Artificial & Computational Intelligence/Published Online: July 2021
https://acors.org/ijacoi/VOL2_ISSUE2_4.pdf

Artificial & Computational Intelligence

Thermal and Hydro Energy Generation Predictive Analysis

Yash Garg[1], Ramanuj Sharma[2]
[1, 2]Amity School of Engineering and Technology
[1]ygarg704@gmail.com, [2]ramanujs3@gmail.com
[1, 2]Amity University, Noida, Uttar Pradesh, India

Abstract

Enhancing the precision of renewable energy forecasts is critical to power system strategy, maintenance, and activities as renewable energy becomes more widespread in the global energy grid. However, due to the intermittent and chaotic existence of renewable energy data, this is a difficult task. In this paper, we analyzed the thermal and hydro energy generation trends in India and present a novel approach that will predict the amount of thermal (nonrenewable) and hydro (renewable) energy generated every single day for the foreseeable future using machine learning regression algorithms like multiple linear regression, decision tree, random forest, etc. Finally, we go over current research developments, complications, and conceivable forthcoming research directions in this range.

Keywords: Regression, Energy Prediction, Hydro Energy, Thermal Energy, Machine Learning

1. Introduction

In recent years, green sources of energy have come into light and constant change is been brought about in order to deal with the climate change. Consequently, renewable (hydro) energy has gotten a lot of attention all over the world. Renewable energy, such as solar, wind, tidal, hydro, and geothermal energy, is energy that can be recycled in nature. The main usage of these energy sources is Power Generation, Heating and Transportation fuels. Nonrenewable energy on the other hand is extracted from sources that can deplete or not be replenished in our lifetimes, or even lifetimes to come.Renewable energy has at least two benefits over fossil fuels.

To begin with, renewable energy supplies are abundant, renewable, and unlimited in the world. Second, renewable energy is clean, green, and emits fewer pollutants, making it more efficient and environmentally friendly. Renewable energy, in particular, can significantly reduce Sulphur dioxide (SO2), carbon monoxide (CO), and dust emissions, lowering the risk of atmospheric pollution and the greenhouse effect [1]. Constant exploitation of fossil fuel reserves has created various problems for the world and several countries are adopting methods to conserve energy. Depletion of fossil fuel have alarmed the researchers about the need of other renewable energy sources. Easy availability of renewable energy reduces the dependency on fossil fuels and also helps in bringing down carbon emissions in the atmosphere.

Although renewable energy is seen as the most viable alternative to fossil fuels because it is clean, green, and naturally replenished over a wide geographic area, it also brings unscheduled volatility, putting energy systems' efficiency and stability at risk, particularly with large-scale renewable energy integration [2]. On one hand, green energy has a high level of volatility, intermittent news, and randomness, both of which will inevitably raise the reserve capacity of electric energy grids, increasing power generation costs. Because of advances in electricity and control infrastructure and administration, demand load forecasting is now one of

the most modern phenomena. As a result, new artificial intelligence (AI) methods have only recently been available [3].

Researchers have started focusing on forecasting energy because of the impact it has over the renewable energy. Machine learning / Artificial Intelligence is extensively exercised in the power engineering for energy load forecasting. The machine learning solution uses the load from the previous testing sample, and generates an appropriate network system, and trains the system to meet the precision norms using a particular training algorithm.

In this paper, we have predicted and analyzed the thermal and hydro energy generation of every single day in India. Initially, after data cleaning, we analyzed the data to get more insights and then we prepared our regression model. Section 2 of the paper consists the related work that has been already done in this field. Following section 3 describes the technique that has been adopted tin this study and the data that has been taken into consideration for this study. Section 4 and 5 consists the complete methodology along with the method that has been used to prepare the model. Section 6 concludes the study and discusses about the work that can be done in the future in order to enhance this study even further.

2. Literature Review

Wang [13] used deep learning methods to forecast renewable energy. Their paper describes a systematic and in-depth scrutiny of deep learning-based renewable energy forecasting approaches, performance, and execution potential while this paper used unsupervised machine learning methods to estimate renewable and non-renewable energy output.

In paper [7], they analyzed the power consumption trends from renewable energy sources and non-renewable energy sources and combined them. This paper proposes a novel machine learning-based hybrid method for power forecasting that combines multilayer perceptron (MLP), support vector regression (SVR), and CatBoost.

