MACHINE LEARNING METHODS IN FORECASTING SOLAR PHOTOVOLTAIC ENERGY PRODUCTION

by

Marina M. MILIĆEVIĆ* and Budimirka R. MARINOVIĆ

Faculty of Production and Management Trebinje, Trebinje, Bosnia and Herzegovina

Original scientific paper
https://doi.org/10.2298/TSCI230402150M

Energy has an effective role in economic growth and development of societies. This paper is studying the impact of climate factors on performance of solar power plant using machine learning techniques for underlying relationship among factors that impact solar energy production and for forecasting monthly energy production. In this context this work provides two machine learning methods: ANN for forecasting energy production and decision tree useful in understanding the relationships in energy production data.

Both structures have horizontal irradiation, sunlight duration, average monthly air temperature, average maximal air temperature, average minimal air temperature and average monthly wind speed as inputs parameters and the energy production as output. Results have shown that used machine learning models perform effectively, ANN predicted the energy production of the PV power plant with a correlation coefficient higher than 0.97.

The results can help stakeholders in determining energy policy planning in order to overcome uncertainties associated with renewable energy resources.

Key words*: solar energy, ANN, decision tree*

Introduction

Energy issues are a key factor for sustainable development. The solar energy is the largest and most important source of renewable energy [1]. This energy is clean, free and abundant in most places throughout the year [2]. On the other hand, a concern about the unpredictability and reliability of solar energy is one of the most challenging issues [3] in the planning and operation of energy systems.

Solar energy have attracted the greatest attention in many application such as space heating, solar radiation prediction, electricity generation [4] and in recent decades it had big annual growth rate mainly because of technological innovation, improved cost effectiveness and government support. Solar energy and energy generated in PV power plant has been investigated in many literature researches. So, Ramsami and Oree [5] used the hybrid technique with conventional linear regression and ANN models for forecasting the 24 hours energy output of PV system. Dumitru *et al*. [6] also used ANN model and Elman neural network for forecasting energy production. Khatib *et al*. [2] used ANN model to predict the clearness index, and used

*Corresponding autor, e-mail: marina.milicevic@fpm.ues.rs.ba

it for predicting global solar irradiation. Shen *et al*. [7] analyzed variation of global solar irradiance and how it influences the optimal tilt angle of grid connected PV power plants.

Forecasting techniques based on type of inputs can be grouped as physical and statistical approaches [5]. While physical methods models are function of independent variables (cell temperature, cell characteristics, *etc*.), statistical methods analyze historical dataset in forecasting. This methods use techniques such as ANN, regression analysis, time-series analysis, *etc*. In the field of artificial intelligence (AI), machine learning (ML) integrates statistics and computer science to build algorithms that get more efficient when they are subject to relevant data rather than being given specific instructions [7]. The ML methods are best - known to be for achieving desired results in prediction tasks [8]. El Maghraoui *et al*. [8] used ML algorithms for predicting mine energy consumption in order to determine the best model for a variety of users who must construct predictive models. Sharma *et al*. [9] used support vector machine, Gaussian processes approaches, and the adaptive neuro-fuzzy inference system to assess the flexural strength of concrete that include waste marble. In [10], the authors used ANN to predict the velocity of crude oil. Liyew and Melese [11] used ML techniques for daily rainfall forecasting. For measuring the performance they used RMSE and mean absolute error methods. Results showed that the best ML algorithm in this case is gradient boosting ML algorithm.

In this context, this work is investigating the energy forecasting using two ML algorithms: ANN and decision tree (DT). Since climate change affects the yield of renewable energy systems (as well as solar energy) and due to the sensitivity toward environmental parameters, this research is analyzing some climatic factors and their impact on PV power output and uses them for forecasting energy from PV power plant. Both structures (ANN and DT) have six independent variables: horizontal irradiation, sunlight duration, average monthly air temperature, average maximal air temperature, average minimal air temperature, and average monthly wind speed as inputs and the energy production as output. This dataset was collected on a monthly basis during the period of five years (from 2016 to 2020). For the experimental part of the research, MATLAB software was used along with its classification learner app. Classification learner app is application for ML algorithms and contain tools for data visualization, regression, classification, *etc*. The research had the benefit of this application by using it for modeling DT.

