ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20200425151543
Original scientific papers

Determining of Solar Power by Using Machine Learning Methods in a Specified Region

A. Burak GUHER, Sakir TASDEMIR, Bulent YANIKTEPE*

Abstract: In this study, it is aimed to estimate the solar power according to the hourly meteorological data of the specified location measured between 2002 and 2006 by using different Machine Learning (ML) algorithms. Data Mining Processes (DMP) were used to select the most appropriate input variables from these measured data. Data groups created using DMP were evaluated according to three different ML algorithms such as Artificial Neural Network (ANN), Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). It can be concluded that DMP-ML based prediction models are more successful than models developed using all available data. The most successful model developed among these models estimated the hourly solar power potential with an accuracy of 97%. Also, different error measurement statistics were used to evaluate ML algorithms. According to Symmetric Mean Absolute Percentage Error, 6.12%, 7.22% and 12.72% values were found in the most successful prediction models developed using ANN, KNN and SVR, respectively. In addition, from the meteorological data used in this study the most effective data on solar power as a result of DMP were shown to be Temperature and Hourly Sunshine Duration.

Keywords: data mining processes; machine learning; optimal data analysis; solar power

1 INTRODUCTION

The need of energy, which has become an essential part of rising population, production activities, communication and technological developments, has been rapidly increasing worldwide. The limited availability of carbon-based fuels and their negative effects on the environment lead countries to different and cheap energy sources. Renewable energy sources such as wind, hydro, solar, geothermal, biomass, tidal energy are one of the most important sources. Solar energy came at the head of the renewable energy sources that countries most invest in because of its being unlimited and environmentally friendly [1-3].

The basic condition for obtaining correct and high energy is knowing the solar radiation values of the relevant region before establishing a solar based system. Solar power is directly associated with the solar radiation affecting a region. The greater the amount of radiation, the greater the increase of solar power produced depending on the radiation. For this reason, the solar radiation potential of the target point must first be examined before a solar investment is made [4]. Radiation is measured with the help of devices such as Pyranometer or Pyrheliometer. These devices are sensitive and require expertise and costs are high. Hence, it is a complicated process to install and operate measuring stations everywhere. Different estimation methods have been formed due to the difficulty of the measurement process [5]. The newest and most common of them are ML algorithms. ML is known as a subfield of artificial intelligence (AI). They are often applied in classification and prediction applications. In current years, ML has become very prevalent in the energy field and especially in predicting solar power compared to other methods as performance [6].

It is very important to select and arrange the data suitable for the purpose when making predictions with ML. There are many types of data that can affect solar power directly and indirectly. ML is substantially affected by the type, structure, length, and size of the data because of its operating logic. For this reason, by putting the raw data into a number of pretreatments such as normalization, scaling, and extraction, happens a positive effect on the predicted performance [7, 8].

If there are too many variables in the data pool, there may be a problem in associating the model with the target output, when using ML algorithms. For this reason, every new variable that exists and was entered into the system has positive or negative effects on the prediction. When working with such sensitive applications, applying all the variables in the dataset may not perpetually present high levels of success as anticipated. Therefore, it is crucial to determine the most effective of the model performances from the available data. This operation provides a significant contribution to the precision, interpretation and evaluation of the model. It also reduces the duration of modelling studies [9].

When the studies on solar energy are considered, it can be deduced that the data-prediction relationship is applied in two different ways. In the first application, in order to find the most successful prediction model it was tried to be determined by selecting the features in the data set by different technique and methods and classifying them in various ways [10-12]. In this way, the most optimal data affecting the solar radiation were identified, and the estimated performance accuracy was evaluated with minimum error and variables. In the second application, it was tried to determine the most successful among ML-based algorithms by applying all of the data features [13-16]. In such a study, model performance becomes more significant than data selection, but, because all variables are used, the intended model precision may not be achieved.

Data mining (DM) is known as the process of discovering valid and useful information from large datasets. Estimation of solar power using DM instead of ML is not a very popular method because of low performance [17]. Still, in recent years, the use of DM has started to be widespread in order to improve the performance of the prediction models, ascertain the most appropriate input data and shorten the prediction time. In literature searches for solar-based, the use of DMP was found in prediction models using ML algorithms, although their number remains little.

Ghofrani et al. [18], have suggested a new DM-supported model increasing the clustering estimate accuracy of solar power. In this model, the best cluster was determined with the data of temperature, extraterrestrial radiation, and wind speed and direction with the TS-Clustering method, and solar radiation was successfully clustered with Recurrent Neural Networks (RNN). In another study intended to estimate the amount of solar power of Bangladesh, optimal properties influencing solar radiation the most from 12 selected data sets were determined by utilizing the Waikato Environment for Knowledge Analysis (WEKA) DM tool. Abedin et al. estimated solar radiation using the ML-based ANN model with five inputs [19].

With the aid of the WEKA program, Yadav et al. [20] defined the geographic and meteorological parameters that have the most significant impact on solar power. Three different ML-based ANN models have been developed. Seven input data in the first ANN model, five input data in the second ANN model, and a four-input model in the third ANN model were created. The best performance was achieved in the second ANN model. In another study [21], the Rapid Miner DM program was applied to select the most efficient on solar radiation out of nine different input parameters to estimate the amount of solar power of the Indian region. The most useful input data were determined as temperature, minimum temperature, maximum temperature, altitude, and sunshine duration. With these five input data, a successful prediction model has been created by using ANN, one of the ML algorithms.

Meenal and Selvakumarin [22], aimed to estimate the daily average monthly global solar radiation in their study. Using the WEKA program, they have discovered that the month, latitude, maximum temperature and sunshine duration have the most significant effect on solar power and depend on the humidity input data of the lowest effect. Then, the most successful prediction models were developed by applying ML algorithms such as Support Vector Machines (SVM) and ANN with the selected input data.