Chen [3] proposed a conventional method of level prediction with a pattern recognition approach was performed by first predicting the actual numerical values using typical pattern-based regression models, hen classifying them into pattern levels (e.g., low, average, and high) while this paper shows regression method to predict thermal and hydro energy without classifying them into any categories. The authors [3] only used one classification algorithm whereas this model uses four regression methods. By estimating the mean monthly wind speed of three wind farms in northwest China, the proposed hybrid approach is tested.

Energy Consumption is essential in order to utilize the resources efficiently and to save the available resources for future use. Stanley [11] have focused their study on real time prediction and short-term forecast using the ANN system to understand the energy generated by PV system, that converts solar energy into electricity without emitting the harmful greenhouse gases and other pollutants, in the form of photovoltaic panels that are installed on a building.

Several other works in ANN have been done to analyze power generation for different renewable resources. In paper [9], three input variables namely wind speed, generation hours and relative humidity has been used to develop a model using the MATLAB toolbox in order to calculate the wind power generation for prediction of energy. Wind Energy is an important aspect of renewable resources and using them to its full potential can help solve the ongoing crisis and prevent us from several future crisis as well.

3. Proposed Technique

In this paper, approach outlined in Fig (1) was followed. Data was collected from publicly accessible National Power Portal reports at https://npp.gov.in/ . After the data was extracted from the web, the data was then prepared for EDA by renaming columns, reindexing of dataset, converting values into float etc. where extensive analysis was done to analyze the data.

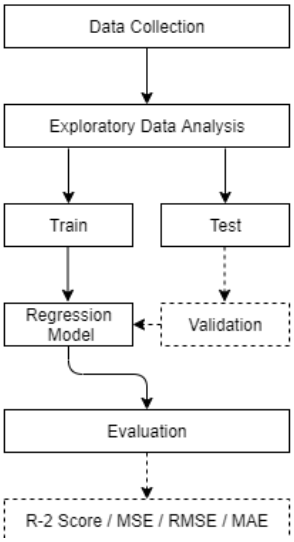

Figure 1. Approach Followed.

The model was validated using the prototype data base. To assess the accuracy of the proposed model, we used a variety of error metrics. After that, we compared the forecasting findings to the real data from the dataset. From 2017 to 2020, regression algorithms were used to forecast the amount of thermal and hydro energy produced on a daily basis.

The information was gathered from India's National Power Portal. The amount of thermal and hydro energy generated per day in India was predicted using the region, day, year, and month. We didn't include nuclear energy because the evidence was inadequate, and our model would have been skewed, so we only used (predicted) thermal and hydro energy.

4. Experimental Data Analysis

The data taken from the National Power Portal consisted of data divided into regions and states and union territories, regions being “Northern”, “North Eastern”, “North Western”, “Southern”, “Central”, “Eastern” and “Western”. The total area covered in each region, as well as their coordinates, were also included.

For the regression model, attributes such as country, year, day, and month were used (prediction). The data was important because we wanted to predict the amount of energy produced on a specific day, which is why we used country, date, year, and month as our key features for prediction.

Fig(2) shows the percent of national share of thermal and hydro energy for every region in our dataset and it can be seen that most of the energy is generated from the northern (27.19%) region of India as presented in fig(3) whereas the smallest contribution is from the north eastern (7.94%) region and the western (15.44%), eastern (12.71%), southern (19.33%) and central (13.48%) region almost produces the same amount of energy.

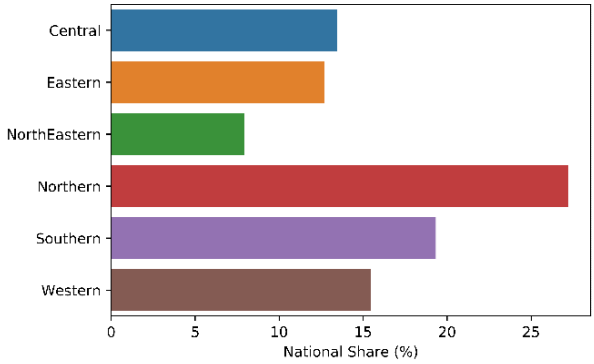

Figure 2. Share of each region.

Fig(3) shows the same data in the form of a bubble map which helps to give a clear look at the national share of each region. From fig(2) and fig(3), it can be concluded that the northern region is responsible for most of the energy production in India.

