Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 4 (No. 1), pp. 53-61, 2024
doi: 10.54569/aair.1535217
Published online: Aug 30, 2024

Artificial Neural Network Parameter Optimization: Improving Meteorological Data Predictions through Machine Learning

Ceyhun Kapucu [1,*], Oğuz Akpolat [2]

[1] Mugla Sitki Kocman University, Department of Informatics, Mugla, Turkey
[2] Mugla Sitki Kocman University, Science Faculty, Department of Chemistry, Mugla, Turkey

Abstract

This study aims to create an artificial neural network (ANN) based model to predict solar irradiance using open-sourced meteorological data. A neural network that is feed-forward with backpropagation was employed to build the model. A large combination of model parameters including learning algorithms, transfer functions, number of hidden layers, and neurons was used to customize the neural network. The data used in this study is a part of the publicly available dataset containing real outdoor measurements provided by The National Renewable Energy Laboratory (NREL). The proposed model has been validated by measuring prediction errors using normalized mean squared error (NMSE) and prediction accuracies using regression value (R). The lowest value of the NMSE error was obtained with a neural network model based on three hidden layers employing 40, 8, and 5 neurons respectively. The R-value of this model was the highest among all models. The results have shown that the ascending/descending distribution of neurons in hidden layers is an important factor among other parameters.

Keywords: *Artificial neural network; solar irradiance; meteorological modeling; curve fitting*

1. Introduction

Today, the reserves of fossil fuels, which are used as the main energy source, are gradually decreasing. However, considering the continuous increase in energy demand and the harm of fossil fuels to the environment all over the world, the tendency towards alternative energy sources becomes increasingly important. Renewable energy is the best alternative energy source for fossil fuels thanks to its limitless resources, environmental friendliness and increasing efficiency with developing the technology. Solar energy is one of the most widely used renewable energy sources. Determining the solar energy potential is important for the proper design of solar applications. Therefore, many countries have measuring stations to measure meteorological data including solar irradiance. It is essential to develop methods for estimating solar irradiance from meteorological data available in areas where measurement stations are not available, or at intervals not measured.

Making a mathematical model of a system is one of the best methods to describe how it functions. It might be required to incorporate experimental or simulation data from the system to be modeled while building a mathematical model. Usually, these data are in the form of numerical quantities and can be visually represented as data points. These data points are essentially independent variables within the system and dependent variables affected by these variables. According to independent variables, the function capable of supplying the dependent variable or variables correctly is the target function. In other words, the target function is a parametric function that will define the closest possible points to all data points. This function defines the relationship between the data and the physical process or actual system represented in the background. The model thus obtained can be used to understand the responses of the system and/or to predict data that has not yet been measured. In this way, choosing a suitable function or selecting the most suitable parameters for a basic function requires a good understanding of the system targeted for the model.

If the target function is indicated as $f(x) = y$, here, x is the independent variable and y is the dependent variable and f is the target function. However, it may not always be possible to reach the perfect target function that defines all these variables continuously. This is a common problem in data analysis. This problem is known as the "Curve Fitting" problem, which can be explained as determining the function closest to the target function or searching for new functions to simplify the calculations by switching the functions that are difficult to use. The function obtained while trying to reach the target function is called a hypothesis or hypothesis function. When the hypothesis function is shown as $h(x) = y'$, x is the independent variable, $y'$is the prediction, and h is the hypothesis.

As computer technology advances, the use of artificial intelligence techniques in curve fitting is gradually increasing compared to traditional iterative approaches. These techniques include artificial neural networks (ANN), genetic algorithm (GA), and fuzzy logic (FL). ANN in curve fitting has been a hot research topic in the literature with its nonlinear nature, flexible structure, self-adaptability, ability to work with missing data

*Corresponding author
E-mail address: ceyhun@mu.edu.tr

Received: 19/Aug/2024; Accepted: 30/Aug/2024.

and ease of use.

Sözen et al. proposed a study that is mapping Turkey's solar potential by using ANNs [1]. The studied ANN models differ from each other in terms of the number of hidden layers, the number of neurons in these layers, and used training algorithms. They employed the logistic sigmoid transfer function in ANN structures together with the Levenberg-Marquardt (LM), scaled conjugate gradient (SCG), and Polak-Ribiére conjugate gradient (CGP) algorithms. They chose to use latitude, longitude, altitude, month, average sunshine duration, and average temperature features for the input layer of the ANN models. They showed that for the estimate of solar radiation, the ANN-based estimation approach is preferable to the traditional regression models put forward in the literature. According to their results, a trained ANN model seems with the potential for estimating solar radiation even in regions not having monitoring stations established.