The main purpose of this study is to predict the hourly solar energy potential of the Mersin province with DMP supported ML algorithms. Three different ML algorithms such as ANN, SVR and KNN have been used in this comparative study due to their high prediction accuracy according to the prediction studies made with classical methods. DMP, which was not used or very rarely used in previous studies is included in the study in order to reduce the prediction accuracy and time. Also, this study is focused on more than one area. The primary focus is to reduce the dimensionality of the input data used in ML-based models and to select the most effective input data on solar power. For this purpose, as a result of solar-based data of the target location arranged with DMP, three different data subsets are formed with different feature selection functions. Another focus of the study is to reduce the error of the developed models, increase their accuracy and shorten the solution time. Therefore, three different data groups and all available input data are divided into training and testing data sets to be used with ML-based algorithms and a large number of prediction models are compared both among themselves and with different algorithms. At the end of this process, in total 252 prediction models from ML algorithms-based were developed and most successful prediction models were obtained. Another aim of this study is to determine the input data that have the least impact on solar energy from the available data. In this way, the model results can be interpreted meaningfully. As a result, the effect of DMP was observed in all ML-based prediction models developed in the present study and the statistical calculations indicate that the results in this study are promising compared to other similar studies on the specific region.

Matlab R2017b program was employed in the process of transferring, combining, calculating and modelling with ML algorithms. The detailed flow chart of the study to be done is presented in Fig. 1.

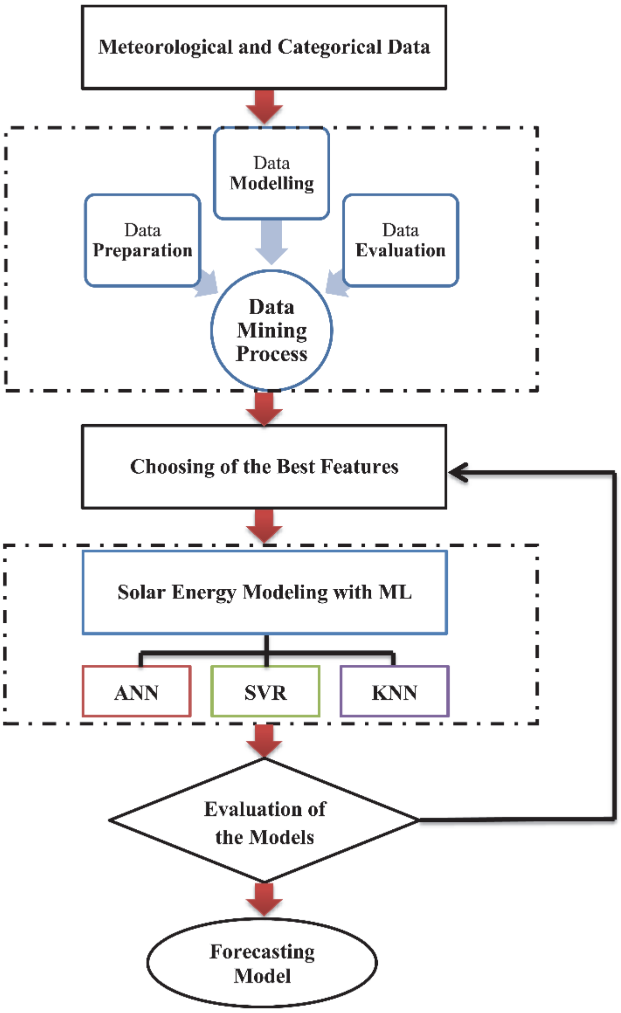

Figure 1 The flowchart of the proposed method

2 MATERIALS AND METHODS

2.1 Location and Database

The meteorological and categorical data of Mersin province in Turkey, in the south of the Mediterranean region, was used in the study. This city is located in the east of the Mediterranean region, as presented in Fig. 2. Mersin is located between 36-37° north latitudes and 33-35° east longitudes. The land border of the province is 608 km, the sea border is 321 km, and its area is 15 853 km^2. The average annual temperature is 18.7° C. The region has a high solar energy potential throughout the year. It is the most significant reason for choosing this location.

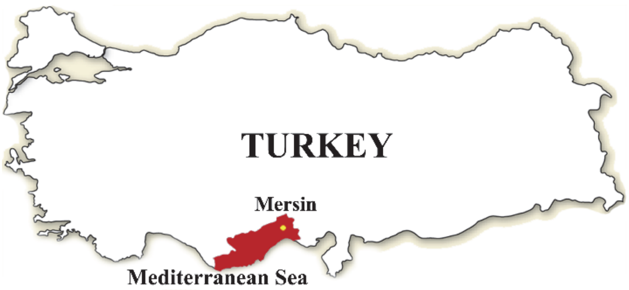

Figure 2 The location of the selected target province in Turkey

For Mersin province, between the years 2002-2006, the 5-year data of hourly meteorological has been requested from the General Directorate of Meteorology of Turkey. The measuring station is located at a height of 36 7808 latitudes, 34 6031 longitude, and 7 meters above sea level. Six different meteorological data like hourly Pressure, hPa; Hourly Sunshine Duration, 0-1; Hourly Average Humidity, %; Hourly Temperature, °C; Hourly Wind Speed, m/s and Hourly Solar Radiation, W/m^2 were obtained from this station. Besides, four different categorical data, such as the year of measurement, the month of the year of measurement, the day of the month of measurement, and the hour of the day of measurement, were involved in the study. Geographical parameters are not involved since they are provincial-based and not changed. As observed in Tab. 1, ten different variables were applied in the study. Based on solar radiation measurement data, data measured between 6-17 hours for January and February and 7-18 hours for March-December were included. In this way, a total of about 4686 data per year were applied in the study.

