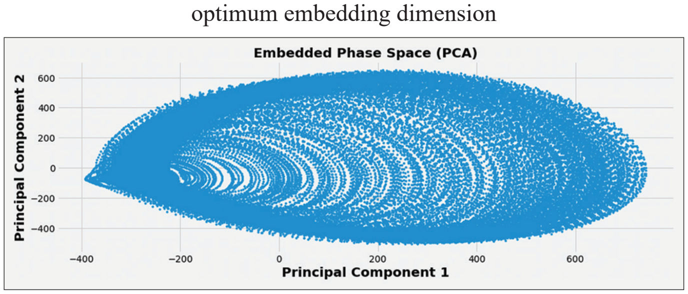
The visualization of the embedded phase space using PCA for dimensionality reduction is presented in Figure 6. Regions where points are clustered closely together may indicate regions of the phase space with similar dynamics, while regions with more scattered points may indicate areas where false nearest neighbors occur. Principal component 1 (PC1) and Principal component 2 (PC2) are the first and second principal components obtained from the PCA of the embedded phase space data. When we apply PCA to the embedded phase space data, each data point in the embedded space is transformed into a new point in the PCA space, represented by its coordinates along the principal components.

In Figure 6, PC1 and PC2 correspond to the horizontal and vertical axes, respectively. These axes are chosen such that PC1 captures the direction of maximum variance in the data, followed by PC2 capturing the direction of the second maximum variance orthogonal to PC1. The scatter plot of the embedded phase space in the PCA space shows the distribution of the data points along these principal components. Each point in the plot represents an embedded vector from the original time series data, projected onto the two principal components.

5.3.2. ESN model

To observe the effect of the chaotic analysis explained in section 5.3.1 more precisely, ESN model which is a chaos based neural network model successful in predicting future steps in nonlinear time series and modeling chaotic dynamic systems, is implemented 2 times: first on the original time series data, and 2nd time on the embedded time series energy production data. As the results reveal in the Table 4, the forecast accuracy across all the evaluation metrics, only except the MAE, has been improved when implementing the ESN model on the embedded time series data with reconstructed phase space comparing to the original data.

Figure 6: Visualization of the embedded phase space with the optimum embedding dimension



6. CONCLUSION

Forecasting solar energy output is essential for optimizing the efficient use of this renewable resource. In this study, our objective is to meet the demand for precise energy production forecasting by using data from a 9.20 MWp SPP. We employ deep learning, machine learning, and chaos-based forecasting models for comparison. The findings confirm that the implemented DL, ML, and chaos-based neural network models perform well for forecasting time series energy production data in our specific scenario. They also provide the optimal parameters for predicting solar energy production with each forecasting algorithm. All the nonlinear implemented forecasting models in this study, with the optimized configuration, succeed in providing accurate forecasts with R-squared metric above 0.97 except the ELM model with R-squared: 0.91. By comparing the forecasting results of ML and DL models, before and after optimization across all the calculated evaluation metrics, it is seen that the accuracy metrics of R-squared, RMSE and MSE is best in the LSTM model among all the models, and the MAPE and MAE metrics achieves their lowest value in the GRU model. It can be concluded that the performance of the implemented DL models in this study outperforms the ML models.

By implementing chaotic analysis and optimizing the embedding dimension using the FNN algorithm, we determined the optimal dimension to be 70. The FNN ratio improved from 0.277 to 0.007 for our dataset, indicating that the reconstructed phase space with this optimal embedding dimension could more accurately capture the system's underlying dynamics generating the time series data. This enhances confidence in using the embedded data for further analysis or modeling with chaos-based approaches, as it better preserves the structure of the original time series. This assumption is validated by the results of implementing the ESN model on the embedded time series data, which show improvements in all accuracy metrics except MAE. Notably, the R-squared metric exceeds 0.97.

By identifying the optimal configuration of the implemented algorithms, particularly the deep learning models LSTM and GRU, which provide the most accurate forecasting results, we offer useful insights for practical applications. These insights can help improve solar energy systems, support a sustainable future, and reduce reliance on non-renewable resources. Additionally, when applied to edge devices for predicting maintenance needs in solar power plants, these insights can enhance operational efficiency.