Şenkal and Kuleli proposed a similar study for the estimation of solar radiation with the help of ANN [2]. In their study, they used different learning algorithms including resilient propagation (RP), the SCG learning algorithms to train ANN models. They used latitude, longitude, altitude, month, average diffuse radiation, and average beam radiation features in the input layer of the ANN models. The dataset in the study comes from twelve stations in twelve different cities in Turkey. Data from stations in nine cities were used to train ANN, while the rest were used to test. They suggested that building solar databases and estimating solar radiation may be done affordably and efficiently by employing ANN.

In order to estimate the solar radiation for the Mediterranean area of Anatolia, Turkey, Koca et al. employed ANN in their work [3]. Data on solar radiation from two cities were used to train the proposed ANN model, while data from five cities were utilized to evaluate the model. ANN models are initialized with feature sets including different numbers of input parameters. Feature sets are formed by selecting from latitude, longitude, altitude, month, average of cloudiness and sunshine duration parameters. The number of parameters in a feature set changes from four to six. They claimed that the most important component in the calculation of solar radiation was the quantity of input parameters.

Wang et al. proposed a short-term solar irradiance prediction model using ANN [4]. They created several neural networks with different structures and then determined the most suitable network by comparing all the models with cross-validation. The models differ from each other in terms of having single or double hidden layers, and by the number of neurons in the hidden layer(s). The solar radiation data from The National Renewable Energy Laboratory (NREL) is used for training and testing the prediction model [5]. The proposed prediction model employs a multi-layered feed-forward neural network with backpropagation. They reported that the model with double hidden layers containing 18-13 neurons respectively was the most successful one.

Ozgoren et al. suggested a research to create an ANN model for calculating the sun radiation of any location in Turkey using a multi-nonlinear regression technique [6]. Using the various combinations of neural network inputs, they produced 10 distinct feed-forward back-propagation neural network models. The input parameters were selected using the stepwise multiple regression analysis. The meteorological data used in the study was collected in 2000-2006 years, from 31 different stations spread throughout Turkey. While the data from 27 stations were used for training, the rest were used for the test. After selecting the most successful model, trial and error procedure was used to determine the best network architecture depending on the number of neurons in the hidden layer. They then decided to use a network consisting of one input layer with 10 input parameters and one hidden layer with 10 neurons, as the final model to estimate the monthly global solar radiation at any location in Turkey.

Renno et al. used ANN in their study to develop a tool in order to estimate the solar energy potential of the University of Salerno [7]. They have investigated two different ANN models to predict the daily global radiation and the hourly direct normal irradiance. The first ANN model for predicting the daily global radiation consists of one hidden layer with ten neurons. The second model for predicting the hourly direct normal irradiance has one hidden layer with five neurons. Both models have a sigmoid transfer function for hidden layers and a linear function for the output layer. The models have been used with feature sets including different numbers of input parameters. They stated that the created ANN models might serve as an effective instrument for evaluating a cleaner energy system, guaranteeing an accurate assessment of the solar potential for various locations.

Bou-Rabee et al. proposed an ANN-based model to forecast for the daily average solar radiation in Kuwait by using average radiation data collected for five years in a row from five different locations in Kuwait [8]. They used a multi-layered feed-forward neural network with the back-propagation learning method. After normalizing the collected data, years 2007-2010 data were used for training the proposed model. To validate the proposed model, the year 2011 data was used. The neural network design utilized in the model was determined after comprehensive testing of several potential configurations. It comprises an input layer with four parameters and a hidden layer with ten neurons. According to the researchers, Kuwait's solar radiation may be accurately predicted using the forecasting model they devised.

Xue, in his study, used an ANN technique to predict daily diffuse solar radiation [9]. The back-propagation neural network model's efficiency and capacity for generalization were enhanced by the suggested study with

the use of two optimization techniques: particle swarm optimization and genetic algorithm. Seven input parameters including the month of the year, duration of sunshine, average temperature, rainfall, relative humidity, wind speed, and daily global solar radiation were used to predict the daily diffuse solar radiation as output. He reported that the performance of the neural network model optimized by particle swarm optimization is better than the plain neural network and the neural network model optimized by genetic algorithm.

Rodríguez et al. proposed an ANN model to forecast the amount of solar energy generated by photovoltaic units [10]. After data analysis, a vector of 146 input values were given as training input to the network. These values consisted of the season of the day, the time of day, and the irradiation values in 10 minute intervals for the last 24 hours that were the remaining 144 input values. The single output of the ANN was the predicted irradiation value. According to the researchers, the suggested method's accuracy was high enough to be implemented in systems with built-in solar generators.

