Hindawi
International Journal of Photoenergy
Volume 2020, Article ID 8843620, 26 pages
https://doi.org/10.1155/2020/8843620

Research Article

Effective Estimation of Hourly Global Solar Radiation Using Machine Learning Algorithms

Abdurrahman Burak Guher,[1] **Sakir Tasdemir**,[2] **and Bulent Yaniktepe**[3]

[1]*Department of Computer Technologies, Osmaniye Korkut Ata University, Osmaniye 80000, Turkey*
[2]*Department of Computer Engineering, Selcuk University, Konya 42130, Turkey*
[3]*Department of Energy Systems Engineering, Osmaniye Korkut Ata University, Osmaniye 80000, Turkey*

Correspondence should be addressed to Bulent Yaniktepe; byaniktepe@osmaniye.edu.tr

Received 31 August 2020; Revised 1 November 2020; Accepted 18 November 2020; Published 10 December 2020

Academic Editor: Gianluca Coccia

The precise estimation of solar radiation is of great importance in solar energy applications with respect to installation and capacity. In estimate modelling on selected target locations, various computer-based and experimental methods and techniques are employed. In the present study, the Multilayer Feed-Forward Neural Network (MFFNN), *K*-Nearest Neighbors (*K*-NN), a Library for Support Vector Machines (LibSVM), and M5 rules algorithms, which are among the Machine Learning (ML) algorithms, were used to estimate the hourly average solar radiation of two geographic locations on the same latitude. The input variables that had the most impact on solar radiation were identified and grouped as a result of 29 different applications that were developed by using 6 different feature selection methods with Waikato Environment for Knowledge Analysis (WEKA) software. Estimation models were developed by using the selected data groups and all input variables for each target location. The results show that the estimations developed with the feature selection method were more successful for target locations, and the radiation potentials were similar. The performance of the estimation models was evaluated by comparing each model with different statistical indicators and with previous studies. According to the RMSE, MAE, R^2, and SMAPE statistical scales, the results of the most successful estimation models that were developed with MFFNN were 0.0508-0.0536, 0.0341-0.0352, 0.9488-0.9656, and 7.77%-7.79%, respectively.

1. Introduction

Energy, which is an effective parameter in the development of countries, is increasing rapidly with industry, technological advances, and increasing population. Not every country has adequate energy resources to meet the need for energy, and the rapid increase in energy consumption forces countries to turn to alternative sources in energy supply. For this reason, countries prefer renewable energy sources such as solar, wind, hydro, bio, hydrogen, geothermal, and tidal energy to meet their energy needs instead of conventional energy sources [1]. Solar energy, which plays a critical role in electricity generation with each passing day, has become one of the promising renewable energy sources attracting the attention of countries because it is clean, unlimited, and sustainable compared to fossil fuels. As a result of this, investments in solar energy for electricity generation are increasing rapidly in recent years with technological advances in solar energy, global climate change, dependence on other countries, and other environmental factors. In this context, photovoltaic (PV), as one of the usages of solar energy application areas, is intensively applied in order to produce electricity [2, 3].

PV, which is used reliably in electricity generation, has been growing rapidly in the world for more than 40 years, and the amount of electrical energy produced from PV power plants has reached 480 GW [4]. Before designing and modelling a PV system in a selected geographical area, solar radiation (SR) data must be measured as the most important input value, where the feasibility of the designs made in terms of investment can be evaluated according to this data. This value is not only necessary in PV designs but is also the most

important parameter in many scientific and engineering works on solar energy practices [5]. For this reason, it is the most accurate method to obtain long-term data in a selected special geographic area. However, measuring the SR everywhere is often not possible, as it requires costly, long, and precise processes. In addition, radiation values cannot be measured in an accurate way in most countries because the measurements can only be made in certain areas. For this reason, experimental, statistical, and Artificial Intelligence-(AI-) based estimation methods were developed to calculate the value of SR worldwide [6–8]. ML algorithms, which are a subfield of AI, are one of the most common methods used in estimation studies.

Many studies have been conducted in recent years based on ML algorithms in different geographic areas of the world to estimate SR. In these studies, algorithms including Artificial Neural Network (ANN), Support Vector Machines (SVMs)/(Support Vector Regression) SVR, *K*-NN, Linear/-Nonlinear Regression, M5, and Random Forests have been used frequently [9]. However, estimation models were developed in these studies by selecting a specific geographical area of a country or different geographical locations in the country [10]. Before the development of estimation models for a selected geographic location, it must be decided which hourly, daily, and monthly average global radiation values that fall onto a certain horizontal surface will be used [11]. Notton et al. [12] recommended that the monthly average data should be used if preliminary modelling or draft design is required to be done, and the daily average data can be used if a more comprehensive design is to be established. However, they also indicated that it is necessary to use hourly average or shorter-scale data in more precise and result-oriented designs. Zhang et al. [13] explained that the estimation processes of studies with hourly data compared with daily and monthly data are more difficult and complex. For this reason, estimation models made with hourly data are less common since they contain more difficult and complex processes. After determining the input data according to the type of work that will be carried out, the SR values of the target area can be estimated by using *one* [14], *multiple* [15], or *hybrid* [16] ML algorithms. Different solutions were sought for Global Solar Radiation (GSR) estimation problems in developed models by making changes on the functional structure and architecture of one single ML algorithm by comparing multiple algorithms or by working two or more AI methods together.

It is possible to classify the planned studies in which the ML method is used in GSR estimation in three different categories according to the measurement time intervals of SR: Monthly Average Global Solar Radiation (MAGSR), Daily Average Global Solar Radiation (DAGSR), and Hourly Average Global Solar Radiation (HAGSR). HAGSR- [17, 18], DAGSR- [19–22], and MAGSR- [23–26] based estimation models were developed by using one single ML algorithm, and it was noticed that the ANN algorithm was used frequently compared to other algorithms because of its flexible structure and accuracy. On the other hand, studies on the methods in which multiple ML algorithms can be analyzed and used together at the same time are increasing rapidly. In these types of studies, a clear idea can be achieved on the effectiveness of each ML algorithm on the dataset used, and the most successful models can be compared and evaluated. In this context, Pang et al. [27] estimated GSR comparatively by using ANN and Recurrent Neural Network (RNN) ML algorithms in 10-, 30-, and 60-minute time zones. Li et al. [28] developed estimation models with the help of seven-year measured hourly data with the Multivariate Adaptive Regression Spline (MARS) ML algorithm to estimate HAGSR and compared their results obtained in Hong Kong with the ANN and logistic regression algorithms. They reported that ANN achieved superior performance compared to the other algorithms. Khosravi et al. [29] developed the most successful estimation models to estimate HAGSR for two different network groups on the Iranian island of Abu Musa by using MFFNN, Radial Basis Function Neural Network (RBFNN), SVR, Fuzzy Inference System (FIS), and Adaptive Neuro-Fuzzy Inference System (ANFIS) ML algorithms. The first network was planned with five inputs, and the second network was planned with one single input, and it was reported that the SVR reached superior estimative accuracy than other algorithms on both networks. Lotfinejad et al. [30] investigated the DAGSR of different cities of Iran by using Bat Neural Network (BNN), Generalized Regression Neural Network (GRNN), and Neuro-Fuzzy (NF) algorithms. They reported that the models developed with the recommended BNN algorithms performed better than other algorithms. Meenal and Selvakumar [31] examined a comparative DAGSR estimation model among SVM, ANN, and experimental models by identifying the most suitable input variables from nine input data from four different cities of India and showed that SVM was more successful than the other algorithms. Loutfi et al. [32] developed ten different HAGSR estimation models in the city of Fes, Morocco, with the help of nine different input variables from 2010 to 2014 five-year with Multilayer Perceptron (MLP) and Neural Autoregressive with Exogenous Inputs (NARX) algorithms. Among the models developed, they contended that the most successful estimation model was the model developed with NARX. Lazzaroni et al. [33] compared the GSR estimation models that were developed according to hourly, daily, and monthly time zones with SVR and Extreme Learning Machine (ELM) ML algorithms by using three-year hourly data in Milan, Italy, with the *K*-NN algorithm. Long et al. [34] investigated the estimation of DAGSR by using ML-based ANN, *K*-NN, SVM, and Multivariate Linear Regression (MLR) algorithms and made a comparative analysis of data-driven algorithms. Ozgoren et al. [35] compared the estimation models developed with the ANN and Multi-Nonlinear Regression (MNLR) algorithms to estimate MAGSR in 31 cities of Turkey using five-year input data collected between 2002 and 2006. Moghaddamnia et al. [36] estimated the DAGSR by using the different meteorological parameters of Britain's Brue Basin by using the Local Linear Regression (LLR), NARX, MLP, Elman Network, and ANFIS ML algorithms.

In the present study, the purpose was to comparatively analyze the HAGSR of two geographical provinces located on the same latitude of the Mediterranean Region by using

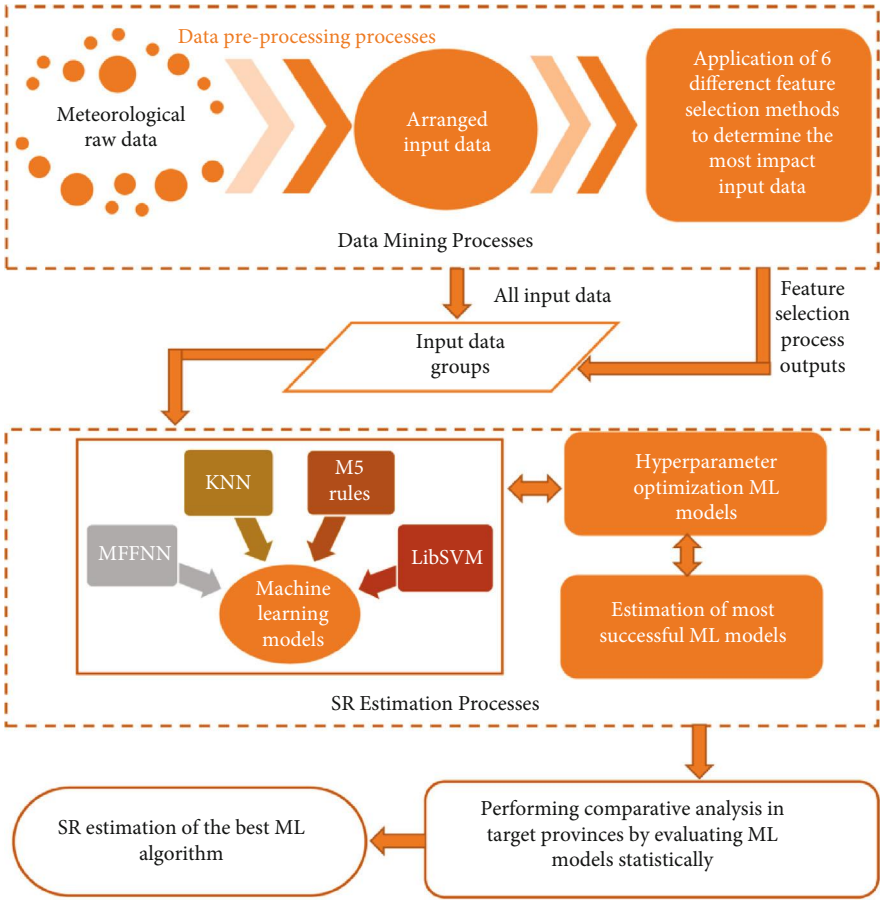


Figure 1: The framework of methodology of the present study.

four different ML algorithms MFFNN, K-NN, M5 rules, and SVR-based LibSVM library. Another purpose was to use the WEKA software program to determine the features of input data that has the most impact on SR. For this purpose, the best features were determined in five groups by developing twenty-nine different applications with the help of six feature selection functions. The eventual models of ML algorithms that were used in the study were developed according to the output groups of feature selection functions. HAGSR estimation models were evaluated with respect to among themselves and also on the basis of the algorithm that was used, and the results were then compared with similar studies. In addition, unlike other previous studies, the present study developed estimation models and evaluated their performance by using the classical SVR algorithm and LibSVM software, which are similar to each other. The framework outlining how the data mining processes and four different ML algorithms are used in this study to evaluate the solar radiation potential of two provinces in the same latitude is shown in Figure 1. The WEKA software was used in data mining processes such as data preprocessing and feature selection, and Matlab R2017b software was made use in modelling studies developed with ML algorithms used in SR estimation.