Some feature engineering processes such as selection, cleaning, transformation, normalization, scaling were applied to raw data before building prediction models with ML. Min-Max normalization was applied in this study. The mathematical Eq. (1) of the Min-Max normalization is also given. In this formula, X_n represents normalized data, X raw data, X_max and X_min the minimum and maximum values in training and testing data. In order for ML algorithms to function appropriately, data properties must have specific requirements and be regulated. The feature engineering operations come into play in this area. Feature engineering mostly serves two purposes. The first is to choose the correct input data properties suitable for ML algorithms, and the second is to increase the performance of ML models to higher levels [23].

$$X_\mathrm{n} = \frac{x - x_\mathrm{min}}{x_\mathrm{max} - x_\mathrm{min}} \tag{1}$$

In this study, approximately 75% of the hourly data has been arranged as training and 25% as testing. This adjustment was made as hourly in day-based. In most ML-based studies, the data are adjusted year-based as daily or monthly. In our studies it has been resolved with existing data that year-based adjustment is not homogeneous since it negatively influences the predict performance. For example, the estimation performance obtained by setting the data between 2002 and 2005 as training, and the data of 2006 as data testing, and the estimation performance obtained by setting the data between 2003-2006 as training, and the data of 2002 as testing data is different. In addition, this method is not very reliable since the model estimation result will be affected when the intervals of training and testing years change. For this reason, the day/hourly distribution was performed in order to make the distribution homogeneous. With a MATLAB code, the daily hourly data were set some as training and some as testing. When Tab. 1 is examined, 17 584 of the five year data set were established as training, and 5858 as testing data in the study and applied a total of 23 442 hourly data.

Table 1 The data of Mersin province used in the study

	Input Data	Output Data	Time Range	Training Dataset	Testing Dataset
Categorical	Year	Solar Radiation	2002 2006	17 584	5858
	Month				
	Day				
	Hour				
Meteorological	Temperature				
	Pressure				
	Mean Humidity				
	Wind Speed				
	H.S. Duration				

2.2 Selection of the Most Optimal Input Data with DMP

DMP covers the processes known as analyzing, reaching, interpretation, and predicting information. It is also known as the extraction of useful information from different data patterns. DM and ML methods use the same algorithms. While DM focuses on the discovery of earlier unknown features, ML algorithms focus on predicting based on known features [24].

There is a positive relationship between ML and the data applied. For this reason, the selection of optimal data clarifies the modelling step. It prevents the extreme compatibility of the algorithm and enables the development of fast and competent models. The desired performance is obtained much more quickly. Contrarily, it may be necessary to develop a much more sophisticated model to reach a high-performance level. This takes more time and labour [23].

Open source WEKA 3.8.2 DMP program was applied to determine the most appropriate input data. This program was developed by the Waikato University applying the JAVA programming language. WEKA has an advanced feature selection module compared to other programs. There are three feature selection methods like correlation-based, wrapper, and filter approaches [25].

The Wrapper approach utilizes the classification method. Hence, there is a strong relationship between data characteristics and classifiers. Accordingly, the wrapper approaches have better feature selection performance than filter approaches [26]. Three different feature selection functions based on the wrapper approach are applied in WEKA. These are Classifier Subset Evaluator (CSE), Wrapper Subset Evaluator (WSE), and Classifier Attribute Evaluator (CAE). The flowchart of the applied feature selection functions is shown in Fig. 3. The selected function is started with (Y^i) initial feature subset. Then, Y^i is evaluated (Z_best) with (T) predefined ML algorithm. The subsets of training data (Y_gen) are created and appraised (Z). The previous and next subsets are evaluated based on T algorithm and their performance is compared ($Z > Z_\mathrm{best}$). The search is repeated until the condition for the stopping criterion (*Q*) is met. Eventually, the best feature subset (Y^i_{best}) for the wrapper-based feature selection function is determined.

CSE, WSE and CAE functions were used to determine the most effective input data on solar radiation. The features and total numbers selected at the end of the DMP process are shown in Tab. 2. In the table, the most effective classifiers on solar radiation are M5 Model Tree (M5P) and Gaussian Processes for Logistic Regression (Elastic Net). Search method is a structural approach in which possible subsets of features are searched. Generally, two methods are used, such as random and comprehensive search. This structure is selected depending on the type of feature selection function. Two different search methods, such as Greedy Stepwise and Ranker, were applied in the study. At the end of the process, three different data sets such as CSE, WSE and RSE were determined to be used in ML-based models.

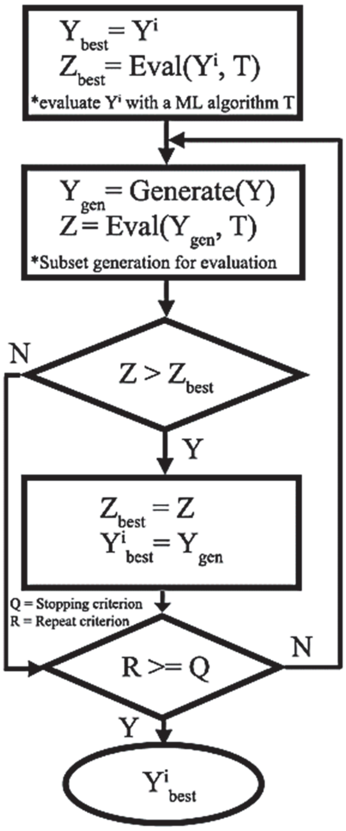

Figure 3 Flowchart of Wrapper-based feature selection functions applied in WEKA

2.3 ML Forecasting Algorithms

Three different ML algorithms, such as ANN, SVR, and KNN, were applied in the study to estimate solar power and to ascertain the most effective dataset at the end of the DMP. Among these algorithms, ANN is the most generally applied ML algorithm in solar radiation prediction studies and it has a very high-performance. It is very successful in resolving nonlinear problems since computers are different from traditional calculation methods when calculating. Because of its flexibility, it works hybrid with different ML algorithms [7]. In the study, Multi-Layer Feedforward Backpropagation (MLFFBB) based neural network was practiced. The architecture of the neural network is shown in Fig. 4.