This research presents the development of an ANN that can model and forecast the intensity of solar radiation utilizing independent variables such as air temperature, relative humidity, atmospheric pressure, cumulative daily total precipitation, and exact time of measurement. MATLAB [11] software was used to generate the produced ANN, and the parameters pertaining to the solution approach were tuned.

2. Artificial Neural Network

ANN is a distributed, parallel information processing architecture, which consists of interconnected processing units, each with its memory and inspired by the human brain [12]. This structure consists of a combination of artificial nerve cells –neurons– obtained by mimicking the ways in which biological nerve cells work. Artificial neurons are completely nature-inspired units and information systems that can mimic the learning ability of the human brain. The network structure formed by the connection of neurons to each other is called a neuron network or neural network [13]. The binding of neurons to each other can be likened to the binding of biological neurons to each other via synapses. Data from the external environment or from the outputs of other neurons are applied to the inputs of the next linked neuron. These inputs are then processed by the neuron and then the output is exported as the neuron output or as an input for another neuron.

ANN has the ability to learn and generalize. The ability to generalize can be summarized as the ability of a trained ANN to make accurate predictions against data it has never encountered before. An ANN consists of input, output, and intermediate layers. There are independent variables in an input layer. The output layer includes dependent variables or variables that are expected to be predicted. Increasing the number of intermediate layers helps to solve more complex problems but makes it difficult to train the network.

The neural network is successful in modeling, classifying, estimating and finding the most appropriate value. Conventional computers are faster and more successful in precise arithmetic operations and calculations based on a particular algorithm. However, ANN is successful in data that contains noise or missing parts, unlike traditional computers. While traditional computers perform only the tasks they are programmed to, the ANN does not need to be programmed. ANN can learn the instructions itself. While conventional computers store information in a specific location in memory, ANN stores information in a distributed manner throughout the network, in fact the information is distributed over the connections and weights of the network.

The objective of an ANN is to learn the system that is being modeled. ANN generally performs the learning process as in the flow chart shown in Figure 1. As shown in the figure, it is necessary to present the data of this system to the neural network and compare the response of the neural network with the expected response. The difference is used by a cost function to update the weights of all connections in the network. The learning algorithm recalculates and updates the weights in each cycle. These cycles continue until the neural network's response and the predicted response diverge less than acceptable bounds.

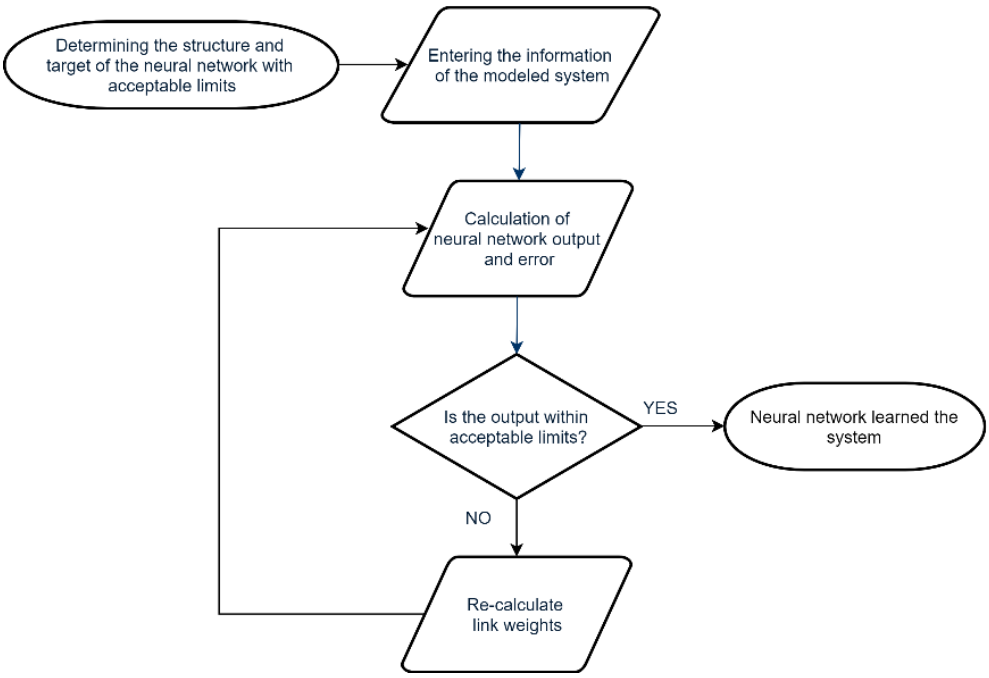

Figure 1. *Learning process of a neural network.*

3. Materials and Methods

3.1. Dataset

The raw data set used in this study contains meteorological data as well as electrical measurement data provided by NREL [14], [15]. Since the main purpose of the study was to model solar irradiance using some meteorological data, electrical data in the raw data set were not used. The meteorological data used are solar irradiance, air temperature, relative humidity, atmospheric pressure and total daily precipitation. In addition to these data, the month and hour of the measurement were used.