The rest of the study is organized as follows. The provinces for which the models were developed and the meteorological and categorical dataset used in the study are defined in Section 2. The details of the feature selection processes that were used to determine the most appropriate input data groups, the methodologies of the MFFNN, SVR-LibSVM, K-NN, and M5 rules algorithms, the architectural and functional structures of the developed models, and the methods applied are also explained in this chapter. Section 3 includes the results and comparative analyses of the estimation models developed with the ML algorithm used for each input data group. The HAGSR estimation performance of the two provinces, which are located in the same latitude, was evaluated with multiple statistical error methods and was also compared with previous similar studies. The results of this study and its contribution to the literature are summarized in Section 4.

2. Materials and Methods

In this section, the evident features of the two selected provinces and the editing of data to be used in ML models are explained. Then, the selection procedures of the most effective input groups are mentioned using feature selection processes. The input data were determined in five different groups at the end of the selection process, and the development processes of the best ML models were explained for each group. In addition, the structural characteristics of the

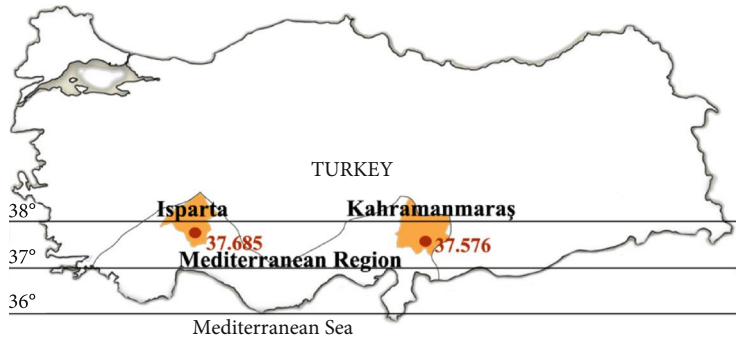


Figure 2: Geographical positions of target provinces and measurement stations.

ML algorithms that were used in the comparative estimation of HAGSR and the statistical scales that were used in evaluating model performance are discussed in detail.

2.1. Study Area and Preparation of Database. The provinces of Kahramanmaras and Isparta were selected as the study areas by considering the climatic characteristics, elevation, various different geographical characteristics, and in particular latitude. The selected provinces are located in the Mediterranean Region and have a high solar power potential with an average annual sunshine time of 2956 hours and an average annual amount of solar energy of 1390 kWh/m^2 [37]. The location of the selected provinces in the Mediterranean Region and the latitude coordinates of meteorological measurement stations are given in Figure 2.

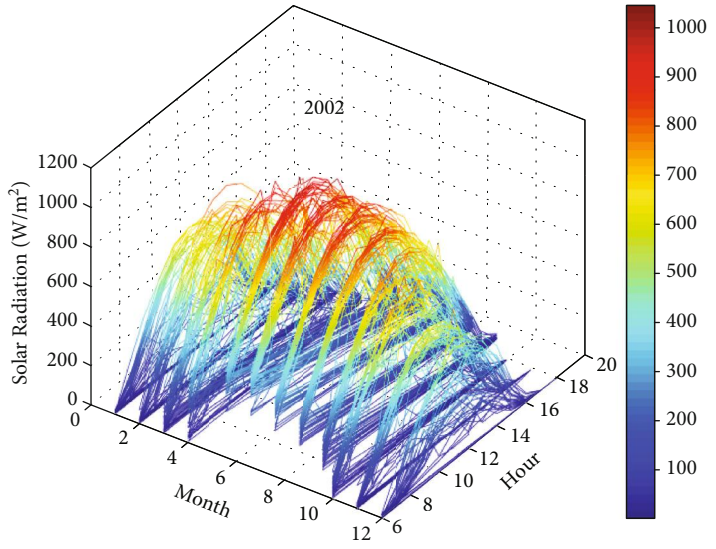
Radiation data is the most important parameter used in solar energy-based systems. However, the radiation data value cannot be measured at every measuring station across the country; instead it is measured at a limited measuring station. SR was measured for certain locations by the Turkish State Meteorological Service (TSMS), which is a government agency with a large network of stations in Turkey. In the study, the data collected for the target provinces consisted of meteorological data measured by TSMS between 2002 and 2006. These data used were the hourly average data that were measured every 5 minutes and meteorological data from measuring stations collected from Hourly Pressure (P), Hourly Sunshine Duration (HSD), Hourly Humidity (H), Hourly Temperature (T), Hourly Wind Speed (WS), and Hourly Solar Irradiance (HSI). 3D plots showing the change of SR for both monthly and seasonal measurements of yearly intervals of Kahramanmaras and Isparta are given in Figures 3 and 4. The annual distributions of the SR values measured in these charts are given in detail on an hourly basis. The specific characteristics of geographical and meteorological data of the target provinces are given in Table 1.

The data preprocessing that will improve the quality of the raw data to be used in the study is one of the most important processes that have a direct positive effect on the performance of all computer science-related algorithms [38]. Since ML algorithms are generally data-focused structures, several operations like cleaning, scaling, reduction, and normalization have significant effects on the accuracy of the estimation [39]. In the present study, four categorical data were included in the meteorological dataset including year of measurement (year), month of the year (month), day of the month (day), and hour of the day (hour). Geographical data were not used since the effect of the latitude was evaluated. Measurement time intervals of the other meteorological data were determined according to HSI measurement time intervals. The data between 06:00-17:00 hours for January and February; 06:00-18:00 hours for March, April, October, November, and December; and 07:00-19:00 hours between May and September were selected. Factors such as measurement time differences between years, variability of the measured time zones of each month, and winter time-summer time were effective in selecting the time ranges. Any missing data was calculated by taking the arithmetic average of the data in the same time frames of the previous and following years, and the data that were calculated in this way constituted approximately 4% of all data. A total of 23442 SR data were obtained for each province. After the raw data were arranged and determined, min-max normalization was applied and scaled. The normalization formula applied is given in equation (1). In this formula, each input (x_i) value was normalized (X_n) linearly between the 0 and 1 range by finding the minimum ($x_{\min}$) and maximum ($x_{\max}$) values of the raw dataset [40].

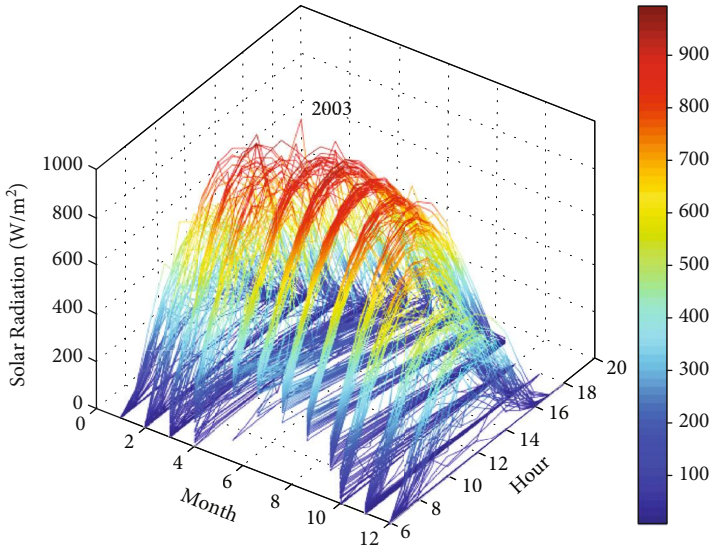
$$X_n = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \quad (1)$$

In data selection, since different estimation results are obtained each time when a certain year range is used in training the model and the remaining years are used to test the model, it was ensured that all data at hand were randomly allocated hourly with a specifically coded program, instead of determining year-based training and test data. In this way, it was aimed that the output results of the estimation models were not affected by the data selection by providing a homogeneous distribution in the input data according to years. The number and basic characteristics of the training and test data, which were determined hourly for each province, are given in Table 2.

2.2. Selection of the Best Input Data Groups with WEKA. Since the data pool used in ML-based GSR estimation studies are quite extensive, the characteristics of the data and their relations with each other affect the output performance of models. Although some data have positive effects on the



(a)



(b)

Figure 3: Continued.