Materials and methods

The data set used in the present study use energy production data from power plant located in B&H, situated on the rooftop of one building. PV system consists of 980255 W polycrystalline solar panels. The panels are fixed and inclined at an angle of 30° (the optimal inclination angle by the defined location is 35° but PV modules are built into an existing roof) of the south to the roof.

Meteorological data: average monthly air temperature, average maximal air temperature, average minimal air temperature, and average monthly wind speed were collected from the Republic Hydro-meteorological Institute of the Republic of Srpska. The PV geographical information system provides us information on horizontal irradiation. Table number shows preliminary statistics evaluation of inputs for analyzed location. Table shows the stabile climate during the year on the analyzed location.

Figure 1 shows the details of instrument for measurement of wind speed. Table 1 shows the stabile climate during the year on the analyzed location.

Machine learning models

The main aspiration of AI is to make computers imitate the human brain instead of blindly following instructions made by humans. To accomplish the required task, people use

their own experience and knowledge with the aim of decision-making, and this is the main goal for which AI strives – to equip computers with intelligence.

Table 1. Preliminary statistical evaluation of inputs for analyzed location

	Range			
	Minimum	Maximum	Mean	St.Dev
Horizontal irradiation [kWh/m^2]	39.66	235.28	126.91	61.74
Sunlight duration [hours]	281.44	464.41	371.90	67.72
Average monthly air temperature [°C]	-0.65	24.56	13.07	7.20
Average maximal air temperature [°C]	5.19	33.26	19.73	7.91
Average minimal air temperature [°C]	-5.17	16.55	7.31	6.19
Average monthly wind speed [m/s]	1.14	3.37	2.30	0.55

The widespread use of AI is a consequence of its applications for providing systems with the ability to automatically learn and predict from experience without programming [12]. A common name for all algorithms and applications of this kind, which have rapidly developed within the last years, is ML. The ML algorithms extract patterns from prearranged dataset in order to make prediction for future and unseen data. The ultimate objective of ML is to find an algorithm that most precisely makes a prediction for the output using input from the future, and all this with a function that is trained on previously observed and known data.

Figure 1. Photograph of instrument that is used for measurement of wind speed

The main categorization of ML algorithms is into classes of supervised and unsupervised learning depending on whether the algorithm uses training data that are labeled or not *i.e.* whether the correct output is given or not, while reinforcement ML as the third category uses trial and estimate error to discover the best prediction model.

The most widely used ML algorithms are Linear regression, Naive Bayes, k-Nearest Neighbors (k-NN), support vector machines, ANN, and DT. Although these are all supervised learning algorithms, they all have different mathematical backgrounds. The ANN is statistical model that discover complex relationships between input and output data by simulating processes among human brain cells. The probabilistic approach to ML with the utilization of Bayes theorem gives a Naive Bayes classifier, best known for its application on text document classification [12]. The classification technique named k-NN identifies the nearest neighbors of every point in a dataset with the aim of determining the class of that point [13]. Support vector machines [14] are algorithms for determining the hyperplane in n-dimensional space with the intention of instances on each side of the hyperplane belonging to the same class. As dimension n is the number of features, for the case of three features the hyperplane in R^3 is a plane. Linear regression is an algorithm based on statistical analysis which learns from training dataset predicting the relationship between two variables. For the unsupervised algorithms, the best known one is clustering. A cluster is a group of data with similar features and it is formed based on analysis of surrounding points in the dataset [15].

There is a wide spectrum of areas and applications of AI and ML – from expert systems, learning systems, fuzzy logic, and genetic algorithms through visual perceptions, tactility,

dexterity, locomotion, and navigation to natural languages, speech recognitions and virtual reality. In addition to its wide application, the main advantages of ML are automation, pattern recognition and efficiency in short time. Nevertheless, ML has disadvantages concerning input data accuracy, high error susceptibility and problems with the interpretation of results and limited resources.

By considering the field of forecasting energy production, it is clear that the ANN has been successfully applied in recent years [16, 17] while DT became more and more popular tools [18]. In this study, these two ML models for prediction monthly solar energy production were developed and compared. The neural network models for predicting solar energy are promising and can be used for forecasting solar energy production for any region [6, 19]. For a presented ANN model, it was selected a number of neurons in the hidden layer, while the number of neurons in the input and output layers is fixed (six for the input and one for the output layer). There are a lot of various factors that impact solar energy production – this paper is considering six of them. For such, multidimensional problems, ANN networks represent the best solution due to their applicability in finding non-linear dependencies in higher dimensions and cardinality. On the other hand, DT proved to be very useful in grasping the underlying relationships among the factors that impact solar energy production and for selecting the most significant factor.