The ANN model to be developed within the scope of the study was used to estimate the intensity of solar irradiance using other data. In terms of the curve-fitting problem, the solar irradiance intensity is the output of the ANN while the other data are used as the inputs of the ANN as shown in **Figure 2**.

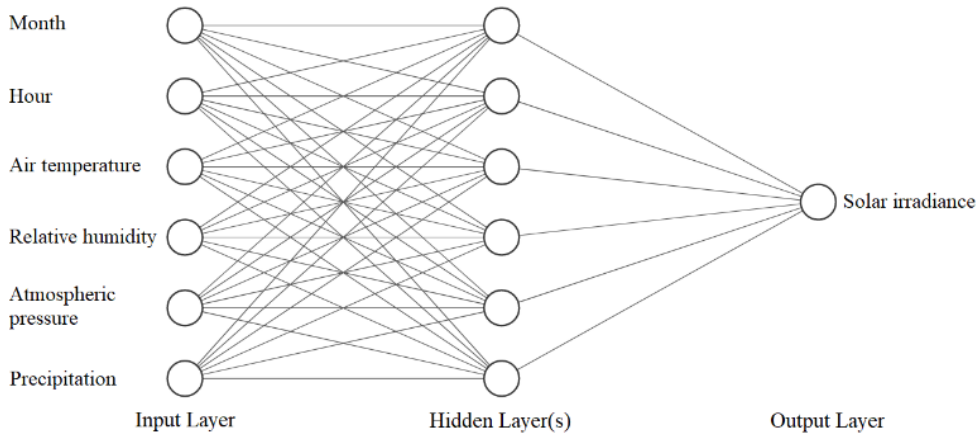

Figure 2. *Inputs and output for the target ANN structure.*

The timestamp of the measurement is a text-type statement and it is stored in the raw format of year-month-day, hour : minute : second. To use in the input variables, this raw string needs to be converted to numerical values. For this reason, the month and hour values in the timestamp were extracted and normalized. Month value varies between one and twelve, while hour value is between zero and one. For example, the hour of a measurement taken at 12:00 p.m. is assumed to be 0.5 after normalization.

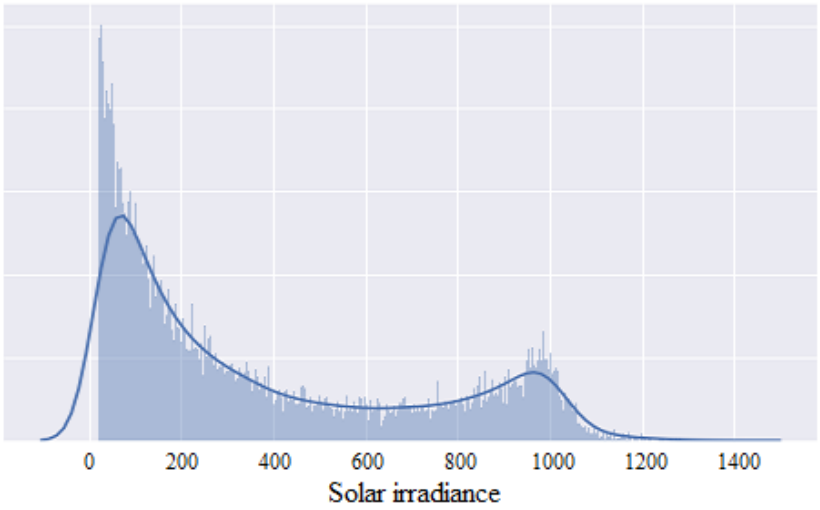

Figure 3. *Distribution of the output variable.*

Figure 3. shows the solar irradiance values' distribution in the dataset. The minimum and the maximum values for the solar irradiance data are 20W/m^2 and 1377.9W/m^2 respectively. As shown in the figure, there is a non-normal distribution in the solar irradiance data. Because of the non-normal distribution, the Spearman test has applied to see the correlations between the input variables and solar irradiance.

Figure 4. shows the correlation coefficients obtained from the Spearman correlation test, between all of the input variables and the solar irradiance output variable. As shown in the figure, there are significant positive and negative correlations between some of the input variables and solar irradiance.