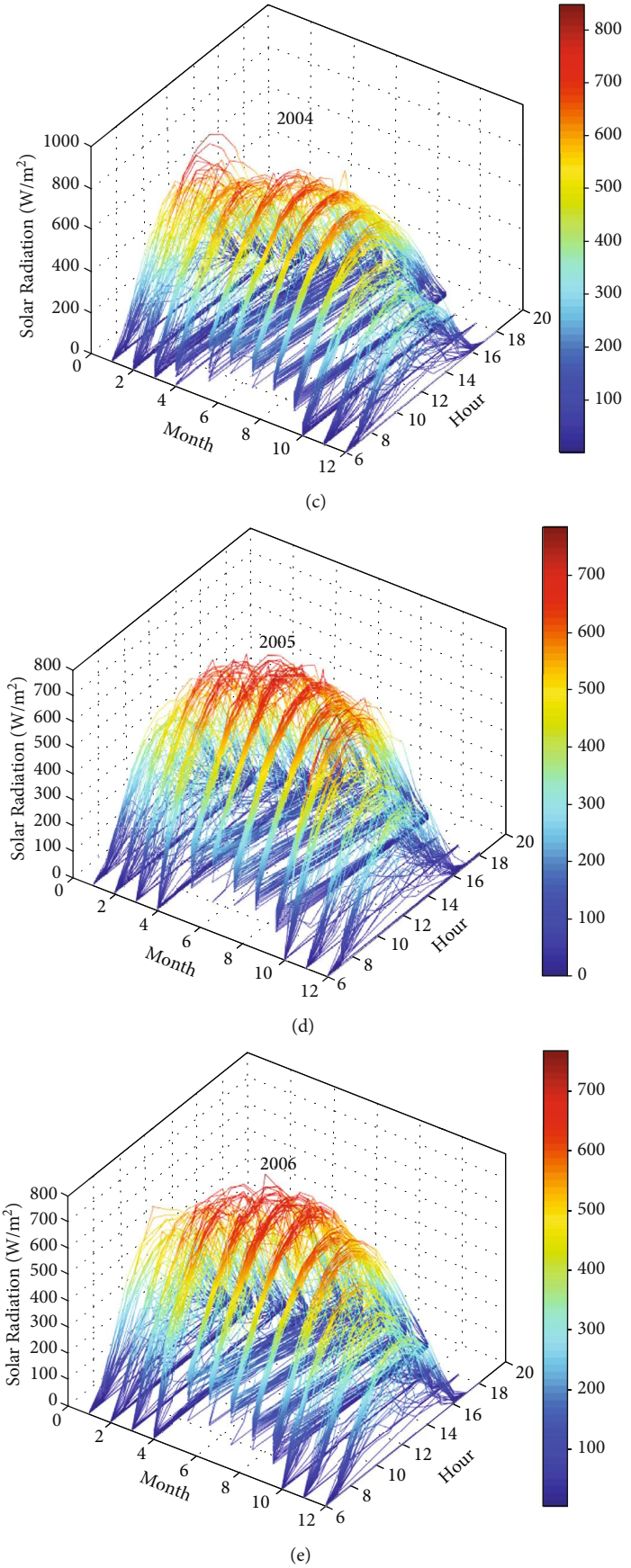
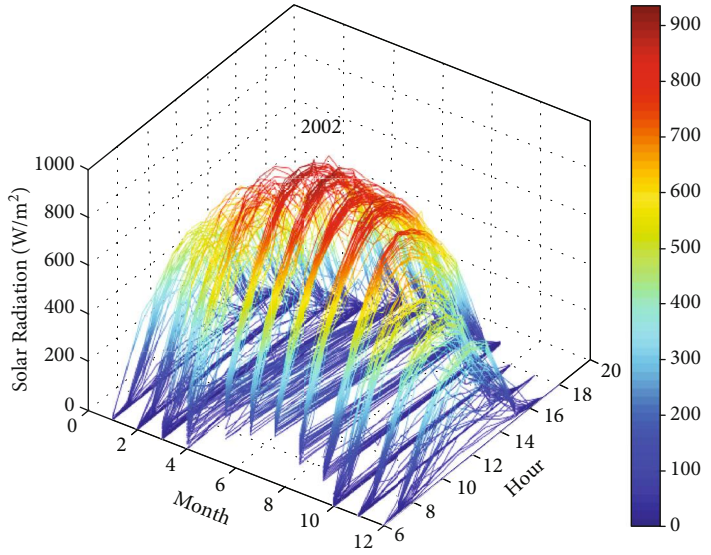
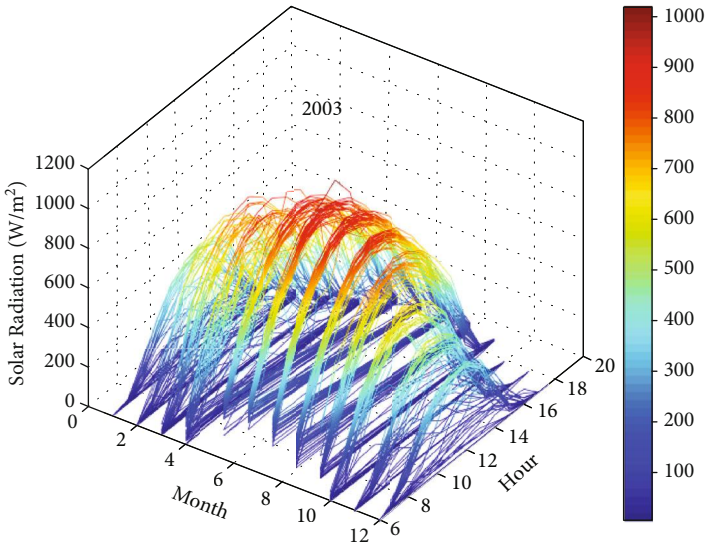


Figure 3: The 3D distribution plots of HAGSR of Isparta measured for the years (a) 2002, (b) 2003, (c) 2004, (d) 2005, and (e) 2006.

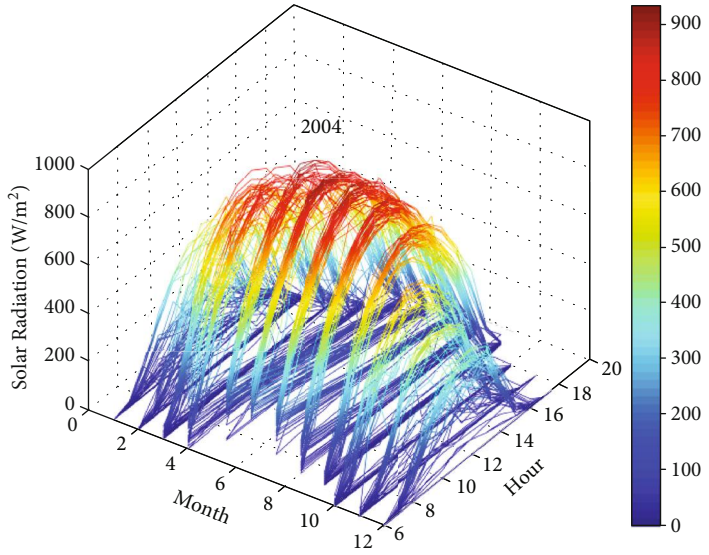


(a)

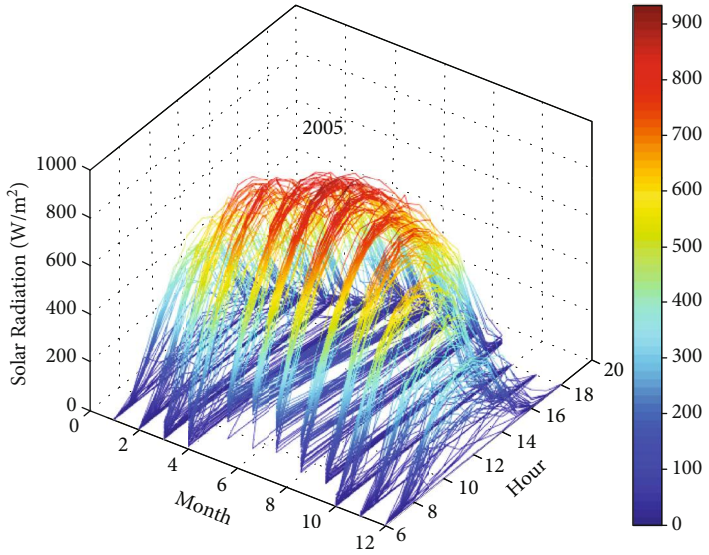


(b)

Figure 4: Continued.

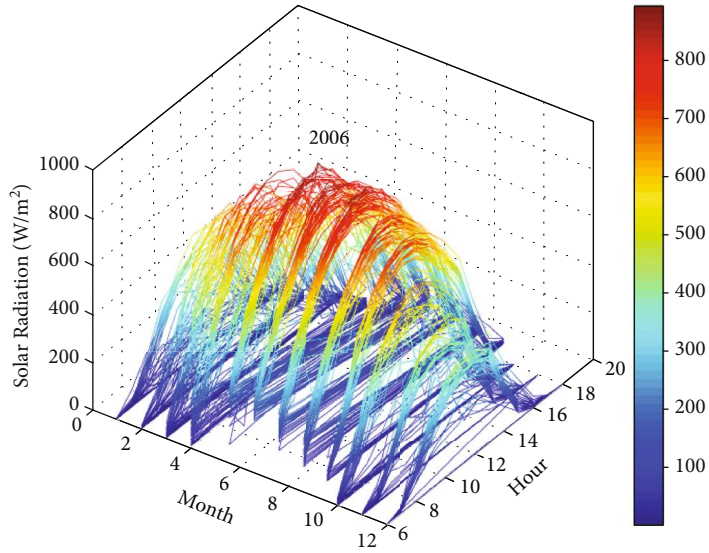


(c)



(d)

Figure 4: Continued.



(e)

Figure 4: The 3D distribution graphs of HAGSR of Kahramanmaras measured for the years (a) 2002, (b) 2003, (c) 2004, (d) 2005, and (e) 2006.

Table 1: Specific characteristics of geographical and meteorological data of each province.

Selected location	Geographical data				Meteorological data											
	Lat. (Deg.)	Lon. (Deg.)	Alt. (m)	Area (km^2)	T (°C)		P (hPa)		H (%)		WS (m/s)		HSD (hour)		HSI (W/m^2)	
					Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
K.Maraş	37.576	36.915	872	14336	-8.2	43.9	928.3	966.1	0	99	0	9.9	0	1	0.2	1028.8
Isparta	37.785	30.568	1050	8933	-14	36.8	876	918.9	7	99	0	10.1	0	1	0.1	1046.7

Table 2: Basic structure of the training and test input data determined hourly to be used in modelling.

	Meteorological and Categorical		Time range
	Training data	Test data	
Total	17584	5858	2002-2006
Rate	±75%	±25%	
Selection type	Random	Random	
Input data	Year, Month, Day, Hour, T, P, H, WS, HSD		
Output data	SR		

output, some others may have negative effects, and some have no effect. For this reason, determining the most effective data features on output prior to the modelling process will decrease the dimensionality of the data employed in this process, facilitate the interpretation of the estimation, and shorten the modelling process increasing the estimation accuracy [10, 41]. The methods and techniques used in the selection of the features affecting SR the most as well as the methods and applications used in the current study are compared with similar ML-based studies in Table 3. In feature selection methods used commonly, the features that have the greatest effect on the SR data are found and a new input dataset is determined. Unlike in previous studies, different input data groups were created in this study by evaluating the different input parameters affecting SR data with multiple applications that were developed by the selection methods applied.

In this study, the open-source WEKA program was used to select the most affected features of SR. This program was developed by Waikato University by using the JAVA programming language. Two feature selection methods based on the wrapper and filter approaches were used to select features that most influenced the SR data in the program. Although the filter approach uses simple, fast, and scalable methods, the wrapper approach processes the data by using classification-based techniques. The relations between different features selected in each application and the classifier models were evaluated in this study in the selection processes [43].

Instead of processing the data with one single feature selection method, it was aimed to evaluate the effect of input variables on SR by developing multiple applications in each function by using six different feature selection functions based on filter and wrapper approaches. Three of these work as wrapper approach-based functions, and the other three work as filter-based functions. Classifier Subset Evaluator

Table 3: Comparison of feature selection methods and techniques used in previous studies and those used in the present study.

ML algorithm	Authors [Ref.]	Feature selection method/algorithm/application	Selection number of the best input data groups
ANN	Yadav et al. [42]	WEKA/J48	1
SVM-ANN	Meenal and Selvakumar [31]	WEKA/Greedy Stepwise	1
MLP-NNARX-ANFİS	Moghaddamnia et al. [36]	Gamma test	1
ANN	Ozgoren et al. [35]	MNLR	1
MARS	Li et al. [28]	Sensitivity analysis/MARS	1
ANN-SVM-KNN-MLR	Long et al. [34]	Linear, Pace, and SVM regression	3
BNN-GRNN-ANFIS	Lotfinejad et al. [30]	Principal component analysis	1
MFFNN-SVR-KNN-M5 rules	Present study	WEKA/6-feature selection function/10-ML algorithm/29-application	6

Table 4: Output of a sample feature selection process applied to the dataset of Isparta.