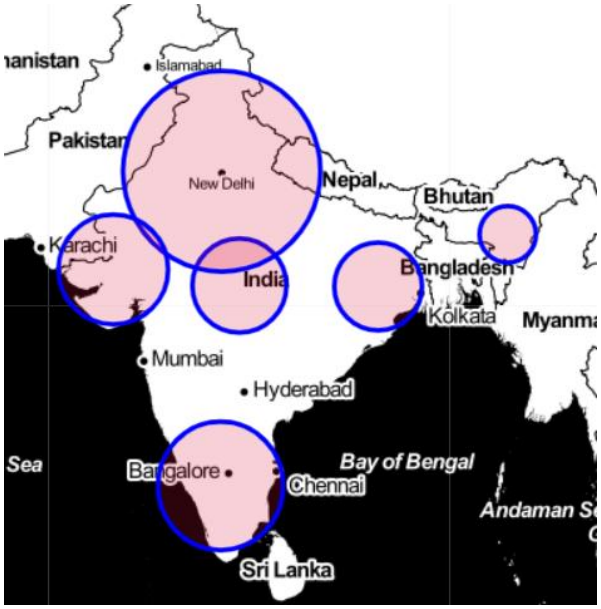

Figure 3. Bubble Map for each region.

Fig(4) shows the national share of each state and union territory of India. Rajasthan (10.55%), which lies in the northern region, covers most of the national share of energy production while the lowest comes from Puducherry (0.003%), which lies in the southern region.

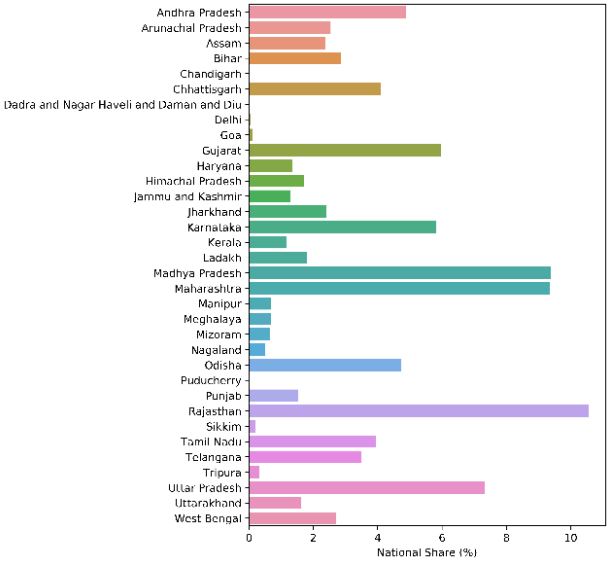

Figure 4. State / UT National Share.

Hydro Energy generated every year, since 2017 to 2020 has dominated the charts. It is two times greater in production than thermal energy as shown in fig(5) and fig(6).

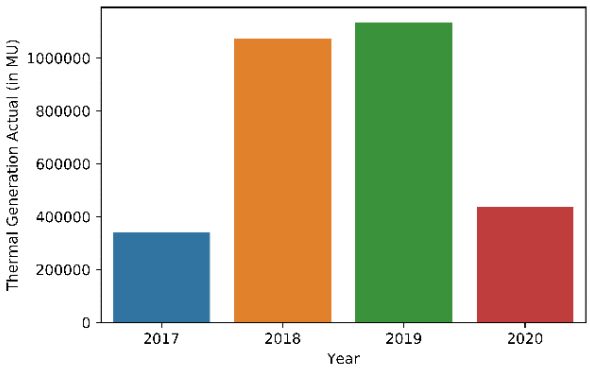

Figure 5. Thermal Energy per year.

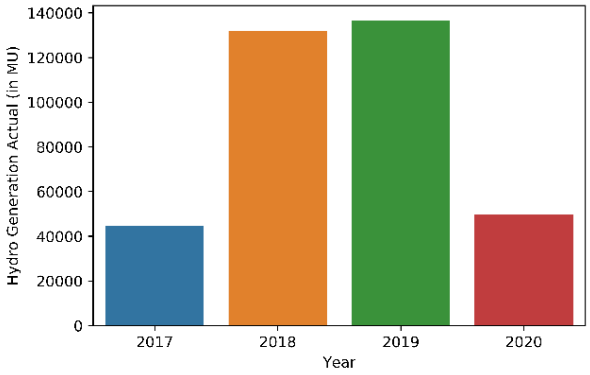

Figure 6. Hydro Energy per year.

Fig(7) and fig(8) shows the total amount (sum) of thermal and hydro energy generated for each twelve months (from 2017 to 2020).