In order to attain a better understanding of the proposed models, the next two sections are covering the detailed insight into DT and neural network algorithms, including DT construction and topology of neural networks.

Decision tree

The DT is one of the most common supervised ML techniques used for building the model to classify data into predefined class labels. For its simplicity and accuracy, it is widely used for prediction output based on the available data. Research has shown that DT is much better in prediction category data than forecasting numerical values [19]. Its application does not necessitate significant calculation understanding, but its main disadvantage is that it can lead to significant deviations between predictions and actual results [20]. It has a form of a tree-like flowchart in which nodes of the tree are the input parameters. Tree branches are values of the parameters. The leaves of the tree represent the output parameters depending on the input parameters. The root node, which is on the top of the tree, represents the sample (one data row of independent variable). Splitting is a process of dividing node into other nodes with a lower position in the tree according to some rule. Pruning is reverse process which eliminates some nodes from tree in order to prevent overfitting.

The first prediction model DT related the six independent variables to a single dependent variable. The model divided the output domain into three classes (under, average, and above) which classify produced energy depending on whether produced energy was less, greater, or approximately equal to average production. Figure 2 utilizes a scatter plot to illustrate dependencies among horizontal

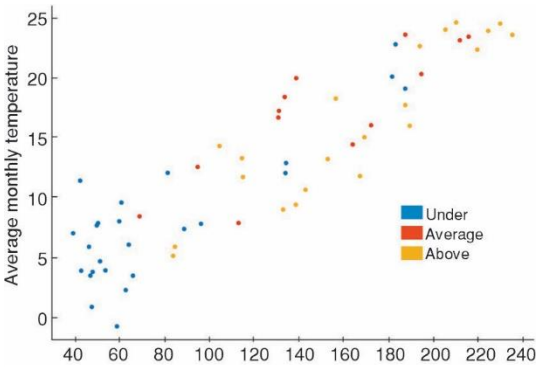

Figure 2. Scatter plot between the energy production and horizontal irradiation

irradiations and average monthly air temperature during the research period in which different dot colors indicate one of three classes of energy production.

Confusion matrix was used to describe the performance of a DT classification model. The order of confusion matrix is $m \times m$ where m is the number of classes, and each column and row shows the number of instances labeled as another class. Diagonal matrix elements represent the correctly classified number of instances.

As a performance evaluation metrics and measures in a DT the following is commonly used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Error Rate} = 1 - \text{Accuracy} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{True Positive Rate (Recall)} = \frac{TP}{TP + FN}; \text{False Negative Rate} = \frac{FN}{FN + TP} \tag{4}$$

where TP stands for true positives count, TN stands for true negative count, FP for false positive count, and FN for false negative count.

Artificial neural network

The ANN is a simplified mathematical representation of the biological neural network. They learn from examples, recognize a pattern in the data, adapt solutions over time, and process information rapidly [20]. They do not require any prior knowledge between inputs and output; they learn relationship between inputs and output [21]. The ANN consist of three layers: input, output and hidden layer. The number of neurons in input and output layer is equal to the number of input/output parameters while the number of hidden layers and number of neurons in hidden layers can be adjusted to finally get an optimum structure [22].

In this research ANN are used to forecast the monthly PV production. A single hidden layer feed forward neural network with Levenberg-Marquardt algorithm is developed. The best results were obtained from the ANN model using the sigmoid activation function in the hidden layer and linear activation function in output layer.

Before the training process, all the data have to be normalized. Various normalization methods are used for ANN to increase the reliability of the trained network [23]. In this research the normalization is applied as:

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5}$$

where $x_{\max}$, $x_{\min}$, and x_n represent the maximum, minimum, and scaled values of the x data sample. Figure 3 shows the relationship between the energy production and sunlight duration and energy production and average monthly air temperature.