Furthermore, this research contributes to the Goal 7: Affordable and Clean Energy, and Goal 13: Climate Action, of the United Nations Sustainable Development Goals. By enhancing the predictability and reliability of solar power generation, this work supports the global effort to increase renewable energy adoption, improve energy efficiency, and mitigate the impacts of climate change.

7. ACKNOWLEDGMENTS

We would like to extend our gratitude to "Tegnatia Enerji" for generously providing the data essential to our research. Their support and contribution have been invaluable to the success of this study.

8. FUNDING

This work was supported by the Scientific Research Projects (BAP) Coordination Unit at Istanbul Technical University [Project ID: 44133, Project code: MDK-2022-44133].

9. RESEARCH DATA

Original data of this study including solar energy production data, exploratory data analysis, all the implemented models and algorithms are available at Mendeley Data (DOI: 10.17632/2tpv28kr83.1).

REFERENCES

Aggarwal, C.C., Aggarwal, C.C. (2017), An introduction to outlier analysis. Germany: Springer International Publishing. p1-34.

Aggarwal, S.K., Saini, L.M. (2014), Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013-14 Solar Energy Prediction Contest. Energy, 78, 247-256.

Amiri, B., Gómez-Orellana, A.M., Gutiérrez, P.A., Dizène, R., Hervás-Martínez, C., Dahmani, K. (2021), A novel approach for global solar irradiation forecasting on tilted plane using Hybrid Evolutionary Neural Networks. Journal of Cleaner Production, 287, 125577.

Assouline, D., Mohajeri, N., Scartezzini, J.L. (2017), Quantifying rooftop photovoltaic solar energy potential: A machine learning approach. Solar Energy, 141, 278-296.

Aybar-Ruiz, A., Jiménez-Fernández, S., Cornejo-Bueno, L., Casanova-Mateo, C., Sanz-Justo, J., Salvador-González, P., Salcedo-Sanz, S. (2016), A novel grouping genetic algorithm–extreme learning machine approach for global solar radiation prediction from numerical weather models inputs. Solar Energy, 132, 129-142.

Baiardi, D. (2023), What do you think about climate change? Journal of Economic Surveys, 37(4), 1255-1313.

Basurto, N., Arroyo, Á., Vega, R., Quintián, H., Calvo-Rolle, J.L., Herrero, Á. (2019), A hybrid intelligent system to forecast solar energy production. Computers and Electrical Engineering, 78, 373-387.

Bebis, G., Georgiopoulos, M. (1994), Feed-forward neural networks. IEEE Potentials, 13(4), 27-31.

Bianchi, F.M., Maiorino, E., Kampffmeyer, M.C., Rizzi, A., Jenssen, R., Jenssen, R. (2017), Other recurrent neural networks models. In: Recurrent Neural Networks for Short-Term Load Forecasting: An Overview and Comparative Analysis. Germany: Springer. p31-39.

Bizzarri, F., Bongiorno, M., Brambilla, A., Gruosso, G., Gajani, G.S. (2012), Model of photovoltaic power plants for performance analysis and production forecast. IEEE Transactions on Sustainable Energy, 4(2), 278-285.

Breiman, L. (2001), Random forests. Machine Learning, 45, 5-32.

Brownlee, J. (2018), Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Vermont: Machine Learning Mastery.

Chen, T., Guestrin, C. (2016), Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining. p785-794.

Cheung, Y.W., Lai, K.S. (1995), Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business and Economic Statistics, 13(3), 277-280.

Cho, K. (2014), On the Properties of Neural Machine Translation: Encoder-decoder Approaches. [arXiv Preprint].

Dambreville, R., Blanc, P., Chanussot, J., Boldo, D. (2014), Very short term forecasting of the global horizontal irradiance using a spatio-temporal autoregressive model. Renewable Energy, 72, 291-300.

Das, U.K., Tey, K.S., Seyedmahmoudian, M., Mekhilef, S., Idris, M.Y.I., Van Deventer, W., Stojcevski, A. (2018), Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews, 81, 912-928.