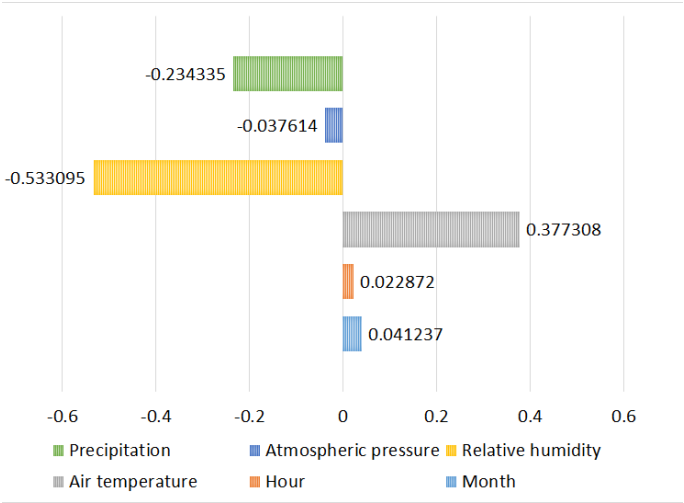

Figure 4. *Correlation coefficients between input variables and output.*

3.2. Creating and Training ANN Models

ANN models to be used in the study are distinguished from each other by parameters such as training algorithms, hidden layer numbers, number of neurons in the hidden layers, and transfer functions used in hidden layer neurons. However, some parameters were set the same in all ANN models. These parameters, which do not change according to the ANN model, are the type of neural network, the transfer function used in neurons in the output layer, and the distinction between training, validation, and test data set size. All ANN models were created in the type of feedforward and backpropagation neural networks and the linear transfer function was used in the output layer. This function is called "purelin" in MATLAB. Transfer functions are functions that calculate the output using the plain input of a neuron [16]. In addition, 70% of the data set was utilized for training, 15% for validation and the rest 15% for testing in all models.

Different ANN models have been created by using hyperbolic tangent sigmoid transfer function (*tansig*), and logarithmic sigmoid transfer function (*logsig*) in the hidden layers. This is the first parameter (*Parameter*

1) used for the optimization of the models. However, the number of hidden layers (*Parameter 2*) of the ANN models, the number of the neurons (*Parameter 3*) in these layers and the training algorithms (*Parameter 4*) are the other optimization parameters used to create different ANN models. With these parameters, 51 different ANN models were created. The neural network/data management tool (*Matlab nntool*), which provides more detailed optimization options than the curve fitting tool (*Matlab nftool*), was used to create these models. Prediction outputs and errors that were obtained after training of all ANN models are listed in the rightmost pane on the neural network/data management tool screen shown in **Figure 5**.

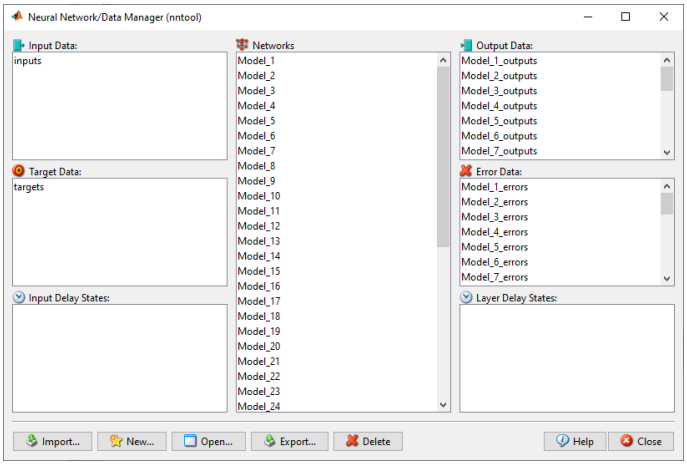

Figure 5. *Creating and Training ANN Models.*

In the study, regression R-value and the normalized mean squared error (NMSE) used as a measure of the neural network's prediction performance are commonly used statistical metrics [17]. An R-value approaching 1.0 means that the solar irradiance values predicted by the neural network have a close relationship with the actual values. If the R-value approaches 0.0, that means a random relationship. The MSE and NMSE values can be calculated with the following equations. Lower error values are better.

$$MSE(t,y) = \frac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2 \tag{1}$$

$$NMSE(t,y) = \frac{MSE(t,y)}{MSE(t,0)} \tag{2}$$

In both equations, the variable t is the actual solar irradiance values; the y variable is the solar irradiance values predicted by the network.

The results including solar irradiance predictions, errors for predictions, NMSE values, and regression curves presented in MATLAB environment, R-values, the elapsed times, and the numbers of cycles for the training of the ANN models were recorded. Evaluations regarding these recorded results will be discussed in the following section.