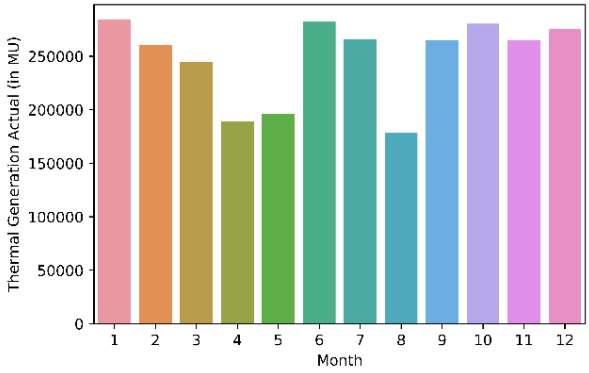

Figure 7. Thermal Energy per month.

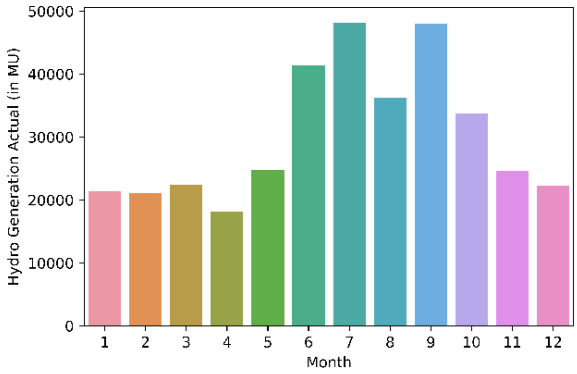

Figure 8. Hydro Energy per month.

This gives us a better look at the thermal and hydro energy generated every month. June has the highest amount of thermal energy production whereas in July and September, most of the hydro energy is generated.

5. Regression

After performing exploratory data analysis on whole dataset, we prepare the dataset for regression. Table (1) summarizes the energy wise data in our dataset. It shows the count, mean, standard deviation, minimum and maximum load of thermal and hydro energy.

	Count	Mean	Std	Min	Max
Thermal	4,945	603.97	383.53	12.34	1395.97
Hydro	4,945	73.30	74.48	0.00	348.72

Table 1. Energy Summary Data

The dataset was split into independent and dependent variables X and y respectively. The “Region” column was transformed using OneHotEncoder [4]. It splits one column into n columns, n being the number of different values it holds. After preprocessing, we then split our dataset using train test split. The test size was 20% and the train size was 80%.

After splitting our dataset, we used different algorithms for forecasting. We used four different

algorithms for our regression model which were Multiple Linear Regression, Decision Tree Regression, Random Forest Regression and finally, XG Boost.

a. Multiple Linear Regression

The equation of the regression algo mentioned above is defined in equation (1).

$$yi = \beta o + \beta 1xi1 + \beta 2xi2 + \cdots + \beta pxip + \in \quad (1)$$

$$Where\ for\ i = n\ operations:$$

$$yi = dependent\ variable$$

$$xi = explanatory\ variable$$

$$\beta o = y - intercept$$

$$\beta p = slope\ coeff.for\ each\ explanatory\ variable$$

$$\in = error\ term$$

In order to find the best-fit line for each independent variable, multiple linear regression considers three factors:

- The regression coefficients that result in the least total model error.
- The final model's t-statistics.
- The p value that goes with it.

It then calculates the *t*-statistic and *p*-value for each regression coefficient in the model [2].

b. Decision Tree Regression

A decision tree is constructed from the top down, starting from a origin node, and involves segmentation of the data into subsets of instances with indistinguishable values (homogenous). The similarity of a numerical sample is premeditated using standard deviation. The standard deviation of a numerical sample that is fully homogeneous is zero.

- For tree building, the standard deviation (S) is used (branching).
- When deciding when to avoid branching, the Coefficient of Deviation (CV) is used.
- We may also use count(n).
- The value in the leaf nodes is called the average (Avg).

c. Random Forest Regression

A Random Forest is an collaborative technique that uses numerous decision trees and a method called Bootstrap and Aggregation, also known as bagging, to implement both regression and classification tasks [6]. Instead of depending on individual decision trees, the simple theory is to merge several decision trees to determine the final production. N_estimators for random forest were 200.

d. XG Boost for Regression

Gradient boosting is an 'ensemble machine learning algorithm' which can be used to solve classification and regression predictive problems. Ensembles are built from decision tree structures. To fix the prediction errors produced by prior models, trees are added to the ensemble one at a time and fitted. The boosting model is a kind of ensemble machine learning model. Models are equipped using a gradient descent optimization algorithm and any arbitrary differentiable loss function.