Table 2 Optimal features obtained at the end of the DMP process

Feature Selection Function	Classifier	Search Method	Selected Inputs	Number of Best Attribute
CAE	M5Rules	Ranker	Month, Hour, Pressure, Temperature, Humidity, Wind Speed, Hourly S.Duration	7
CSE	M5Rules	Greddy Stepwise	Year, Month, Day, Hour, Temperature, Hourly S.Duration	6
WSE	Elastic Net	Greddy Stepwise	Year, Month, Hour, Temperature, Humidity, Wind Speed, Hourly S.Duration	7

A mathematical formula of an MLFFBB neural network is given in Eq. (2). The optimal output is calculated until error rate between output and target has been minimal, according to the data set available and the architectural structure of the network. In this formula, the activation function, weights, input data and bias are *f*, *X*, *W* and *b*, respectively.

$$\text{Output}_k = f\left(\sum_{k=1}^{N}\left(W_{ik}X_k + b_k\right)\right) \tag{2}$$

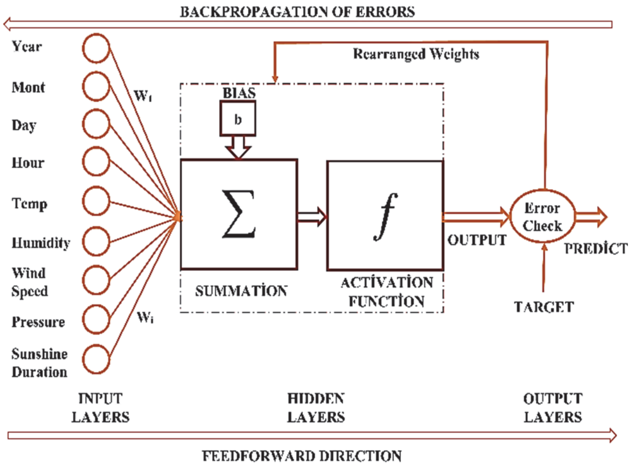

Figure 4 The architecture of the MLFFBB neural network used in the study

Another ML algorithm used in the study is SVR. Support Vector Machines (SVM) have been developed in a structure commonly established for classification found by Vapnik [27]. It is pretty good at solving nonlinear problems. SVR has been developed to solve the regression-based problems of SVM. There are five core functions commonly applied in SVR. These functions are Linear Kernel, Polynomial Kernel, Normalized Polynomial Kernel, Sigmoid Kernel, and Gaussian RBF (Radial Basis Function) Kernel [28]. Polynomial, Normalized Polynomial, and RBF Kernel were employed in the study. The mathematical formulas of the used kernels are given in Eqs. (3), (4) and (5), respectively.

$$k(X_a, X) = [(X_a, X) + 1]^d \tag{3}$$

$$k\left(X_a, X\right) = \frac{\left[\left(X_a X\right) + 1\right]^d}{\sqrt{\left(X_a\right)^2 \left(X\right)^2}} \tag{4}$$

$$k\left(X_a, X\right)) = \exp\left(-\frac{1}{2\sigma^2} \| X_a - X \|^2\right) \tag{5}$$

KNN is an ML algorithm known as a non-parametric classification and regression algorithm. It is preferred because it is a simple but efficient algorithm. The purpose of this algorithm is to make predictions by using the sample data set in a dataset with certain classes. In the KNN study, a *K* value is determined according to the number of data to be analyzed. *K* is randomly determined constant and this constant directly influences the forecast performance. For each new data, the nearest *K* data is taken, and its distance to the new data is calculated. Generally, three types of functions are applied in the distance calculation. These are Euclidean, Manhattan, and Minkowski functions [28]. After calculating the distance, an estimate is performed according to the incoming value. Euclidean distance function was applied in this study and its Eq. (6) is also given. In the formula, x_i, y_i, *K* represent the input data, the output found for each input data, and the neighborhood rate, respectively. *K* constants were determined for values between 1 and 10. Decreases in performance above 10 were observed for the *K* constant.

$$E = \sqrt{\sum_{i=1}^{K} \left(x_i - y_i\right)^2} \tag{6}$$

2.4 Statistical Comparison of ML Algorithms Models

Six different statistical indicators were applied to assess the performance of ML algorithms used in the study in order to assess the solar power. Mean Square Error (*MSE*), Root Mean Square Error (*RMSE*), Mean Absolute Error (*MAE*) and Symmetric Mean Absolute Percentage Error (*SMAPE*) are standard error measurement statistics. Statistical analysis methods such as Correlation Coefficient (*R*) and Coefficient of Determination (R^2) were applied to evaluate the performance of ML models. Formulas of statistical indicators are presented in Eqs. (7) to (12). O_i, P_i, $\overline{O}$ and $\overline{P}$ in the formulas represent, respectively, measured, predicted, measurement and averages of predictions.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(O_i - P_i\right)^2 \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(O_i - P_i\right)^2} \tag{8}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | O_i - P_i | \tag{9}$$

$$SMAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{| O_i - P_i |}{\left(| O_i | + | P_i |\right) \cdot 0.5} \tag{10}$$

$$R = \frac{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)\left(P_i - \overline{P}\right)}{\sqrt{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2} \sqrt{\sum_{i=1}^{n} \left(P_i - \overline{P}\right)^2}} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(O_i - P_i\right)^2}{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2} \tag{12}$$

3 RESULTS AND DISCUSSIONS

In order to predict the solar power of 5-year (2002-2006) hours in specified location, the data to be used in the study were initially determined and then requested from the General Directorate of Meteorology of Turkey. A total of 10 data were used in this study. Six of these data are meteorological (Pressure, Hourly Sunshine Duration, Average Humidity, Temperature, Wind Speed and Solar Radiation), and 4 of them are categorical (Year, Month, Day, Hour) data. Since solar radiation is the target feature, the time intervals that the radiation during the day are determined, and the data patterns in this range are transferred to the study. In this way, a total of 23 442 hourly data between 2002 and 2006 were included in the study.