In order to estimate PV energy production the subject of the analysis was ANN with a different number of neurons in hidden layer from 1 to 35. Parameters RMSE and R to attain minimum error were evaluated to assess the fitness of the implemented ANN. Figure 4 shows that the minimum RMSE reached was 0.02 and this corresponding to the network with 11

neurons in hidden layer. The RMSE at any other network structure was higher. This neural network also had the best regression value $R = 0.99$.

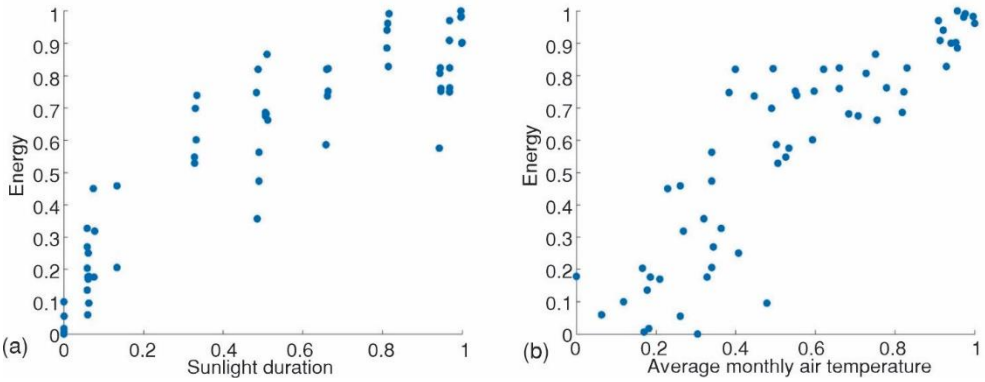

Figure 3. Scatter plot between the energy production and (a) Sunlight duration and (b) average monthly air temperature

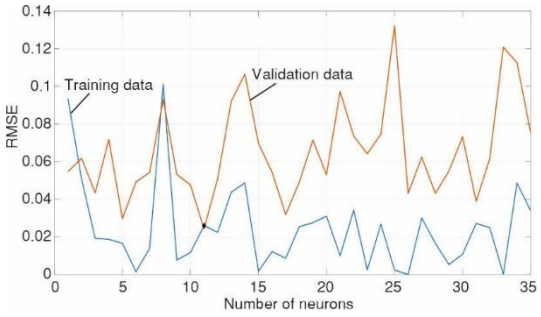

Figure 4. Number of neurons *vs.* RMSE

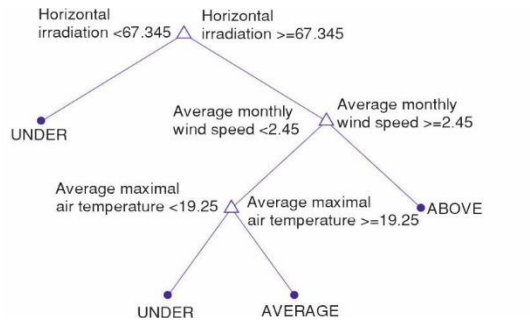

Figure 5. Pruned DT for classification energy production

Results and discussion

Figure 5 shows a pruned DT on energy production. As depicted in the fig. 4, the most significant independent variable for forecasting energy production is horizontal irradiation, which gives us the first split. The classification algorithm estimated that the most important independent variables are horizontal irradiation, average monthly wind speed and average maximal air temperature. The nodes that represent other independent variables from dataset are pruned.

The DT model for forecasting energy production gained accuracy of 66.7%, with the error rate 0.334, MSE 0.4834 and RMSE 0.6952. It is allowed maximal 100 splits with Gini index for splitting the data.

The confusion matrix illustrated on fig. 6 reflects the correctly classified instances and the misclassification of the energy production. It can conclude that:

– From the total of 25 instances with the value of UNDER, the DT has correctly classified 19 (76%) instances, 5 (20%) misclassified as AVERAGE and 1 (4%) as ABOVE.
– From the total of 13 instances with the value of AVERAGE, the DT has correctly classified 8 (62%) instances, 2 (15%) misclassified as UNDER and 3 (23%) as ABOVE.
– From the total of 22 instances with the value of ABOVE, the DT has correctly classified 13 (59%) instances, 2 (9%) misclassified as UNDER and 7 (32%) as AVERAGE.