Evaluator	weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search	weka.attributeSelection.BestFirst -D 1 -N 5
Instances	17584
Attributes	10 (Year, Month, Day, Hour, Pressure, Temp, Humidity, W.Speed, H.S.Duration, S.Radiation)
Evaluation mode	10-fold cross-validation
Number of folds (%)	Attribute
10 (100%)	1 Year
0 (0%)	2 Month
0 (0%)	3 Day
0 (0%)	4 Hour
0 (0%)	5 Pressure
0 (0%)	6 Temp
10 (100%)	7 Humidity
0 (0%)	8 W.Speed
10 (100%)	9 H.S.Duration

(CSE), Wrapper Subset Evaluator (WSE), and Classifier Feature Evaluator (ClassAE) are based on the wrapper approach; and Correlation-based Future Selection Subset Evaluator (CfsSE), Correlation Feature Evaluator (CorrAE), and Relief Feature Evaluator (RAE) are filter-based selection functions. In addition, two basic methods (i.e., random and comprehensive search) were used based on the type of feature selection function. Some selection functions support multiple search methods, and some others support only one. Search methods such as Best-First (BF), Evolutionary Search (ES), Firefly Search (FS), Elephant Search (ELS), Ant Search (AE), Linear Forward Selection (LFS), Greedy Stepwise (GS), and Ranker were used in this respect. However, ten different ML algorithms like Multi-Layer Perceptron (MLP), Linear Regression (LR), Simple Linear Regression (SLR), M5 rules, M5P, Decision Table (DT), Random Forest (RD), Additive Regression (AR), Elastic Net Regularization (ENR), and K-Nearest Neighbors Classifier (IBk) were used as classifiers in wrapper-based feature selection functions.

In the selection process of the most effective input variables, 29 different applications were developed by using a total of six different feature selection functions. A 10-fold cross-validation method was used in all applications. The screenshot of the application developed by using the CfsSE feature selection function with the BF search method is given in Table 4. It is seen in the table that the year, H, and HSD data were most effective on SR, and the other data had no effect.

Six different data groups were created for each feature selection function, with variables that most affected the SR. The input variables that affected the SR the most according to the selection functions for the provinces for which the models were developed are given in Tables 5 and 6. In processes where more than one selection was applied, the selection of the most effective features was determined by evaluating the number of applications and the impact totals of the selected variables. Consequently, the feature was not included in the data group if the impact level on SR was negative, neutral, or very low. Some feature variables calculated for selection functions and the number of inputs were similar in selection processes. The results of the CSE and WSE feature selection functions in Isparta and the results of the CSE and RAE feature selection functions in the data of Kahramanmaras were similar.

Table 5: Input features affecting the SR the most in Isparta.

Feature selection function	Selected features	Total number of feature
CfsSE	Year, H, HSD	3
ClassAE	Year, Month, Hour, P, T, H, WS, HSD	8
CSE	Year, Month, Day, Hour, P, T, HSD	7
CorrAE	Day, T, WS, HSD	4
RAE	Month, Hour, HSD	3
WSE	Year, Month, Day, Hour, P, T, HSD	7

Table 6: Input features affecting the SR the most in Kahramanmaras.

Feature selection function	Selected features	Total number of feature
CfsSE	Year, T, H, WS, HSD	5
ClassAE	Month, Hour, T, H, HSD	5
CSE	Month, Hour, T, HSD	4
CorrAE	Day, T, WS, HSD	4
RAE	Month, Hour, T, HSD	4
WSE	Month, Hour, T, H, WS, HSD	6

For each province, the final data groups and feature numbers created to be used in estimation models to be developed with ML algorithms as a result of feature selection processes are given in Table 7.

2.3. MFFNN Algorithm. ANN is an ML algorithm developed based on nerve cells specific to humans. This structure is known as a computer-modelled version of the biological and intellectual structure of the brain and is used frequently in solving problems such as estimations which cannot be calculated by nonlinear and classical calculation methods, time series problems, pattern recognition, and classification [44]. For the past 50 years, many neural network architectures have been developed based on Feed-Forward and Recurrent Networks to be used for various purposes and in a number of fields. Each architectural structure does not reach the same level of success on input data [45]. For this reason, the MFFNN ML algorithm, which was based on the Feed-Forward architectural structure, which is suitable for the available data structure and exhibits high performance, was used. Since MFFNN works with the backpropagation learning algorithm to minimize error, it has greatly increased learning success [25]. The Matlab R2017b software program was used in the development and modelling of this network. The architectural structure and working principles of the MFFNN that was used in the modelling studies are given in Figure 5.

GR1-GR5 represents the input data groups selected at the end of the feature selection process, and GR6 represents all input data that did not undergo any selection process. The architectural structure of the neural network was created in three layers, and a 5-iteration training model was developed for each neuron by using 1-50 neurons in the hidden layer. No significant increase was detected in the operating performance in neurons over 50, and the working time became considerably longer. In the developed MFFNN models, each input data (X_j) connected to neurons between layers was multiplied by a weight value (W_{ij}), added by bias (b_i), and the net input values (N_i) were calculated. The formula of the net input is given in

$$N_i = \sum_{j=1}^{m} \left(W_{ij} X_j\right) + b_i. \tag{2}$$

Net input is activated with a transfer function once it is calculated [46]. A hyperbolic tangent sigmoid transfer function (Tansig) was used between the input layer and the hidden layer and between the hidden and the output layer. By using Tansig, net-input values are scaled in the -1 to +1 range. When determining the transfer function, the logistic sigmoid (Logsig) or Tansig function was determined to be available in the hidden layer, while Tansig or Linear (Purelin) functions were available in the output layer. Choosing a function other than these significantly reduced the performance. The formula for the Tansig transfer function is given in

$$\text{Tansig}(N_i) = \left(\frac{2}{1 + e^{-2N_i}} - 1\right). \tag{3}$$

The Levenberg-Marquardt Backpropagation (Trainlm) training function was used in the MFFNN. Other training functions such as Trainbr (Bayesian Regularization Backpropagation) and Traincgb (Conjugate Gradient Backpropagation) were also tested. However, since the best performance was provided with Trainlm, this training function was selected.

2.4. SVR Algorithm. SVM is known as the ML algorithm that was developed by Vapnik and commonly used in classification problems. The smallest subsets of training data are used to find the best prediction model between two classes with SVM [47]. However, since it was not adequate in multiclass

Table 7: Final data groups created to be used in estimation models of the two provinces.

Selected groups	Kahramanmaras		Isparta	
	Feature selection function	Number of feature	Feature selection function	Number of feature
GR1	CfsSE	5	CfsSE	3
GR2	ClassAE	5	ClassAE	8
GR3	CSE-RAE	4	CSE-WSE	7
GR4	CorrAE	4	CorrAE	4
GR5	WSE	6	RAE	3

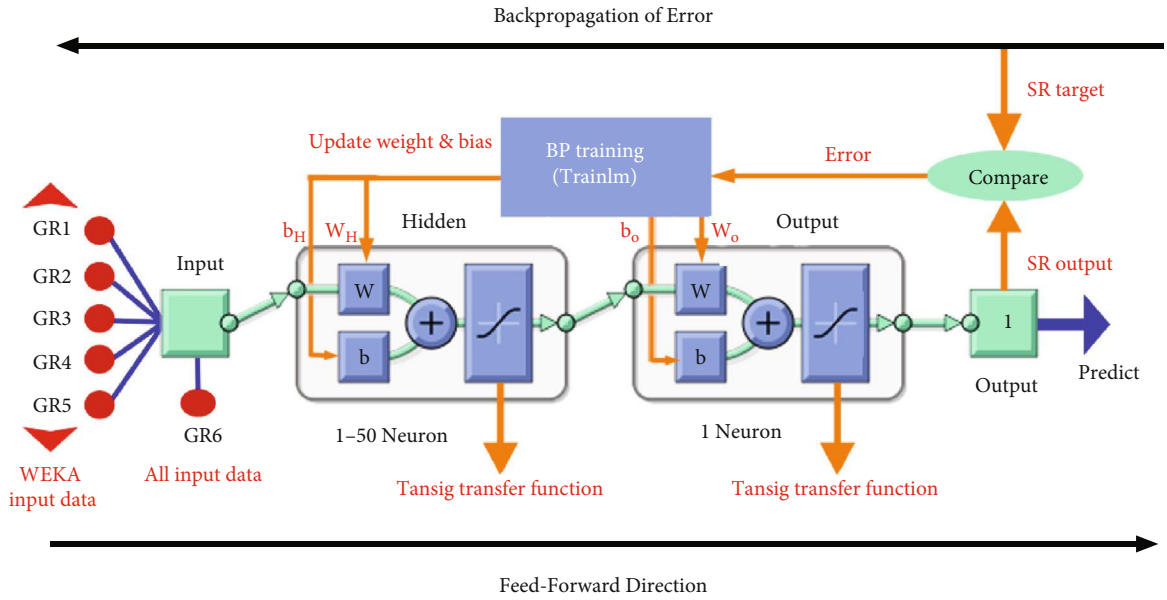


Figure 5: Architectural structure of MFFNN used in modelling.

estimation problems, the SVM-based SVR method was developed. SVR uses a technique based on regression problems and based on calculating a linear regression function in a multidimensional feature set [48]. The architectural structure of the SVR used in modelling studies is given in Figure 6.

The gaps between the data are kept wide in the SVR algorithm, ideal locations are found, and errors are minimized. In a dataset with a certain number of elements, $\{(x_a, y_a), a = 1, 2, 3, \cdots, M\}$ represent the input vector $x_a \epsilon R^d$, respectively, $y_a \epsilon R\, x_i$ represents the corresponding output vector, and M represents the total number of elements [49]. The formula of the SVR linear function is given in

$$y = f(x) = W^t \varphi(x) + b. \tag{4}$$

$\varphi(x)$ represents the nonlinear mapping function which converts multidimensional data structures into a two-dimensional chart, W represents the weight vector, and b represents bias. The error function is given in equation (5). The constant C and the ε values are determined by the user and are defined as the estimation accuracy of the training data.

$$R(x) = \frac{1}{2}\|W\|^2 + C\frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|_\varepsilon. \tag{5}$$

The equation that minimizes the error function is given in equation (6). α_a^* and α_a are the LaGrange multipliers and are referred to as the support vectors if the training vector has a value other than zero. This structure is known as the critical values for SVR algorithms [50]. The $K(x_a, x)$ structure is called the kernel function and converts the data it receives as input into an available form. Different types of kernel functions are used in SVR. Three different types of kernels, i.e., Polynomial (POL), Normalized Polynomial (NOR-P), and Gaussian Radial Basis Functions (RBF), were used in the models that were developed with SVR, and formulas for these functions are given in equations (7)–(9), respectively.

$$f(x) = \sum_{a=1}^{A}(\alpha_a^* - \alpha_a)K(x_a, x) + b, \tag{6}$$

$$K(x_a, x) = [(x_a x) + 1]^d, \tag{7}$$

$$K(x_a, x) = \frac{[(x_a x) + 1]^d}{\sqrt{(x_a)^2 (x)^2}}, \tag{8}$$

$$K(x_a, x) = \exp\left(-\frac{1}{2\sigma^2}\|x_a - x\|^2\right). \tag{9}$$

The classic SVR algorithm was also evaluated in the study by developing estimation models with LibSVM, which is another SVR-based method, and which is also an SVM-SVR-based algorithm software supporting single-class SVMs, two or multiclass SVMs, and SVRs [51]. LibSVM is preferred because it is a method that is used quite frequently in academic studies but not much preferred in SR prediction studies. Two different SVR types and kernels (Epsilon SVR (E-SVR) and Nu-SVR) were used in the estimation models that were developed with LibSVM, and RBF was used as the kernel. All prediction models were developed with Matlab R2017b software using LibSVM library interface software plugin.