Three different ML algorithms, such as ANN, SVR, and KNN, were applied to estimate the amount of solar power. In order to increase the estimation accuracy and to find data patterns powerful on data set, DMP, which is rarely used in this field, has been applied. Firstly, the data was adjusted by applying DMP and then arranged to achieve high-level performance from the selected ML algorithms. At the end of the DMP, three different data groups based on CAE, CSE, and WSE feature selection functions were created. A total of 7 features with CAE, 6 with CSE, and 7 with WSE were decided to be the most useful data on solar radiation. It is adjusted hourly to make the distribution homogeneous. For example, on the 1st day of January, 6 am, 7 am, 8 am, 10 am, 11 am, 12 pm, 14 pm, 15 pm, and 16 pm are devoted to training, and 9 am, 13 pm and 17 pm to testing.

Table 3 Architectural structure and training performance of the neural network used in models

Data Groups	Input	Hidden	Trans. Func.	Train Func.	Train MSE	Test MSE
CAE	7	45	Tansig	Trainlm	0.0025	0.0028
CSE	6	47			0.0022	0.0024
WSE	7	30			0.0026	0.0027
ALL	9	31			0.0022	0.0024

In this study, four different ANN models were developed. While modelling, the MATLAB R2017b program was utilized. For estimation, a program code was written in MATLAB, and ANN models were trained by creating one hidden layer neural network models with 1-50 neurons. According to the input data selected for each hidden layer neuron, the highest training and testing results were found with 5 different trials. Tab. 3, shows the architectural structures and training performance results of four different ANNs. The most optimal data at the end of the DMP process were taken as input data to the ANN models. The first model was trained between 0-1000 epochs with 7 inputs and 45 hidden neurons, the second model with 6 inputs and 47 hidden neurons, the third model with 7 inputs and 30 hidden neurons. The last model was trained using all available input data with 9 inputs and 31 hidden neurons. In Tab. 3, MSE performances are seen at the end of the training period of the models. CSE feature selection data and ALL Data MSE performances are similar for training and testing with values of 0.0022 and 0.0024, respectively. In ANN modelling, selection of transfer function has been found to have a positive or negative effect on performance. Following the data structure in the study, performance losses are quite low when using logistic sigmoid (logsig) or hyperbolic tangent sigmoid (tansig) in the hidden layer and hyperbolic tangent sigmoid (tansig) or linear (purelin) in the output layer. Since its performance is better than other transfer functions, tansig was chosen in both layers. The best performance was obtained in the training function known as Trainlm (Levenberg-Marquardt Backpropagation).

In Tab. 4, the models are shown to estimate solar radiation by applying test data that they have not tryed before. Then, the results measured by the test results were compared by different statistical methods. In this way, the success performance of each model was evaluated. SMAPE was employed in the study to evaluate the models at a statistical scale. Basically, MAPE is a more favoured method in similar studies. However, if there are zero or very close to zero values in the data set, it produces insignificant results. If the *SMAPE* value is lower than 10%, it can be deduced that the model is predicted on the best scale and between 10% and 20% on good scale [7].

Table 4 The statistical comparison of most successful ANN models

Training			Testing			
Data Groups	*RMSE*	*R*	*RMSE*	*MAE*	*SMAPE* / %	R^2
CAE	0.0518	0.9823	0.0535	0.0375	7.57	0.9627
CSE	**0.0471**	**0.9848**	**0.0484**	**0.0335**	**6.12**	**0.9694**
WSE	0.0515	0.9825	0.0525	0.0369	7.44	0.9641
ALL	0.0476	0.9850	0.0489	0.0337	7.01	0.9689

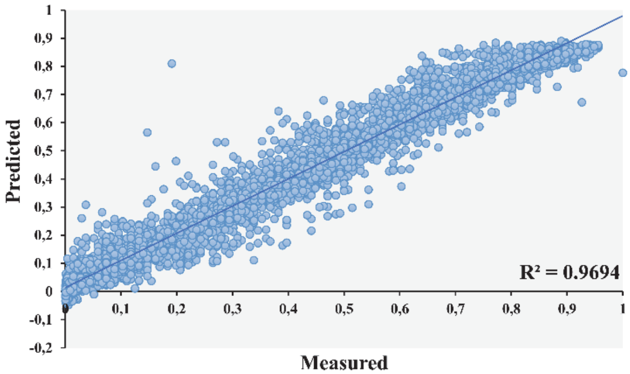

Figure 5 Regression graph of ANN model according to CSE data group with 6 inputs and 1 output

In Tab. 4, the ANN model with 6 inputs CSE datagroup with *SMAPE* value of 6.12%, *MAE* value of 0.0335, *RMSE* value of 0.0484 and a R^2 value of 0.9694 successfully estimated the solar radiation of specified location. The model can be used successfully in an hourly solar power prediction for the Mersin province. The prediction performance graph of the model is presented in Fig. 5.

Another ML algorithm used to evaluate the solar energy of Mersin is KNN. *K* constant was taken between 1 and 10, and 10 different KNN models were developed for each feature group. The Euclidean kernel was used in all models as the kernel. For CEA, CSE, WSE, and ALL data groups, it was tried to determine the most successful one among 40 different KNN models. In Tab. 5, evaluation results of the developed models are shown. From this table, *SMPAE* and R^2 values for $K = 3$ of the most successful model were found for CSE data groups with 7.22% and 0.9469, respectively.