David, M., Ramahatana, F., Trombe, P.J., Lauret, P. (2016), Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. Solar Energy, 133, 55-72.

de Freitas Viscondi, G., Alves-Souza, S.N. (2019), A systematic literature review on big data for solar photovoltaic electricity generation forecasting. Sustainable Energy Technologies and Assessments, 31, 54-63.

Ding, M., Wang, L., Bi, R. (2011), An ANN-based approach for forecasting the power output of photovoltaic system. Procedia Environmental Sciences, 11, 1308-1315.

Dolara, A., Grimaccia, F., Leva, S., Mussetta, M., Ogliari, E. (2015), A physical hybrid artificial neural network for short term forecasting of PV plant power output. Energies, 8(2), 1138-1153.

Doorga, J.R.S., Dhurmea, K.R., Rughooputh, S., Boojhawon, R. (2019), Forecasting mesoscale distribution of surface solar irradiation using a proposed hybrid approach combining satellite remote sensing and time series models. Renewable and Sustainable Energy Reviews, 104, 69-85.

Erdoğan, S., Çakar, N.D., Ulucak, R., Danish, D., Kassouri, Y. (2021), The role of natural resources abundance and dependence in achieving environmental sustainability: Evidence from resource-based economies. Sustainable Development, 29(1), 143-154.

Erdoğan, S., Gedikli, A., Cevik, E.I., Erdoğan, F. (2022), Eco-friendly technologies, international tourism and carbon emissions: Evidence from the most visited countries. Technological Forecasting and Social Change, 180, 121705.

Gala, Y., Fernández, Á., Díaz, J., Dorronsoro, J.R. (2016), Hybrid machine learning forecasting of solar radiation values. Neurocomputing, 176, 48-59.

Gandelli, A., Grimaccia, F., Leva, S., Mussetta, M., Ogliari, E. (2014), Hybrid model analysis and validation for PV energy production forecasting. In: 2014 International Joint Conference on Neural Networks (IJCNN). United States: IEEE. p1957-1962.

Garnett, R. (2023), Bayesian Optimization. United Kingdom: Cambridge University Press.

Hanna, R., Kleissl, J., Nottrott, A., Ferry, M. (2014), Energy dispatch schedule optimization for demand charge reduction using a photovoltaic-battery storage system with solar forecasting. Solar Energy, 103, 269-287.

Hernández-Torres, D., Bridier, L., David, M., Lauret, P., Ardiale, T. (2015), Technico-economical analysis of a hybrid wave power-air compression storage system. Renewable Energy, 74, 708-717.

Hochreiter, S., Schmidhuber, J. (1997), Long Short-term Memory. Neural Computation. 9(8):1735-80.

Huang, G.B., Zhu, Q.Y., Siew, C.K. (2004), Extreme learning machine: A new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541). Vol. 2. United States: IEEE. p985-990.

Huang, Y., Lu, J., Liu, C., Xu, X., Wang, W., Zhou, X. (2010), Comparative study of power forecasting methods for PV stations. In: 2010 International Conference on Power System Technology. United States: IEEE. p1-6.

Jaeger, H. (2001), The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148(34), 13.

Jiang, H., Dong, Y., Xiao, L. (2017), A multi-stage intelligent approach based on an ensemble of two-way interaction model for forecasting the global horizontal radiation of India. Energy Conversion and Management, 137, 142-154.

Jiménez-Pérez, P.F., Mora-López, L. (2014), Modeling daily profiles of solar global radiation using statistical and data mining techniques. In: Advances in Intelligent Data Analysis XIII: 13th International Symposium, IDA 2014, Leuven, Belgium, October 30-November 1, 2014. Proceedings 13. Germany: Springer International Publishing. p155-166.

Jolliffe, I. (2005), Principal Component Analysis. In: Encyclopedia of Statistics in Behavioral Science. United States: John Wiley and Sons, Ltd.

Kennel, M.B., Brown, R., Abarbanel, H.D. (1992), Determining embedding dimension for phase-space reconstruction using a geometrical construction. Physical Review A, 45(6), 3403.

Keskin, A., Ersin, İ., Atalan, A. (2024), Price estimation of selected grains products based on machine learning for agricultural economic development. In: Türkiye. Journal of Animal and Plant Sciences, 35(5), 0815.