2.5. KNN Algorithm. This algorithm is widely preferred in classification problems. However, a regression-based method was used in the present study. KNN is a nonparametric lazy

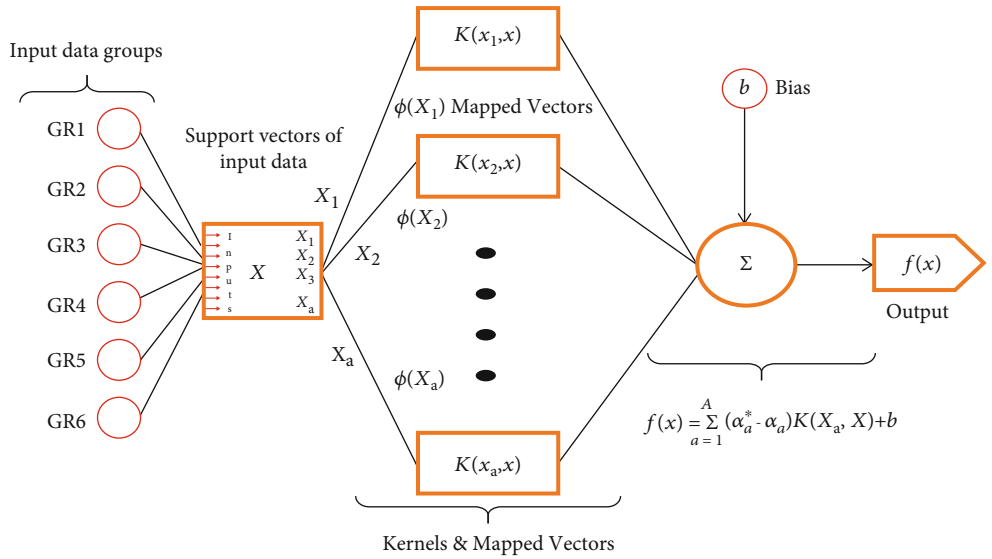


Figure 6: SVR's architectural structure.

ML-based learning algorithm and estimates by searching for the closest neighbors in the training dataset. KNN's nonparametric equation is given in equation (10), where each $N_k(x)$ was taken as the neighbor K of x data. In the formula, the y_i value represents the target output for each x_i training data.

$$f_k(x) = \frac{1}{K} \sum_{i \in N_k(x)} y_i. \tag{10}$$

Each new data intended to be estimated is looked at in the neighborhood of K from the previous data with a KNN. The distance between any data value and all values in the training dataset is calculated and then the nearest K training data values were determined. The average of the target output values is estimated for these values [52, 53]. The Euclid function was used for the calculation of the distance. The formula for the Euclid Function is given in equation (11). Care should be paid in choosing the K value; small values should be used since the model tends to overfitting if the selection is too high [54]. In the present study, the K value was taken as 1, 2, 3, 4, 6, and 10, and six different KNN models were developed for each data group. The model was deemed to over fit with a K value of more than 10. The linear nearest neighbor (LinearNN), which is a rough force-based search algorithm, was also used in the study. With this structure, the distance between each point pairs was found in the dataset.

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{K} (x_i - y_i)^2}. \tag{11}$$

2.6. Rule-Based M5 Algorithm. The M5 algorithm was developed by Quinlan as an advanced version of the Classification and Regression Tree (CART) [55], which is based on a binary decision tree structure developing a relation between dependent and independent variables of tree leaves creating a linear regression model on each leaf to estimate the value of the samples reaching the leaf. The algorithm is established on two structures, which are the decision tree and the linear regression. The best leaf is determined as the rule in the M5 algorithm, and pruning and dividing occur in two stages. In the dividing operation, the dataset at hand is divided into subsets to create a decision tree. It is also ensured in this process that numerical features are constantly estimated on each node by using a linear regression function in leaf nodes [56]. Standard deviation is used to find the error in the relevant node, and the error is seen to decrease here at the desired rate for each feature. The division ends if there is little change in the values of the samples that reach a node or if the number of samples decreases too much [57]. The Standard Deviation Reduction (SDR) formula is given in equation (12), where T is defined as the set of feature values reaching the node, T_i is the feature values taken from the divided node, and std is the standard deviation [57].

$$\text{SDR} = \text{std}(T) - \sum_i \frac{|T_i|}{|T|} \text{std}(T_i). \tag{12}$$

A rule-based type of the M5 algorithm was used in the present study. In this method, which is also known as M5 rules, a series of M5 trees are created where the best leaf (rule) is hidden, and the sample dataset with the best rule in each cycle is removed from the training dataset without creating the next tree. While the M5 algorithm creates one single decision tree, M5 rules create a complete tree in each cycle. M5 rules develop a series of rules based on the M5 algorithm by using the Partial and Regression Tree (PART) algorithm [58].

2.7. Performance Analysis of ML Algorithm Models. The widely used statistical error measurement and analysis methods were employed in evaluating the performance of the models that were developed with ML algorithms in predicting SR output both themselves and among each other. The Mean Square Error (MSE), Root Mean Square Error

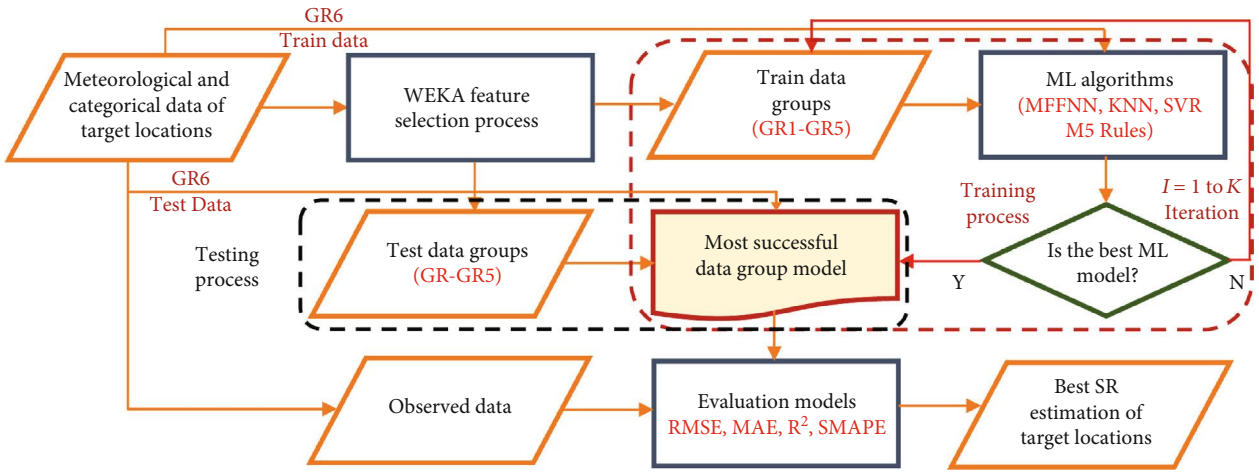
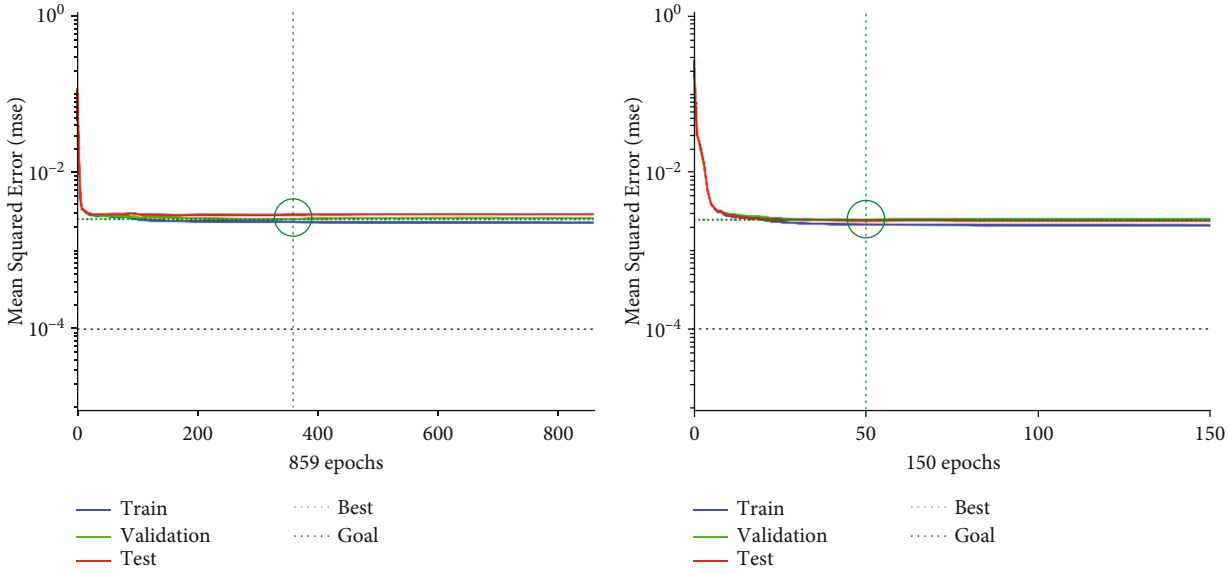


Figure 7: Flowchart of ML-based HAGSR estimation processes.



(a) Best Validation Performance is 0.002638 at epoch 359

(b) Best Validation Performance is 0.0026333 at epoch 50

Figure 8: The MFFNN performance plots of best models: (a) Isparta and (b) Kahramanmaras.

(RMSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (SMAPE) are the error measurement statistics used in the study. Two different statistical analysis methods, Correlation Coefficient (R) and Coefficient of Determination (R^2), were also used. The formulas used for the statistical scales are given in equations (13)–(18), respectively. In the formulas, O_i, P_i, $\bar{O}$, and $\bar{P}$ are the measured, estimated, and measurement and estimation averages, respectively.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2, \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}, \tag{14}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i|, \tag{15}$$

$$\mathrm{SMAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{|O_i - P_i|}{(|O_i| + |P_i|) * 0.5}, \tag{16}$$

$$R = \frac{\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}}, \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}. \tag{18}$$

The percentage errors are used widely to compare the estimation performance of various datasets. MAPE, which is an estimation error calculation method independent from

Table 8: Training and test estimation results of the most successful models that were developed by using the MFFNN algorithm according to input data groups of both provinces.