4. Results and Discussion

When the results are examined, the following findings are obtained. Increasing the number of hidden layers in ANN models increases the accuracy of the predictions but the required time for training also increases. The learning algorithm of seven of the top 10 most successful models is the Levenberg-Marquardt algorithm. Model 8 with 3 hidden layers and 10-20-10 neuron distribution achieved an R-value of 0.90261 in 50.24 seconds, Model 49 with 3 hidden layers and 20-10-10 neuron distribution achieved an R-value of 0.90831 in 22.68 seconds. Here the only difference between Model 8 and Model 49 was the neuron distribution. Only changing the distribution of the neurons from 10-20-10 to 20-10-10 reduced the training time by nearly 50% besides increased the accuracy slightly.

Similarly, Model 50 with 30-10-3 neuron distribution achieved an R-value of 0.91255 in 25.75 seconds, Model 51 with 40-8-5 neuron distribution achieved an R-value of 0.92036 in 33.48 seconds.

Briefly, a decreasing neuron distribution from the first hidden layer to the last, significantly reduced training time and increased the prediction accuracy. The results of the four most successful models among the 51 ANN models obtained with different parameters are shown below in the **Figure 6**. In this figure, the regression R-values, the number of hidden layers, NMSE values, elapsed time, and required epochs for the training are

shown in a visual manner to show the performance of these four most successful models.

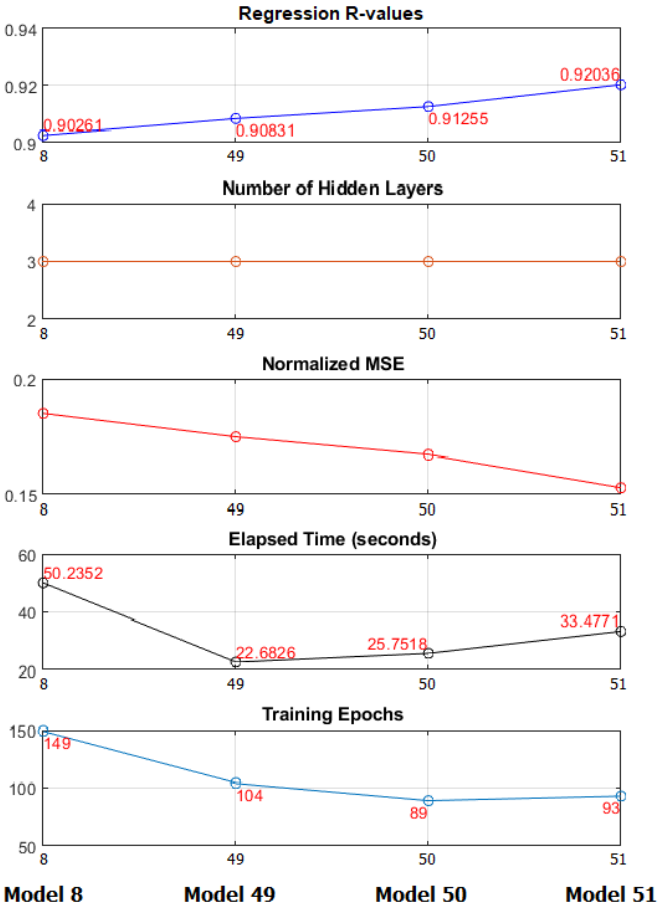

Figure 6. *The results of the four most successful models.*

Model 8 with 3 hidden layers and 10-20-10 neuron distribution achieved an R-value of 0.90261 in 50.24 seconds, Model 49 with 3 hidden layers and 20-10-10 neuron distribution achieved an R-value of 0.90831 in 22.68. Here the only difference between Model 8 and Model 49 was the neuron distribution. Only changing neuron distribution from 10-20-10 to 20-10-10 reduced the training time by nearly 50% and slightly increased the accuracy. Similarly, Model 50 with 30-10-3 neuron distribution achieved an R-value of 0.91255 in 25.75 seconds, Model 51 with 40-8-5 neuron distribution achieved an R-value of 0.92036 in 33.48 seconds. Briefly, a decreasing neuron distribution from the first hidden layer to the last, significantly reduced training time and increased the prediction accuracy.