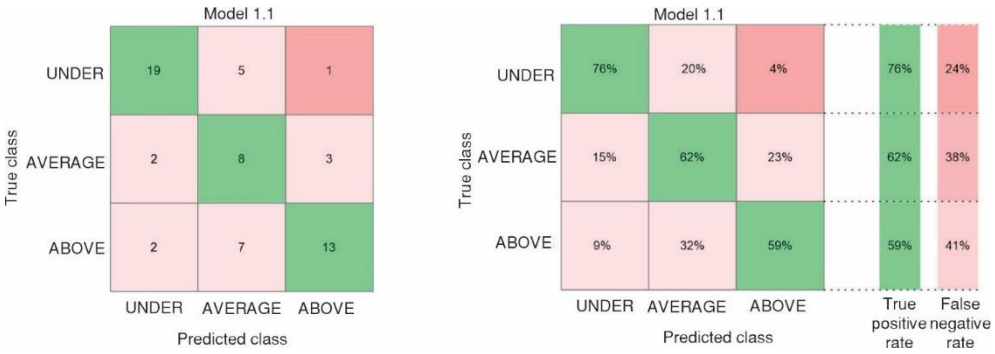

Figure 6. Confusion matrix for DT

The graphic in fig. 7 shows the calculated produced energy *vs.* actual energy of ANN model. It is clear that the observed data almost entirely overlap with the predicted data for training dataset.

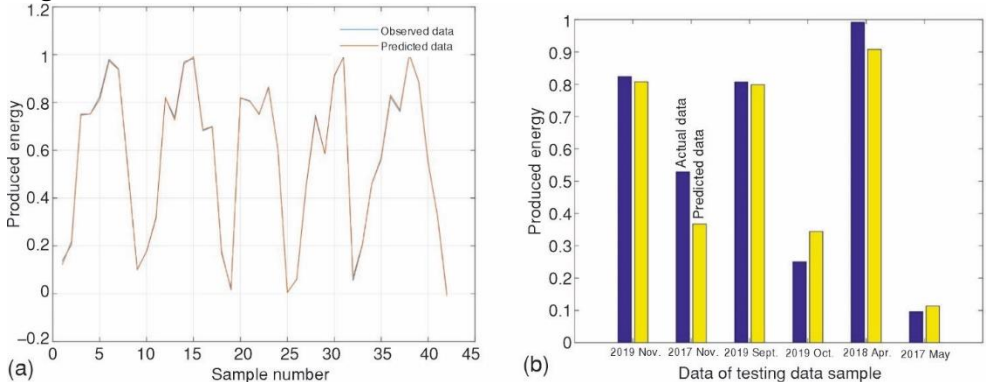

Figure 7. (a) Predictive *vs.* observed Energy – training data and (b) predicted *vs.* observed energy – test data

In order to evaluate the grade of the prediction of ML algorithms for prediction of monthly energy production, a set of evaluation criteria can be used. The correlation coefficient, *R*, and RMSE are statistics used for the ANN. For the quantification of the model performance for DT it is commonly used MSE, accuracy, precision and the recall of the model.

The method of MSE is valuable for making comparison among models and measures the sum of errors:

$$MSE = \frac{1}{n}\sum_{1}^{n} d_i - y_j^{\ 2} \tag{6}$$

The RMSE is measuring the efficiency of model in predicting future individual values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n} d_i - y_j^{\ 2}} \tag{7}$$

Correlation coefficient, *R*, is measure that indicates the association degree between variables [24]:

$$R = \frac{n\sum d_i y_j - \sum y_j \ \sum d_i}{\sqrt{n\ \sum d_i^{\,2} - \sum d_i^{\,2}}\ \sqrt{n\ \sum y_i^{\,2} - \sum y_i^{\,2}}} \tag{8}$$

In eqs. (6)-(8) d_i and y_j are actual and predicted output values, respectively, and n represents the number of samples. Models with smaller MSE and RMSE indicate the high quality of the algorithm. In general, the best value for R is 1 which indicates the high performance of model [4].

Table 2 shows complete measure of performance of the DT by the three classes, as well as weighted metrics. The rate which gives a measure of how often wrongly classify some class into another one is the weighted true positive rate of the model (recall) and its value for the model is 0.68. In addition, a weighted false negative rate is calculated and its value is 0.33. The precision of the model (by class and weighted) carries the information of how often is correct in positive prediction – weighted prediction for this classifier is 0.72.