	Data	Hidden Layer	Hidden Neurons	Training models: Training data MSE	Training models: Test data MSE	Training models: Validation data MSE	Training models: Model MSE	Training models: Model R	Testing models: RMSE	Testing models: MAE	Testing models: SMAPE (%)	Testing models: R^2
Isparta	GR1	1	18	0.0198	0.0202	0.0202	0.0201	0.7970	0.1392	0.1109	15.27	0.6547
	GR2		28	0.0029	0.0031	0.0034	0.0030	0.9724	0.0559	0.0372	8.36	0.9444
	GR3		**48**	**0.0024**	**0.0029**	**0.0026**	**0.0025**	**0.9768**	**0.0536**	**0.035**	**7.77**	**0.9488**
	GR4		32	0.0212	0.0217	0.0216	0.0216	0.7803	0.1446	0.1124	15.29	0.6275
	GR5		21	0.0059	0.0057	0.0063	0.0059	0.9442	0.0754	0.0562	8.97	0.8988
	GR-6		25	0.0027	0.0030	0.0029	0.0028	0.9745	0.0542	0.0363	8.39	0.9477
Kahramanmaras	GR1	1	30	0.0134	0.0137	0.0141	0.0137	0.9049	0.1181	0.0897	14.25	0.8138
	GR2		**40**	**0.0021**	**0.0026**	**0.0026**	**0.002**	**0.9845**	**0.0508**	**0.034**	**7.79**	**0.9656**
	GR3		27	0.0029	0.0031	0.0028	0.0030	0.9800	0.0556	0.0385	8.23	0.9587
	GR4		25	0.0145	0.0145	0.0150	0.0149	0.8959	0.1234	0.0928	14.37	0.7969
	GR5		44	0.0027	0.0030	0.0033	0.0028	0.9812	0.0558	0.0386	8.29	0.9585
	GR-6		39	0.0028	0.0028	0.0029	0.0029	0.9805	0.0551	0.0383	8.11	0.9595

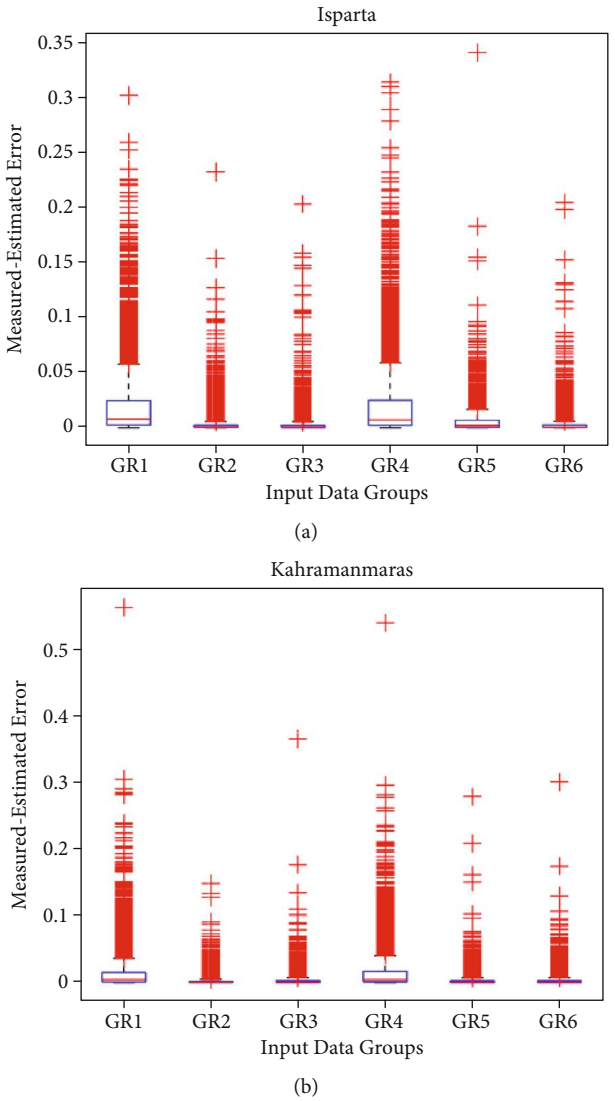


Figure 9: Statistical error box plots between the measured and estimated values of (a) Isparta and (b) Kahramanmaras according to the input data.

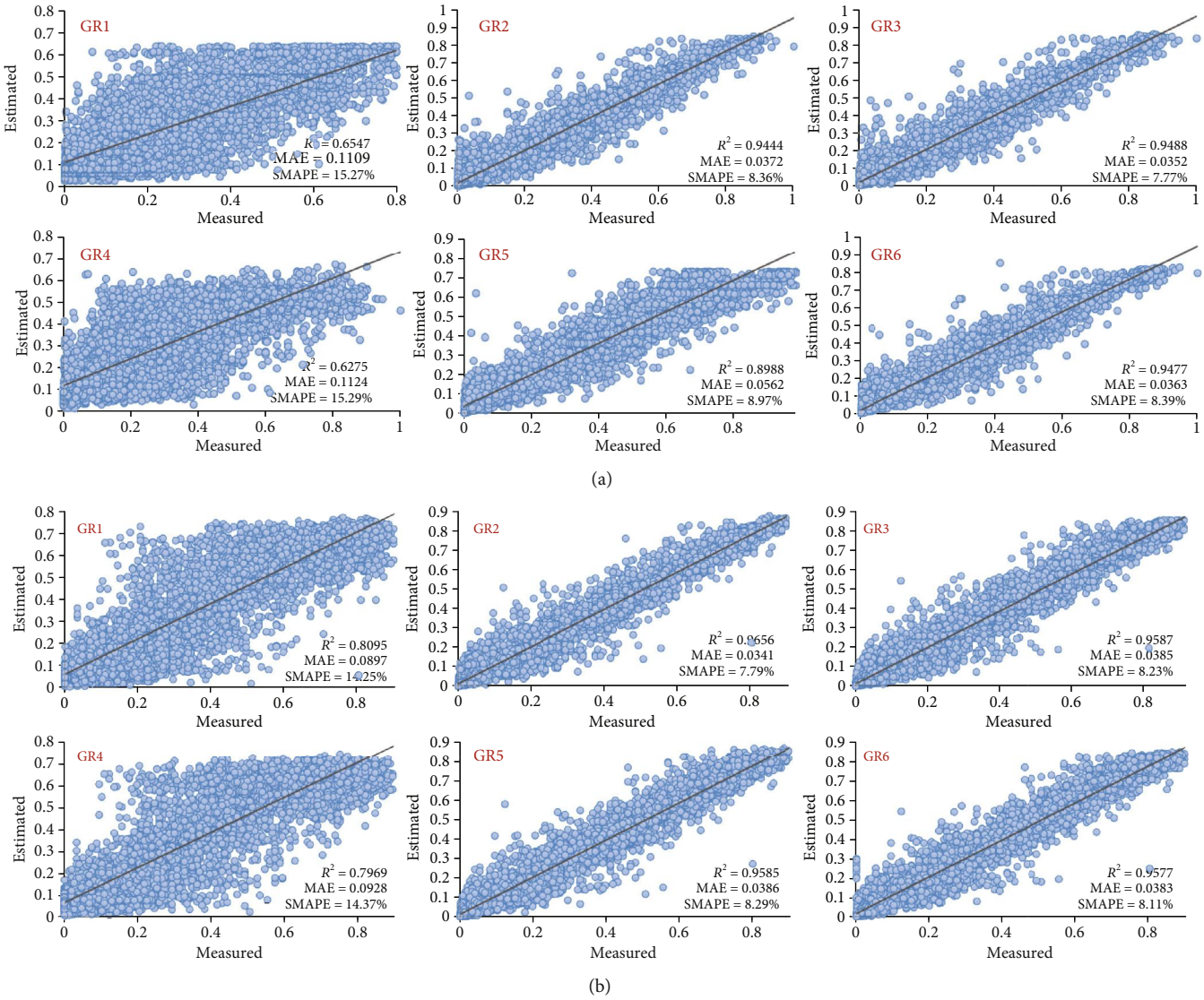


Figure 10: SR estimation scatter plots of MFFNN models for (a) Isparta and (b) Kahramanmaras according to test input data.

the scale, gives an incorrect result when measurement and estimation values are zero or have a value quite close to zero [59]. The SMAPE percentage scale was used to overcome this problem since the measurement and estimation results had values that were zero or quite close to zero.

3. Results and Discussion

Although input data is used in SR estimation studies in many areas and location, it is not common to evaluate SR on the same latitude and at locations that have similar geographical characteristics. Based on the effect of latitude on sunshine duration and the angle of coming solar rays, Darhmaoui and Lahjouji [60] calculated that the annual solar radiation values were at similar levels at the same latitude points of a geographical area, with a strong relationship between optimum tilt angle and target latitude value. Ahlgren et al. [61] emphasized the relationship between annual yield and latitude because there was a directly proportional relationship between latitude and direct normal radiation where parabolic groove collectors were located. For this reason, places that had the same latitude coordinates were selected on the target area, and ML algorithms were employed for high-accuracy GSR estimation. The estimated results of the data groups were compared by using statistical error measurement and analysis methods including SMAPE, MAE, RMSE, and R^2 to evaluate the training and testing estimation performance of the developed models. The closer the value between the measured and estimated in statistical error measurement methods is to 0 and the closer to 1 in analysis methods, the estimation accuracy of the developed models is higher [40]. The flowchart of the HAGSR estimation processes of both provinces is given in Figure 7.

Different features were used for each data group to estimate SR with the MFFNN algorithm by employing data groups in the GR1-GR6 range. During the training process of the models, a five-iteration structure was created for each hidden layer neurons between 1 and 50, and 250

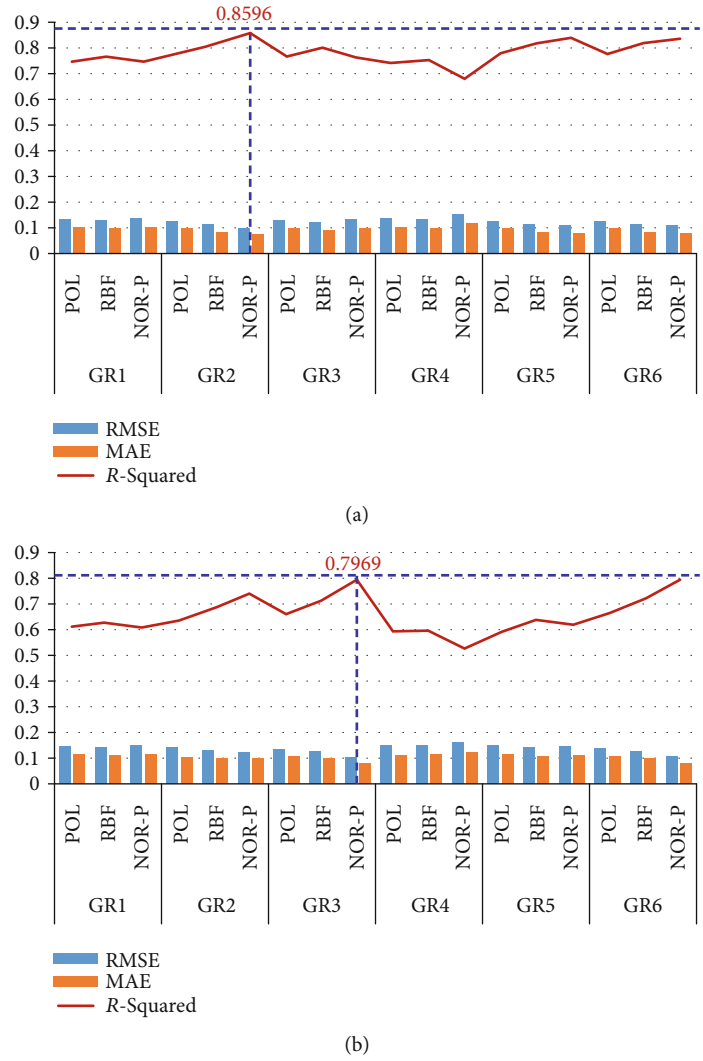


Figure 11: Estimation performance of the models developed with SVR by using 3 kernel functions for (a) Kahramanmaras and (b) Isparta.

different models were developed for each input group, improving 1500 different models in total. In the range of 0 to 1000 epochs, the network performance plots of the models that reached the best estimation results in the training process of both provinces are shown in Figure 8. When the training, validation, and testing SR estimation of each neural network model that was developed was evaluated statistically, the following results of the most successful MFFNN estimation models were determined and are given in Table 8.