The four regression plots in Figure 7 show the relationship between the outputs from the four most successful models and the actual values. The plots show prediction accuracies for the entire data set. The closer predictions to the actual values cause a more homogeneous and collective distribution of the data points on the 45-degree green dashed line in the graphs. The detailed regression plot in Figure 7 has been formed as an alternative to the black-and-white regression plot obtained from MATLAB, which shows the prediction accuracy in the whole data set. Although the data points in the regression plot originating from MATLAB are actually the predicted outputs and the actual target values, they are not distinguishable because they are all drawn in black colour. In the alternative regression plot in Figure 7, red data points are the actual target values, while blue data points show the predicted outputs. The green and dashed line in the figure is the first order polynomial curve between the network's predicted values and actual target data. The black coloured and continuous line in the figure is the regression curve formed from the values obtained as a result of the regression analysis that measures the relationship between the predicted values and the actual data.

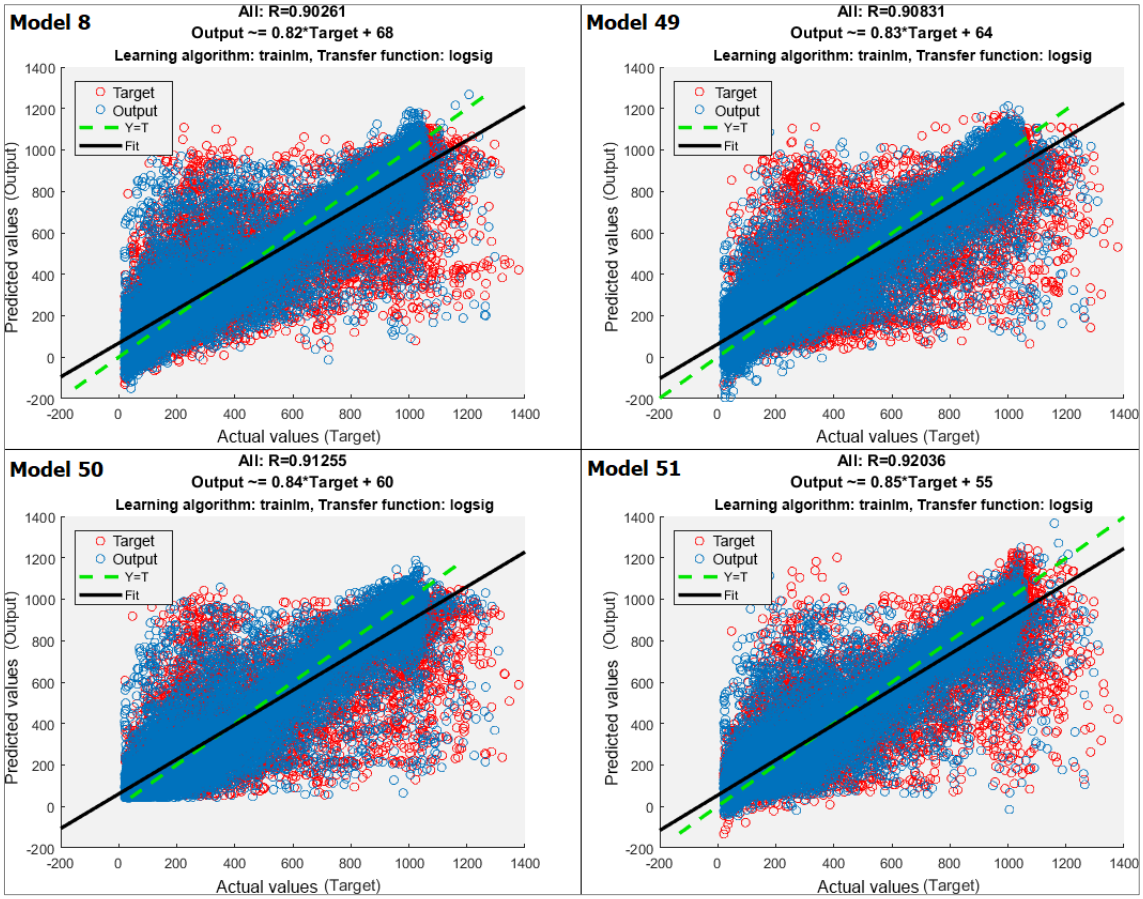

Figure 7. *The regression plots of the four most successful models.*

5. Conclusion

In this study, it was aimed to reveal and optimize the parameters in the development of the most suitable ANN structure for modeling solar irradiance using some climatic data. In the study, 51 different ANN models, which are separated from each other with parameters related to the solution method, were designed and examined.

The dataset was selected from the data collected at the measuring station installed in the Oregon University campus by NREL researchers. The measurement data had been collected between 20 December 2012 and 20 January 2014. Within the scope of the study, only measurement data of 2013 were used. The meteorological data used in the study are the solar irradiance as the target value and the independent variables such as the air temperature, relative humidity, atmospheric pressure, and cumulative daily total precipitation. Along with these data, the month and the hour values extracted from the measurement timestamp was used. After removing the samples with missing values, 70% of the data was used to train the ANN models, 15% for validation, and the remaining 15% were used to test the generalization ability against the new data faced by the trained models. The performances of the models were compared in terms of the regression R-values, NMSE error, the time spent training the models, the number of epochs required in the training.