Gradient boosting gets its name from the fact that the loss gradient is reduced as the model is fitted, much like a neural network [7]. GridSearchCV was used with XGBoost and Random Forest to avoid overfitting or underfitting.

e. Tools Used

NumPy and pandas were used for the classification of the model, cleaning of the dataset and seaborn, matplotlib and folium were used for experimental data analysis. Sklearn module was used for classification and for model scoring. Python language was used for the whole implementation which was done in Jupyter Notebook kernels. Python was used because of high availability of libraries and packages.

6. Results

The dataset consists of 4,945 values of both thermal and hydro energy. For evaluation of the model, R^2 score, mean squared error, root mean squared error and mean absolute error was used. The result summary of all the regression models is shown in Table (2).

	R^2 Score	MSE	RMSE	MAE
Multiple Linear	0.845	1820.97	42.67	28.41
Decision Tree	0.998	71.38	8.44	1.04
Random Forest	0.999	30.44	5.51	1.18
XGBoost	0.998	130.51	11.42	6.98

Table 2. Result Summary

a. R2 Score

The percentage of the variation in the dependent variable that is predictable from the independent variable(s) is determined using the coefficient of determination, abbreviated as R2 or r2 and is shown in equation (2) [8].

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \overline{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \quad (2)$$

b. Mean Absolute Error

The mean absolute error (MAE) is calculated as the dataset's absolute variance mean which represents the difference between the initial and predicted values. Equation (3) is used to measure the MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \quad (3)$$

c. Mean Squared Error

The difference between the initial and estimated values is the mean squared error (MSE). It's calculated by squaring the dataset's mean squared error with Equation (4).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \overline{Y_i}| \quad (4)$$

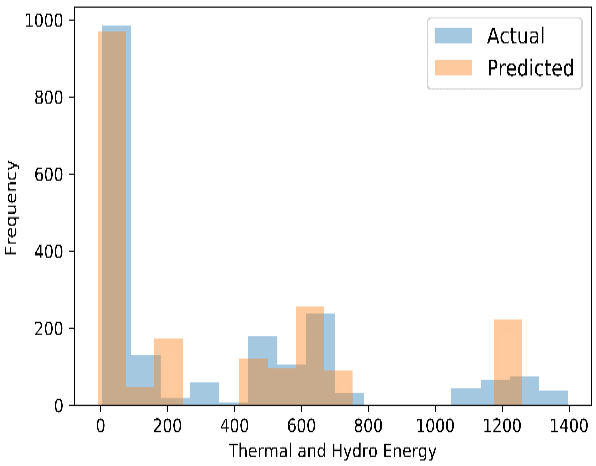

Figure 9. Linear Regression Model.

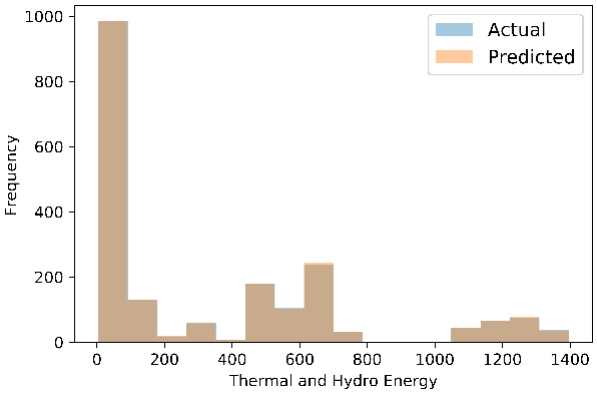

Figure 10. Decision Tree Model.

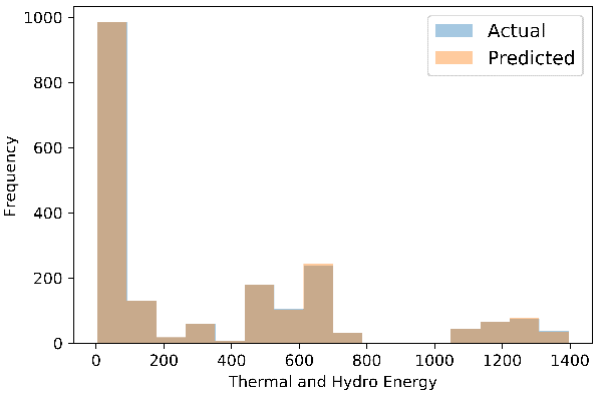

Figure 11. Random Forest Model.