As seen in Table 8, the most successful estimation models were calculated by using GR3 for Isparta and GR2 input data for Kahramanmaras. Although a 48-hidden layer neuron was found as the most successful MFFNN model in the first iteration in Isparta, a 40-hidden layer neuron was the most successful estimation model in the second iteration in Kahramanmaras. The training performance of the most successful models that were developed for Isparta and Kahramanmaras was found to be MSE = **0.0025** and **0.0023** and R = **0.9768** and **0.9845**, respectively. When the best estimation values were compared with the actual values measured by using the test data, the R^2, MAE, and SMAPE values for Isparta were **0.9488**, **0.0352**, and **7.77%**, respectively; and these values were **0.9656**, **0.0341**, and **7.79%**, respectively, for Kahramanmaras. The estimation performance of the two target areas was evaluated with different scales, and both the training and test data reached very similar results.

Boxplots between the measured and estimated values of the study done on the selected provinces are given in Figure 9. In these plots, the statistical error average measurement results between the test input data and the estimated values of each province can be seen. Scatter plots between measured and estimated values of the most successful model developed for each data group are given in Figure 10. It is understood in both plots that a high level of correlation was achieved for GR2, GR3, and GR6 data groups in Isparta, and a similarly high-level relationship was reached for GR2, GR3, GR5, and GR6 data groups in Kahramanmaras.

Another ML algorithm that is employed in estimating HAGSR is SVR. The results calculated with SVR-based estimation models were found to be quite low. Therefore, it

Table 9: Estimation results of the most successful models that were developed according to LibSVM and the input data of both provinces.

Target provinces		Kahramanmaras				Isparta			
Data groups	SVM type	RMSE	MAE	SMAPE (%)	R^2	RMSE	MAE	SMAPE (%)	R^2
GR1	E-SVR	0.1271	0.1003	15.11	0.7876	0.1431	0.1153	15.46	0.6372
	Nu-SVR	0.1257	0.0960	14.97	0.7893	0.1428	0.1144	15.44	0.6377
GR2	E-SVR	0.0742	0.0584	13.43	0.9278	0.0816	0.0637	12.64	0.8818
	Nu-SVR	**0.0675**	**0.0501**	**12.14**	**0.9394**	0.0765	0.0580	12.05	0.8961
GR3	E-SVR	0.0765	0.0604	13.69	0.9229	0.0815	0.0643	12.95	0.8821
	Nu-SVR	0.0697	0.0523	12.42	0.9352	**0.0752**	**0.0573**	**12.11**	**0.8995**
GR4	E-SVR	0.1300	0.1026	15.17	0.7781	0.1495	0.1193	15.66	0.6050
	Nu-SVR	0.1287	0.0984	14.97	0.7792	0.1491	0.1185	15.64	0.6057
GR5	E-SVR	0.0780	0.0616	13.77	0.9197	0.0849	0.0673	12.28	0.8728
	Nu-SVR	0.0717	0.0540	12.59	0.9315	0.0827	0.0621	11.77	0.8792
GR6	E-SVR	0.0802	0.0631	13.80	0.9146	0.0827	0.0645	12.52	0.8790
	Nu-SVR	0.0756	0.0576	12.96	0.9237	0.0779	0.0594	12.15	0.8925

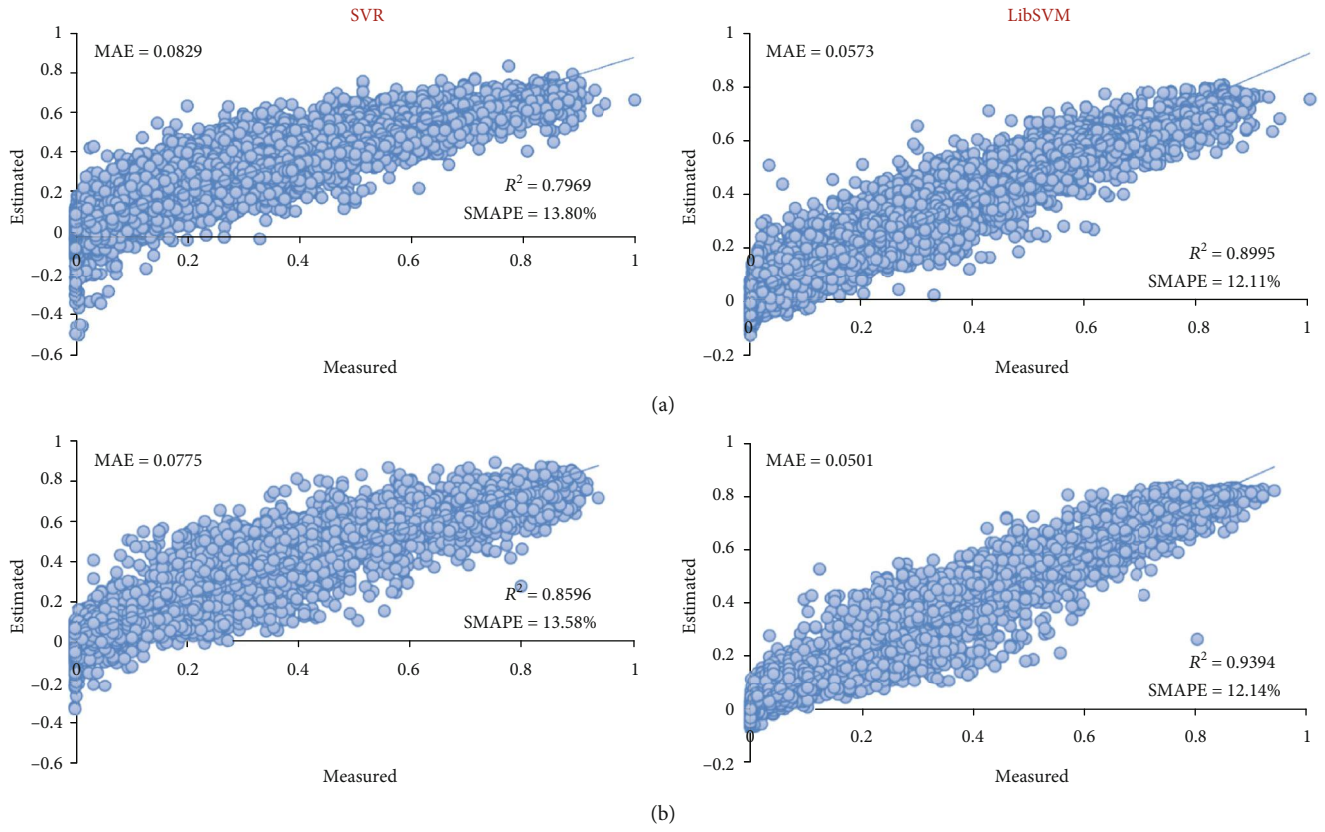


Figure 12: Scatter plots of the most successful models that were developed according to SVR and LibSVM for (a) Isparta and for (b) Kahramanmaras.

was decided that the classic SVR estimation results should be evaluated with LibSVM, which is another SVR-based method. LibSVM was preferred because it is a well-known method in academic literature. In both methods, the most suitable combinations were determined by creating numerous models and the most successful estimation models were developed in the selection of user-defined C (complexity and cost parameter), epsilon (error parameter), and Nu (parameter used instead of C). The performance results of 18 different models that were developed for each province by using the POL, NOR-P, and RBF core functions with SVR are shown in Figure 11. The estimates were obtained between the **0.6786** and **0.8596** range for the province of Kahramanmaras according to the R^2 scale and **0.5273-0.7969** for Isparta. A total of 12 different estimation models were developed with LibSVM for the data groups in each

Table 10: Estimation results of the two most successful models that were developed by using a KNN algorithm according to input data groups of both provinces.

Target provinces		Kahramanmaras				Isparta			
Data groups	The number of neighbors (K)	RMSE	MAE	SMAPE (%)	R^2	RMSE	MAE	SMAPE (%)	R^2
GR1	6	0.1221	0.0903	14.25	0.8022	0.1447	0.1140	15.53	0.6274
	10	0.1205	0.0896	14.20	0.8069	0.1434	0.1131	15.42	0.6338
GR2	4	0.0618	0.0427	8.91	0.9490	0.0692	0.0481	9.34	0.9151
	6	**0.0605**	**0.0419**	**8.84**	**0.9511**	0.0693	0.0485	9.38	0.9150
GR3	4	0.0631	0.0437	9.08	0.9470	**0.0646**	**0.0438**	**8.88**	**0.9261**
	6	0.0631	0.0439	9.13	0.9469	0.0648	0.0450	9.17	0.9258
GR4	6	0.1288	0.0956	14.44	0.7798	0.1506	0.1157	15.42	0.5980
	10	0.1257	0.0937	14.30	0.7898	0.1463	0.1134	15.26	0.6194
GR5	4	0.0645	0.0446	9.37	0.9445	—	—	—	—
	6	0.0649	0.0448	9.32	0.9439	0.0820	0.0614	9.33	0.8805
	10	—	—	—	—	0.0811	0.0610	9.32	0.8829
GR6	3	0.0718	0.0487	10.34	0.9318	—	—	—	—
	4	0.0708	0.0482	10.32	0.9335	0.0728	0.0514	9.79	0.9065
	6	—	—	—	—	0.0730	0.0522	10.07	0.9062

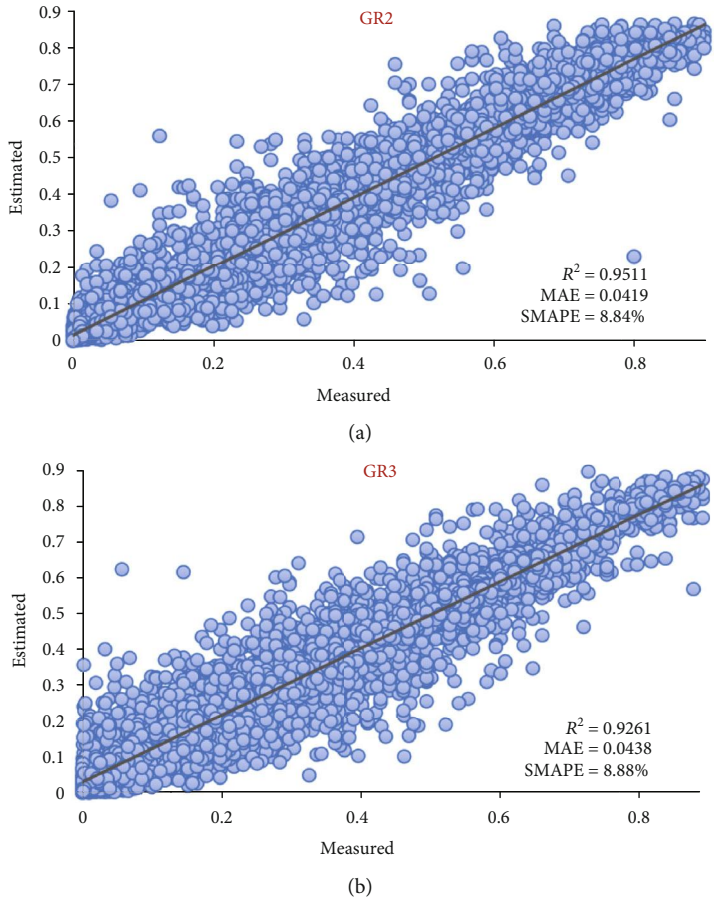


Figure 13: Scatter plots of a KNN-based most successful models according to the (a) GR2 input data for Kahramanmaras and (b) GR3 for Isparta.

province. The statistical results of SR estimation models that were developed by using the RBF kernel function for two different regression-based SVR algorithms are given in Table 9. The models that were developed with Nu-SVR were more successful than E-SVR. The model that was developed with the GR2 data group had the most successful estimation

Table 11: Estimation results of the two most successful models that were developed by using the M5 rules algorithm according to input data groups of both provinces.