Korstanje, J. (2021), Advanced Forecasting with Python. United States: Apress. p243-251.

Kostylev, V., Pavlovski, A. (2011), Solar power forecasting performance–towards industry standards. In: 1st International Workshop on the Integration of Solar Power into Power Systems, Aarhus, Denmark. Langen, Germany: Energynautics GmbH Mühlstraße. p1-8.

Kühnert, J., Lorenz, E., Heinemann, D. (2013), Satellite-based irradiance and power forecasting for the German energy market. Solar Energy Forecasting and Resource Assessment, 2013, 267-297.

Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P. (2015), A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Solar Energy, 112, 446-457.

Li, J., Ward, J.K., Tong, J., Collins, L., Platt, G. (2016), Machine learning for solar irradiance forecasting of photovoltaic system. Renewable Energy, 90, 542-553.

Li, Y., Sun, Q., Lehman, B., Lu, S., Hamann, H.F., Simmons, J., Black, J. (2016). A machine-learning approach for regional photovoltaic power forecasting. In: 2016 IEEE Power and Energy Society General Meeting (PESGM). United States: IEEE. p1-5.

Li, Z., Rahman, S.M., Vega, R., Dong, B. (2016), A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. Energies, 9(1), 55.

Lou, S., Li, D.H., Lam, J.C., Chan, W.W. (2016), Prediction of diffuse solar irradiance using machine learning and multivariable regression. Applied Energy, 181, 367-374.

Öztürk, I., Aslan, A. (2015), Are fluctuations in electricity consumption permanent or transitory? Evidence from a nonlinear unit root test in high-income OECD countries. Energy Sources, Part B: Economics, Planning, and Policy, 10(3), 257-262.

Ozturk, I., Aslan, A., Altinoz, B. (2022), Investigating the nexus between CO2 emissions, economic growth, energy consumption and pilgrimage tourism in Saudi Arabia. Economic Research-Ekonomska Istraživanja, 35(1), 3083-3098.

Pedro, H.T., Coimbra, C.F. (2012), Assessment of forecasting techniques for solar power production with no exogenous inputs. Solar Energy, 86(7), 2017-2028.

Russo, M., Leotta, G., Pugliatti, P.M., Gigliucci, G. (2014), Genetic programming for photovoltaic plant output forecasting. Solar Energy, 105, 264-273.

Schapire, R.E. (2013), Explaining adaboost. In: Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin, Heidelberg: Springer Berlin Heidelberg. p37-52.

Seabold, S., Perktold, J. (2010), Statsmodels: Econometric and statistical modeling with python. SciPy, 7(1), 92-96.

Shahi, S., Fenton, F.H., Cherry, E.M. (2022), Prediction of chaotic time series using recurrent neural networks and reservoir computing techniques: A comparative study. Machine Learning with Applications, 8, 100300.

Su, Y., Lucey, B.M., Jha, A.K. (2024), Finance Research and the UN Sustainable Development Goals–an analysis and forward look. Research in International Business and Finance, 71, 102463.

Tiba, S., Belaid, F. (2021), Modeling the nexus between sustainable development and renewable energy: The African perspectives. Journal of Economic Surveys, 35(1), 307-329.

Vrettos, E., Gehbauer, C. (2019), A hybrid approach for short-term PV power forecasting in predictive control applications. In: 2019 IEEE Milan PowerTech. United States: IEEE. p1-6.

Watanabe, S. (2023), Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. [arXiv Preprint].

Watanabe, S., Awad, N., Onishi, M., Hutter, F. (2022), Speeding up multi-objective hyperparameter optimization by task similarity-based meta-learning for the tree-structured parzen estimator. [arXiv Preprint].

Watanabe, S., Hutter, F. (2022), c-TPE: Tree-structured Parzen estimator with inequality constraints for expensive hyperparameter optimization. [arXiv Preprint].

Yona, A., Senjyu, T., Funabashi, T., Kim, C.H. (2013), Determination method of insolation prediction with fuzzy and applying neural network for long-term ahead PV power output correction. IEEE Transactions on Sustainable Energy, 4(2), 527-533.