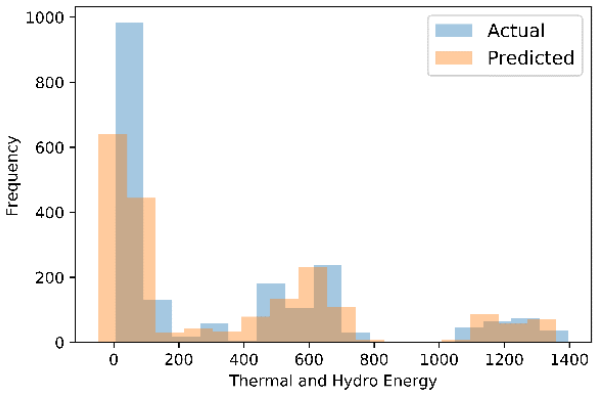

Figure 12. XG Boost Model.

7. Conclusion and Future Work

In this paper, we were able to effort my way through a dataset, create a model and train it to predict the amount of energy generated using four different algorithms.EDA, and train-test-split was done, before training the model. Various metrics were used to test the proposed model which are mean absolute error, mean squared error, root mean squared error, and the coefficient of determination. This model which predicts the energy generated per day will allow better predictions of energy generation.

Future research can look at whether the model would

correctly forecast annual results by including weather conditions (weekly, monthly, and annual), which will reduce the number of specific details used to predict performance [12]. To increase the accuracy of the forecasts, different training data (random) and computational methods (e.g., jack-knife, neural network) should be considered.

References

[1]. Ahmad, Ahmad S., et al. "A review on applications of ANN and SVM for building electrical energy consumption forecasting." Renewable and Sustainable Energy Reviews 33 (2014): 102-109.

[2]. Almeshaiei, Eisa, and Hassan Soltan. "A methodology for electric power load forecasting. " Alexandria Engineering Journal 50.2 (2011): 137-144.

[3]. Chen, Yu-Tung, Eduardo Piedad, and Cheng-Chien Kuo. "Energy consumption load forecasting using a level-based random forest classifier." Symmetry 11.8 (2019): 956.

[4]. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Tiede, D.; Aryal, J. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote. Sens. 2019, 11, 196

[5]. Hu, Jianming, Jianzhou Wang, and Guowei Zeng. "A hybrid forecasting approach applied to wind speed time series." Renewable Energy 60 (2013): 185-194.

[6]. Jain, R.K.; Smith, K.M.; Culligan, P.J.; Taylor, J.E. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl. Energy 2014, 123, 168–178.

[7]. Khan, Prince Waqas, et al. "Machine learning-based approach to predict energy consumption of renewable and nonrenewable power sources." Energies 13.18 (2020): 4870.

[8]. Lee J, Lee I, Kim S. Multi-site photovoltaic power generation forecasts based on deep-learning algorithm. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2017, pp. 1118–20.

[9]. M. Carolin Mabel; E. Fernandez. (June 2007). Analysis of wind power generation and prediction using ANN: A case study. Renewable Energy 33 (2008) 986–992.

[10]. Mendonça de Paiva, G.; Pires Pimentel, S.; Pinheiro Alvarenga, B.; Gonçalves Marra, E.; Mussetta, M.; Leva, S. Multiple Site Intraday Solar Irradiance Forecasting by Machine Learning Algorithms: MGGP and MLP Neural Networks. Energies 2020, 13, 3005.

[11]. Stanley K.H. Chow; Eric W.M. Lee; Danny H.W. Li. (August 2012). Short-term prediction of photovoltaic energy generation by intelligent approach. Energy and Buildings 55 (2012) 660–667.

[12]. Tsekouras, G.J.; Kanellos, F.D.; Mastorakis, N. Short term load forecasting in electric power systems with artificial neural networks. In Computational Problems in Science and Engineering; Springer: Berlin, Germany, 2015; pp. 19–58

[13]. Wang, Huaizhi, et al. "A review of deep learning for renewable energy forecasting." Energy Conversion and Management 198 (2019): 111799.

[14]. Zhang J, Wang Y, Sun M, Zhang N, Kang C. Constructing probabilistic load forecast from multiple point forecasts: a bootstrap based approach. In: 2018 IEEE Innovative Smart Grid Technologies – Asia (ISGT Asia), Singapore, 2018, pp. 184–9.