Table 2. Detailed accuracy of the DT **(by class and weighted)**

Model\measure	True positive rate (Recall)	False negative rate	Precision	Accuracy of the classifier	Class
Decision tree	0.76	0.24	0.83		Under
	0.62	0.38	0.40		Average
	0.59	0.41	0.76		Above
	0.68	0.33	0.72	0.667	WEIGHTED

Table 3 shows the results of the ANN model for the training, testing and validation phases. Regarding the correlation, R, model approximates energy with a higher accuracy.

Table 3. Performance results for ANN model

Model	Measure	Datasets		
		Testing	Validation	Training
Neural Network (6-11-1)	R	0.9647	0.9773	0.9998
	RMSE	0.1032	0.0612	0.0063

Conclusions

There is an increasing interest in estimating energy production, especially in forecasting renewable energy production such as solar energy. Since ML methods have a high accuracy, short computing time, they do not need an experiment to figure out the input/output relation - they are widely used in renewable energy planning applications.

In this research, application of two ML methods, ANN, and DC for modeling solar PV energy production have been used. Both models have six variables as inputs and one as output. The developed DT model is categorical - six independent variables are referred to as predictor variables, and the target variable is under, average, and above classes of produced energy. Compared to other ML algorithms (as well as the ANN presented in this study), the DT classifier is more efficient and it requires less training time. Nevertheless, it is less accurate than a ANN. Its error rate is 0.333 and MSE is 0.4834 which is far higher than MSE in the ANN. The ANN model is based on the feed forward neural network with a single hidden layer feed forward neural network using Levenberg-Marquardt algorithm and with 11 neurons in hidden layer.

The results indicate that the both ML methods have the ability to predict the monthly energy production with great accuracy and short computational time, but ANN has better results. These results can be used for effective preliminary energy planning and algorithms like this can play a significant role concerning energy planning methods.

Future scope of the work

This work investigates the influence of some climatic factors on the energy production in PV power plants. The PV industry is developing rapidly and the share of energy produced in PV power plants in the total energy production is becoming more and more significant year by year. With AI and ML tools constantly developing, new possibilities for forecasting solar energy became available.

The investigation carried out in this paper could be extended in more than one direction. To avoid overfitting which is characteristic of DT and generally gain better accuracy, a random decision forest could be incorporated into solar energy prediction. For reference, see study [25].

The available studies also suggested that future research could examined some hybrid systems which incorporates the features of both a DT and ANN. In this case, DT will be used for the selection of the most relevant factors for accurate prediction of solar energy production, afterwhile ANN will make a prediction using as input parameters DC results. One similar ML models combination could be found in [26].

The model presented in this research could be improved if, apart from analyzed climatic factors, in the future, includes some other factors like pressure, humidity, wind direction, etc. Besides, the parameters, topologies, and architecture of ANN could be varied, and the accuracy metrics analyzed and compared afterward.

Nomenclature

AI – artificial intelligence
DT – decision tree
MSE – mean square error
RMSE – root mean square error
R – correlation coefficient