Table 5 The statistical comparison of most successful KNN models

Testing					
Data Groups	*K* Constant	*RMSE*	*MAE*	*SMAPE* / %	R^2
CAE Data	10	0.0639	0.0465	7.51	0.9468
CSE Data	**3**	**0.0639**	**0.0446**	**7.22**	**0.9469**
WSE Data	6	0.0679	0.0496	7.75	0.9399
ALL Data	4	0.0749	0.0551	8.47	0.9273

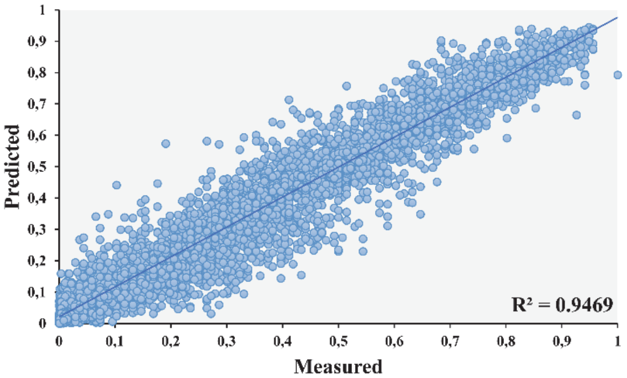

Figure 6 Regression graph of KNN model according to CSE data group with 6 inputs and 1 output

A total of 12 different applications were developed with SVR by applying three different kernel Polynomial, Normalized Polynomial, and RBF. In Tab. 6, the effects of the models developed in the selected feature groups are observed comparatively. In this table, the model applied to the CSE data group developed by using Normalized Polynomial kernel is the most successful and the highest with $R^2 = 0.8088$ according to other models. The regression graphs of the most successful models developed using KNN and SVR ML algorithms are shown in Fig. 6 and Fig. 7.

Table 6 The statistical comparison of most successful SVR models

Testing					
Data Groups	Karnel	*RMSE*	*MAE*	*SMAPE* / %	R^2
CAE	RBF	0.1362	0.1073	12.75	0.7582
CSE	**N. Polly**	**0.1223**	**0.0950**	**12.72**	**0.8088**
WSE	RBF	0.1364	0.1075	12.83	0.7574
ALL	N.Polly	0.1242	0.0968	12.80	0.8019

All ML algorithms used in the study were compared with each other on the basis of the most successful solar energy estimate model. In addition, the impact of DMP on output performance was evaluated. As a result of all these operations, it can be concluded that DMP worked compatible with ML-based algorithms.

A comparative graph of ML algorithms with each other was presented in Fig. 8. The data set used in this graph is randomly selected data for July 2006. Hourly measured and predicted values are compared in these graphs. The test data set contains data corresponding to the data obtained at 7 am, 8 am, 9 am, 10 am, 11 am, 12 pm, 13 pm, 14 pm, 15 pm, 16 pm, 17 pm, 18 pm per day on an hourly basis. At least 3 data were compared with each other approximately for each day and a total of 101 hourly data were used.

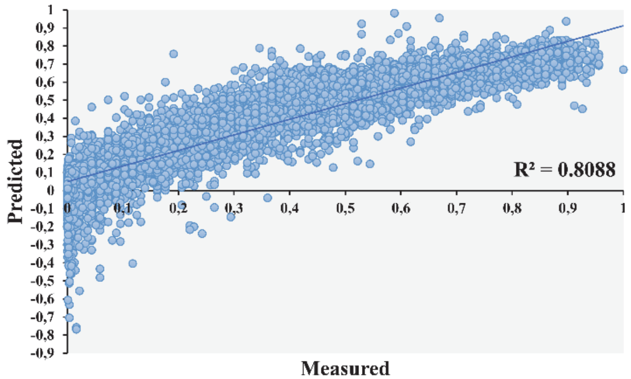

Figure 7 Regression graph of SVR model according to CSE data group with 6 inputs and 1 output

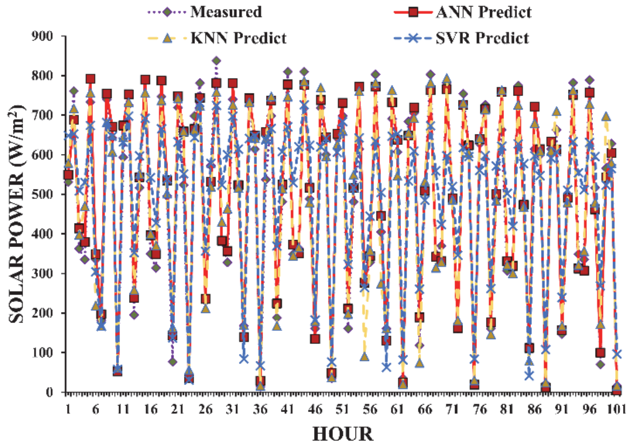

Figure 8 Comparison of predicted and measured most successful of ML algorithm models for July 2006

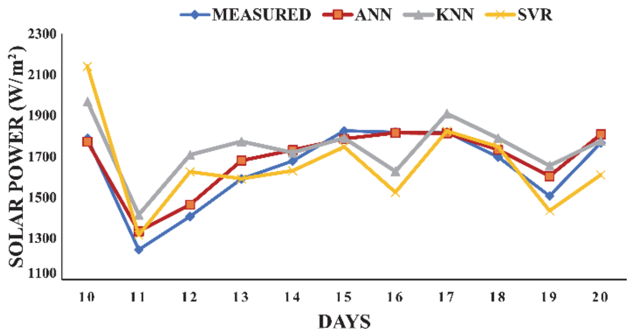

Figure 9 Evaluation of the models performance of the daily average hourly data of May 2004

ANN, KNN and SVR estimation models are compared by calculating the daily average values of 10-day hourly data selected randomly in May 2004 as shown in Fig. 9. In the graph, the sum of hourly data set is taken randomly for the relevant day. Regression graph of most successful ML based method between estimated and measured data is shown using the daily average data of May in Fig. 10. Also, the estimation performance and regression results are shown ANN-based according to daily average data of July 2002 in Fig. 11. According to the graph, the daily average hourly data came out with a result similar to the most accomplished model performance with 97%.