The results obtained from all ANN models showed that the selected training algorithms and the number of hidden layers forming the architectural structure of the network and the number of neurons within these layers have important effects on the network's predictive performance. Since the transfer functions used in neurons were selected according to regression problems, they have been less determinant of the performance of the models. When using a transfer function that is not suitable for regression problems, the network has lost almost all of its prediction capabilities. Therefore, transfer functions suitable for problems such as classification and clustering were not addressed. Certain training algorithms such as trainlm and trainrp showed a significant increase in their predictive quality by increasing the number of hidden layers used in the ANN model or changing the neuron distribution in the hidden layers. Both algorithms achieved the highest predictive performance with the ANN model with 3 hidden layers. When neurons in hidden layers were distributed from the first hidden layer to the last in descending order, both accuracy increased and processing time shortened significantly. Using the advantage of reduced processing time, adding more neurons to hidden layers has led

to more successful models. This advantage can be used until the processing time-accuracy tradeoff is broken. As a further study, models using deep learning can be created and compared with the models obtained in this study.

Declaration of interest

There is no conflict of interest, according to the authors.

Nomenclature

Abbreviations

ANN	artificial neural network
CGP	polak-ribiére conjugate gradient algorithm
FL	fuzzy logic
GA	genetic algorithm
LM	levenberg-marquardt algorithm
MSE	mean squared error
NMSE	normalized mean squared error
NREL	national renewable energy laboratory
R	regression value
SCG	scaled conjugate gradient algorithm

References

[1] A. Sözen, E. Arcaklioğlu, M. Özalp, and E. G. Kanit, "Use of artificial neural networks for mapping of solar potential in Turkey," *Applied Energy*, vol. 77, no. 3, pp. 273–286, Mar. 2004.

[2] O. Şenkal and T. Kuleli, "Estimation of solar radiation over Turkey using artificial neural network and satellite data," *Applied Energy*, vol. 86, no. 7–8, pp. 1222–1228, 2009.

[3] A. Koca, H. F. Oztop, Y. Varol, and G. O. Koca, "Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8756–8762, 2011.

[4] Z. Wang, F. Wang, and S. Su, "Solar irradiance short-term prediction model based on BP neural network," *Energy Procedia*, vol. 12, pp. 488–494, 2011.

[5] B. Marion *et al.*, "Data for Validating Models for PV Module Performance," 2014.

[6] M. Ozgoren, M. Bilgili, and B. Sahin, "Estimation of global solar radiation using ANN over Turkey," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5043–5051, 2012.

[7] C. Renno, F. Petito, and A. Gatto, "ANN model for predicting the direct normal irradiance and the global radiation for a solar application to a residential building," *Journal of Cleaner Production*, vol. 135, pp. 1298–1316, 2016.

[8] M. Bou-Rabee, S. A. Sulaiman, M. S. Saleh, and S. Marafi, "Using artificial neural networks to estimate solar radiation in Kuwait," *Renewable and Sustainable Energy Reviews*, vol. 72, no. November 2016, pp. 434–438, 2017.

[9] X. Xue, "Prediction of daily diffuse solar radiation using artificial neural networks," *International Journal of Hydrogen Energy*, vol. 42, no. 47, pp. 28214–28221, 2017.

[10] F. Rodríguez, A. Fleetwood, A. Galarza, and L. Fontán, "Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control," *Renewable Energy*, vol. 126, pp. 855–864, 2018.

[11] "Matlab 2017b," 2017. [Online]. Available: https://ww2.mathworks.cn/en/. [Accessed: 04-May-2019].

[12] Ç. Elmas, *Yapay Zeka Uygulamaları Yapay Sinir Ağları – Bulanık Mantık– Genetik Algoritma*, 4th ed. 2011.

[13] S. Haykin, *Neural Networks and Learning Machines, 3d Edition*, 3rd ed. ew Jersey: Pearson Education, 2008.

[14] "Photovoltaic Research, NREL," 2019. [Online]. Available: https://www.nrel.gov/pv/index.html. [Accessed: 07-May-2019].

[15] B. Marion *et al.*, "New data set for validating PV module performance models," in *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, 2014, pp. 1362–1366.

[16] C. Kubat, *MATLAB Yapay Zeka ve Mühendislik Uygulamaları*. Abaküs, 2019.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103, no. 4. New York, NY: Springer New York, 2013.