References

[1] Batić, I. M., *et al*., Impact of Air Temperature and Wind Speed on the Efficiency of a Photovoltaic Power Plant: An Experimental Analysis, *Thermal Science*, *27* (2023), 1A, pp. 299-310
[2] Khatib, T., *et al*., Solar Energy Prediction for Malaysia Using Artificial Neural Networks, *International Journal of Photoenergy*, *2012* (2012), ID419504
[3] Sedai, A., *et al*., Performance Analysis of Statistical, Machine Learning and Deep Learning Models in Long-Term Forecasting of Solar Power Production, *Forecasting*, *5* (2023), 1, pp. 256-284
[4] Elsheikh, A. H., *et al*., Modeling of Solar Energy Systems Using Artificial Neural Network: A Comprehensive Review, *Solar Energy*, *190* (2019), Mar., pp. 622-639
[5] Ramsami P., Oree V., A Hybrid Method for Forecasting the Energy Output of Photovoltaic Systems, *Energy Conversion and Management*, *95* (2015), May, pp. 406-413
[6] Dumitru, C.-D., *et al*., Solar Photovoltaic Energy Production Forecast Using Neural Networks, *Procedia Technology*, *22* (2016), 1, pp 808-815
[7] Shen, Y., *et al*., Impact of Solar Radiation Variation on the Optimal Tilted Angle for Fixed Grid-Connected PV Array - Case Study in Beijing, *Global Energy Interconnection*, *1* (2018), 4, pp. 460-466
[8] El Maghraoui, *et al*., Smart Energy Management: A Comparative Study of Energy Consumption Forecasting Algorithms for an Experimental Open-Pit Mine, *Energies*, *15* (2022), 13, 4659
[9] Sharma, N., *et al*., Assessing Waste Marble Powder Impact on Concrete Flexural Strength Using Gaussian Process, SVM, and ANFIS, *Processes*, *10* (2022), 12, 2745
[10] Makinde, A., *et al*., Prediction of Crude Oil Viscosity Using Feed-Forward Backpropagation Neural Network (FFBPNN), *Petroleum and Coal*, *54* (2012), 2, pp. 120-131
[11] Liyew, M., Melese, A., Machine Learning Techniques to Predict Daily Rainfall Amount, *Journal of Big Data*, *8* (2021), 153
[12] Mitchell, R., *et al*., *An Artifcial Intelligence Approach*, Springer, New York, USA, 2013
[13] Cunningham, P., Delany, S. J., k-Nearest Neighbor Classifiers, *Multiple Classifier Systems*, *34* (2013), 8, pp. 1-17

[14] Suthaharan, S., *Support Vector Machine Learning Models and Algorithms for Big Data Classification*, Springer, New York, USA, 2016, pp. 207-235
[15] Kassambara, A., *Practical guide to cluster analysis in R: Unsupervised machine learning*, vol. 1, Sthda, http://www.sthda.com, 2017
[16] Femila Roseline, J., *et al*., Neural Network Modelling for Prediction of Energy in Hybrid Renewable Energy Systems, *Energy Reports, 8* (2022), Suppl. 8, pp. 999-1008
[17] Fadare, D. A., Modelling of Solar Energy Potential in Nigeria Using an Artificial Neural Network Model, *Applied Energy, 86* (2009), 9, pp. 1410-1422
[18] Ahmad, M. R.,, Yacine, J. R., Predictive Modelling for Solar Thermal Energy Systems: A Comparison of Support Vector Regression, Random Forest, Extra Trees and Regression Trees, *Journal of Cleaner Production*, 203 (2018), Dec., pp. 810-821
[19] Sedai, A., *et al*., Performance Analysis of Statistical, Machine Learning and Deep Learning Models in Long-Term Forecasting of Solar Power Production, *Forecasting, 5* (2023), 1, pp. 256-284
[20] Cobaner, M., *et al*., Prediction of Hydropower Energy Using ANN for the Feasibility of Hydropower Plant Installation to an Existing Irrigation Dam, Water Resources Management, *An International Journal, 22* (2008), June, pp. 757-774
[21] Hamzacebi, C., *et al*., Forecasting of Turkey's Monthly Electricity Demand by Seasonal Artificial Neural Network, *Neural Compute & Applic, 31* (2019), Aug., pp. 2217-2231
[22] Adil, M., *et al*., Effect of Number of Neurons and a Layers in an Artificial Neural Network for Generalized Concrete Mix Design, *Neural Computing and Appllications*, 34 (2020), Sept., pp. 8355-8363
[23] Jayalakshmi, T., Santhakumaran, A., Statistical Normalization and Back Propagation for Classifcation, *International Journal of Computer Theory and Engineering, 3* (2011), 1, pp. 1793-8201
[24] Lopes, M. N. G., *et al*., Artificial Neural Networks Approaches for Predicting the Potential for Hydropower Generation: a Case Study for Amazon Region, *Journal of Intelligents and Fuzzy Systems, 36* (2019), 6, pp. 5757-5772
[25] Da, L., Kun, S., Random Forest Solar Power Forecast Based on Classification Optimization, *Energy, 187* (2019), 115940
[26] Tsai, C. C., *et al*., Decision Tree-Based Classifier Combined with Neural-Based Predictor for Water-Stage Forecasts in a River Basin During Typhoons: A Case Study in Taiwan, *Environmental Engeenering Scence, 29* (2012), 2, pp. 108-116

Paper submitted: April 2, 2023
Paper revised: May 23, 2023
Paper accepted: May 31, 2023