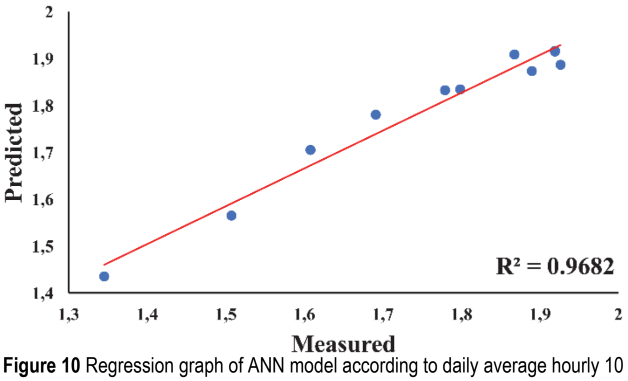

Figure 10 Regression graph of ANN model according to daily average hourly 10 days data in May 2004

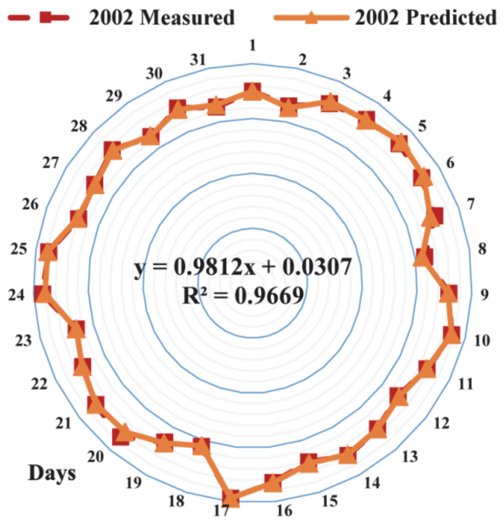

Figure 11 ANN performance evaluation according to daily average hourly data of July 2002

Table 7 Comparison of the current study with previous studies according to specified location

Studies	Working Area	ML Algorithm	Statistical Indicator
Koca [29]	Mersin	ANN	*RMSE*: 0.0690
Kaba [30]		Deep Learning Neural Network	*MAE*: 0.45 *RMSE*: 0.58
Kisi [31]		Fuzzy Genetic (FG), ANN, ANFIS	*MAE*: 1.72 *RMSE*: 2.54
Present Study		**DMP-ML based (ANN, SVR, KNN)**	***MAE*: 0.0335** ***RMSE*: 0.0484**

The literature searches based on the specified location were compared with similar studies in Tab. 7. When the present study is compared with previous studies modelled with single or multiple ML algorithms specific to Mersin Province, the best prediction performance with *MAE* = 0.0335, *RMSE* = 0.0484 was obtained by using ANN prediction model. The obtained results are promising in performance appraisal according to other studies.

4 CONCLUSIONS

In this study, it was proposed to comparatively predict the solar energy potential of specified location by an effective method. In line with this goal, ML algorithms that perform quite well in solar-based prediction studies such as ANN, KNN and SVR, were used. In addition, a DMP supported method was applied to increase the precision of prediction models developed with ML algorithms, to interpret the effect of input data on output and to shorten the modelling time. It has been observed that Sunshine Duration and Temperature input data are the most influencing parameters on solar power in all prediction models developed. Besides, it is understood that Pressure, Humidity and Wind Speed input data do not have a significant effect on the performance of the prediction models. The estimation models of the most successful ANN, KNN, and SVR algorithms, according to the R^2 statistics scale, have 0.9694, 0.9469, and 0.8088 values, respectively. According to the SMAPE scale, the same algorithms were calculated as 6.12%, 7.22%, and 12.72%, respectively. ANN demonstrated superior prediction performance compared to the other two ML algorithms according to the stated results and successfully predicted the solar radiation of the Mersin province.

DMP improved the prediction performance and it was provided to find out the most effective data on solar radiation. It was seen that the most successful feature selection data set in all three groups, CAE, CSE, and WSE, was CSE. However, it has been indicated that using ALL data group with different ML algorithms does not have a positive effect on prediction performance. The proposed study is distinguished from the current studies in this aspect and it adds a new dimension to ML-based models on the prediction of solar power. Moreover, two different applications of the data-prediction relationship in the literature are merged in one study and the performances of DMP-ML based prediction models are evaluated.

5 REFERENCES

[1] Yaniktepe, B., Kara, O., & Ozalp, C. (2017). The global solar radiation estimation and analysis of solar energy: Case study for Osmaniye, Turkey. *International Journal of Green Energy, 14*(9), 765-773. https://doi.org/10.1080/15435075.2017.1329148

[2] Şahin, M. (2019). Determining Optimum Tilt Angles of Photovoltaic Panels by Using Artificial Neural Networks in Turkey. *Technical Gazette, 26*(3), 596-602. https://doi.org/10.17559/TV-20160702220418

[3] Skrúcaný, T., Kendra, M., Stopka, O., Milojević, S., Figlus, T., & Csiszár, C. (2019). Impact of the electric mobility implementation on the greenhouse gases production in Central European countries. *Sustainability, 11*(18), 4948. https://doi.org/10.3390/su11184948

[4] Ayodele, T. R., Ogunjuyigbe, A. S. O., & Chukwuka, G. M. (2016). On the global solar radiation prediction methods. *Journal of Renewable and Sustainable Energy, 8(*2), 023702. https://doi.org/10.1063/1.4944968

[5] Khatib, T., Mohamed, A., & Sopian, K. (2012). A review of solar energy modeling techniques. *Renewable and Sustainable Energy Reviews, 16*(5), 2864-2869. https://doi.org/10.1016/j.rser.2012.01.064

[6] Mosavi, A., Salimi, M., Faizollahzadeh Ardabili, S., Rabczuk, T., Shamshirband, S., & Varkonyi-Koczy, A. R. (2019). State of the art of machine learning models in energy systems, a systematic review. *Energies, 12*(7), 1301. https://doi.org/10.3390/en12071301

[7] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one, 13*(3). https://doi.org/10.1371/journal.pone.0194889

[8] Tasdemir, S., Yaniktepe, B., & Guher, A. B. (2018). The effect on the wind power performance of different normalization methods by using multilayer feed-forward backpropagation neural network. *International Journal of Energy Applications and Technologies, 5*(3), 131-139. https://doi.org/10.31593/ijeat.464210

[9] Obando, E. D., Carvajal, S. X., & Agudelo, J. P. (2019). Solar Radiation Prediction Using Machine Learning Techniques: A Review. *IEEE Latin America Transactions, 17*(04), 684-697. https://doi.org/10.1109/TLA.2019.8891934

[10] Chiteka, K. & Enweremadu, C. C. (2016). Prediction of global horizontal solar irradiance in Zimbabwe using artificial neural networks. *Journal of Cleaner Production, 135,* 701-711. https://doi.org/10.1016/j.jclepro.2016.06.128

[11] Çelik, Ö., Teke, A., & Yıldırım, H. B. (2016). The optimized artificial neural network model with Levenberg-Marquardt algorithm for global solar radiation estimation in Eastern Mediterranean Region of Turkey. *Journal of cleaner production, 116*, 1-12. https://doi.org/10.1016/j.jclepro.2015.12.082

[12] Almaraashi, M. (2018). Investigating the impact of feature selection on the prediction of solar radiation in different locations in Saudi Arabia. *Applied Soft Computing, 66*, 250-263. https://doi.org/10.1016/j.asoc.2018.02.029

[13] Hassan, M. Z., Ali, M. E. K., Ali, A. S., & Kumar, J. (2017, December). Forecasting day-ahead solar radiation using machine learning approach. *4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 252-258. https://doi.org/10.1109/APWConCSE.2017.00050

[14] Feng, Y., Gong, D., Zhang, Q., Jiang, S., Zhao, L., & Cui, N. (2019). Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. *Energy Conversion and Management, 198*, 111780. https://doi.org/10.1016/j.enconman.2019.111780

[15] Khosravi, A., Koury, R. N. N., Machado, L., & Pabon, J. J. G. (2018). Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. *Journal of Cleaner Production, 176*, 63-75. https://doi.org/10.1016/j.jclepro.2017.12.065

[16] Yamamoto, Y., Tsuruta, S., Muranushi, T., Muranushi, Y. H., Kobashi, S., Mizuno, Y., & Knauf, R. (2015, November). Improvement of Sun Flare Prediction by SVM Integrated GA. *11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 719-724. https://doi.org/10.1109/SITIS.2015.37

[17] Mayilvahanan, M. & Sabitha, M. (2013, January). Estimating the availability of sunshine using data mining techniques. *International Conference on Computer Communication and Informatics*, 1-4. https://doi.org/10.1109/ICCCI.2013.6466298

[18] Ghofrani, M., Niromand, N., Azimi, R., & Ghayekhloo, M. (2017, September). A novel data mining method for high accuracy solar radiation forecasting. *North American Power Symposium (NAPS)*, 1-6. https://doi.org/10.1109/NAPS.2017.8107259

[19] Abedin, Z., Barua, M., Paul, S., Akther, S., Chowdhury, R., & Chowdhury, M. S. U. (2017, February). A model for prediction of monthly solar radiation of different meteorological locations of Bangladesh using artificial neural network data mining tool. *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 692-697. https://doi.org/10.1109/ECACE.2017.7912993

[20] Yadav, A. K., Malik, H., & Chandel, S. S. (2014). Selection of most relevant input parameters using WEKA for artificial

neural network based solar radiation prediction models. *Renewable and Sustainable Energy Reviews, 31*, 509-519. https://doi.org/10.1016/j.rser.2013.12.008

[21] Yadav, A. K., Malik, H., & Chandel, S. S. (2015). Application of rapid miner in ANN based prediction of solar radiation for assessment of solar energy resource potential of 76 sites in Northwestern India. *Renewable and Sustainable Energy Reviews, 52*, 1093-1106. https://doi.org/10.1016/j.rser.2015.07.156

[22] Meenal, R. & Selvakumar, A. I. (2018). Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renewable Energy, 121*, 324-343. https://doi.org/10.1016/j.renene.2017.12.005

[23] Zheng, A. & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly Media, Inc.

[24] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

[25] Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). *Weka: Practical machine learning tools and techniques with Java implementations.* Computer Science Working Papers.

[26] Kumar, V. & Minz, S. (2014). Feature selection: a literature review. *Smart CR, 4*(3), 211-229. https://doi.org/10.6029/smartcr.2014.03.007

[27] Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

[28] Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy, 105*, 569-582. https://doi.org/10.1016/j.renene.2016.12.095

[29] Koca, A., Oztop, H. F., Varol, Y., & Koca, G. O. (2011). Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey. *Expert Systems with Applications, 38*(7), 8756-8762. https://doi.org/10.1016/j.eswa.2011.01.085

[30] Kaba, K., Sarıgül, M., Avcı, M., & Kandırmaz, H. M. (2018). Estimation of daily global solar radiation using deep learning model. *Energy, 162*, 126-135. https://doi.org/10.1016/j.energy.2018.07.202

[31] Kisi, O. (2014). Modeling solar radiation of Mediterranean region in Turkey by using fuzzy genetic approach. *Energy, 64*, 429-436. https://doi.org/10.1016/j.energy.2013.10.009

Contact information:

A. Burak GUHER, Instructor
OsmaniyeKorkut Ata University,
Osmaniye Vocational School, Karacaoglan Campus,
80000, Osmaniye, Turkey
E-mail: burakguher@osmaniye.edu.tr

Sakir TASDEMIR, Professor
Selcuk University,
Technology Faculty, Computer Engineering, Alaaddin Keykubat Campus,
Selcuklu, 42075, Konya, Turkey
E-mail: stasdemir@selcuk.edu.tr

Bulent YANIKTEPE, Associate Professor
(Corresponding author)
Osmaniye Korkut Ata University,
Engineering Faculty, Energy Systems Eng. Dept., Karacaoglan Campus,
80000, Osmaniye, Turkey
E-mail: byaniktepe@osmaniye.edu.tr