Data groups	Target provinces							
	Kahramanmaras				Isparta			
	RMSE	MAE	SMAPE (%)	R^2	RMSE	MAE	SMAPE (%)	R^2
GR1	0.1216	0.0915	14.37	0.8026	0.1400	0.1114	15.46	0.6514
GR2	0.0625	0.0431	9.07	0.9479	0.0666	0.0447	8.77	0.9218
GR3	0.0625	0.0430	9.09	0.9479	**0.0650**	**0.0441**	**8.42**	**0.9254**
GR4	0.1254	0.0944	14.56	0.7903	0.1454	0.1140	15.44	0.6236
GR5	**0.0610**	**0.0418**	**9.01**	**0.9506**	0.0814	0.0593	9.37	0.8822
GR6	0.0614	0.0420	9.08	0.9500	0.0655	0.0445	8.43	0.9245

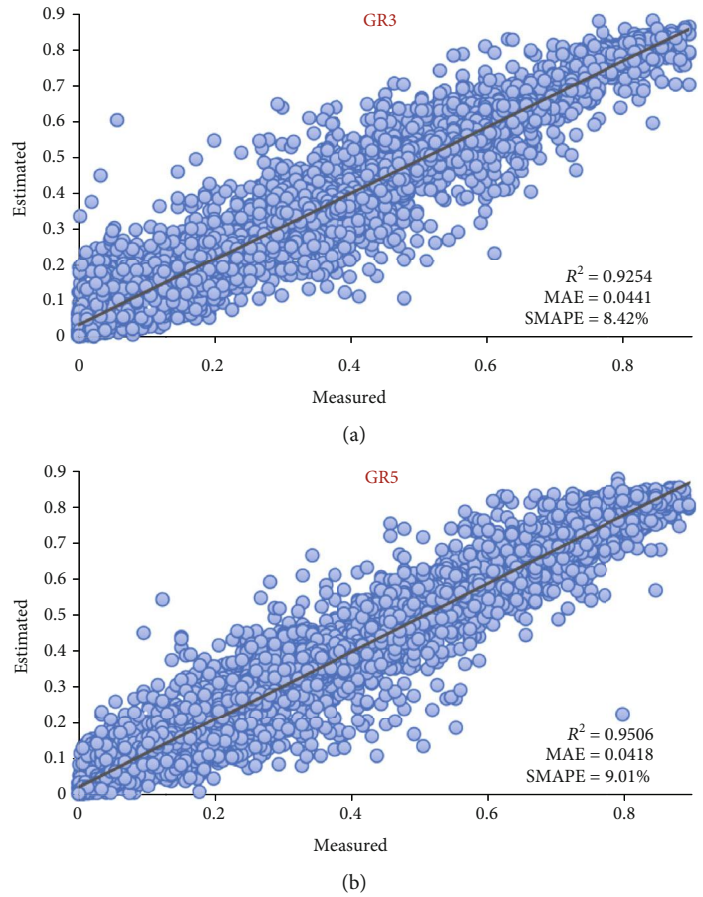


Figure 14: Scatter plots of the M5 rules-based most successful models according to (a) GR3 input data for Isparta and (b) GR5 for Kahramanmaras.

performance with **0.0675**, **0.0501**, **12.14%**, and **0.9394**, respectively, according to the RMSE, MAE, SMAPE, and R^2 scales in the Kahramanmaras target area. Similarly, the model that was developed by using the GR3 data group in Isparta was successful with **0.0752**, **0.0573**, **12.11%**, and **0.8995**, respectively. Comparative scatter plots of the most successful models developed with two different SVR methods used in the study according to the selected provinces are given in Figure 12.

As understood in Figure 12, the best SR estimation results of the models that were developed with LibSVM from both similar methods were found to be more successful than the classic SVR. For this reason, it was decided to use the estimation results of LibSVM in comparative evaluation of ML algorithms.

A total of 36 different estimation models were developed by using selected input data for each province based on six different K-neighbor coefficients between 1 and 10 with the KNN ML algorithm. The estimation performance results of the two most successful models that were developed in each data group with user-defined K parameters are given in Table 10. It was determined that the K

Table 12: Comparison of the most successful HAGSR estimation models that were developed by using MFFNNN, KNN, M5 rules, and LibSVM for both provinces.

Statistical indicators	Target provinces							
	Kahramanmaras				Isparta			
	MFFNN	KNN	M5 rules	LibSVM	MFFNN	KNN	M5 rules	LibSVM
RMSE	**0.0508**	0.0605	0.0610	0.0675	**0.0536**	0.0646	0.0650	0.0752
MAE	**0.0341**	0.0419	0.0418	0.0501	**0.0352**	0.0438	0.0441	0.0573
R^2	**0.9656**	0.9511	0.9506	0.9394	**0.9488**	0.9261	0.9254	0.8995
SMAPE (%)	**7.79**	8.84	9.01	12.14	**7.77**	8.88	8.42	12.11
Data groups	**GR2**	GR2	GR5	GR2	**GR3**	GR3	GR3	GR3

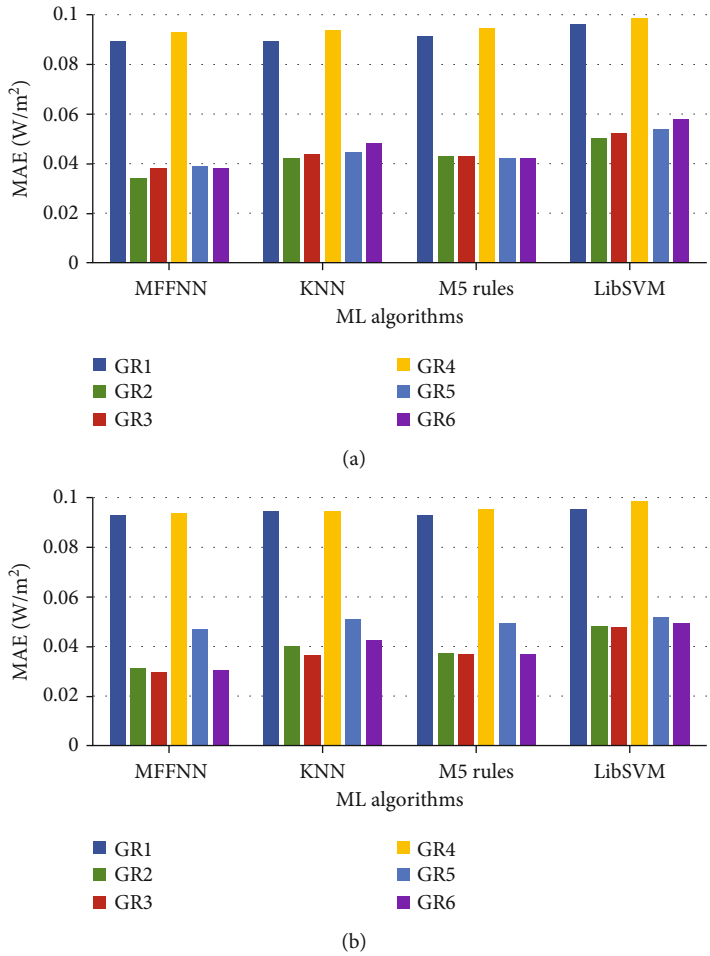


Figure 15: MAE performance comparisons of the best models developed by using ML algorithms according to data groups for (a) Kahramanmaras and (b) Isparta.

parameter was a defining feature in the estimation models, but there was not always a correct proportion towards an increase. No significant performance increases were detected in all models developed with over 10 K parameters, and modelling time was extended. In the relevant table, the most successful model that was developed for Kahramanmaras estimated SR with **0.0605**, **0.0419**, **0.9511**, and **8.84%**, respectively, with the GR2 data group according to the RMSE, MAE, R^2, and SMAPE scales. For Isparta, similarly, it was estimated with the GR3 data group resulting in **0.0646**, **0.0438**, **0.9261**, and **8.88%**, respectively. The scatter plots of estimation results are given in Figure 13. It is seen in the SMAPE scale that the SR estimations of the provinces used in the study are very close and similar.

Six different rule-based estimation models were developed for each province by using selected data groups of the targeted cities with the M5 rules algorithm. The estimation performance of the developed models is given in Table 11. The best model developed for Kahramanmaras was

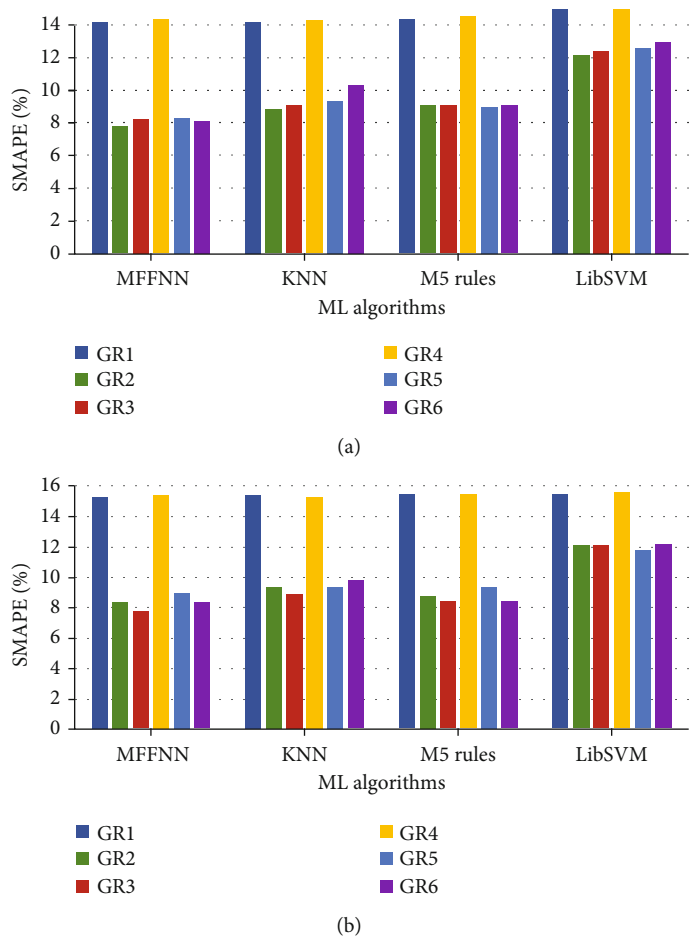


Figure 16: SMAPE performance comparisons for the best models developed by using ML algorithms according to data groups for (a) Kahramanmaras and (b) Isparta.

estimated by using the GR5 data group, which is unlike other ML algorithms employed in the study. Isparta, on the other hand, was estimated similarly by using the GR3 data group. According to the RMSE, MAE, R^2, and SMAPE statistical scales, the values of **0.0610**, **0.0418**, **0.9506**, and **9.01%**, respectively, were obtained in the performance of the best model for Kahramanmaras. Similarly, **0.650**, **0.0441**, **0.9254**, and **8.42%** were obtained for the province of Isparta. Scatter plots of the most successful models that were developed in the target cities are given in Figure 14. According to the plots, it is understood that the data distributions and performance measurement metrics of the target provinces were very close to each other.

Aside from the trial studies in all ML algorithms used in the target provinces to increase estimation accuracy and to select the most successful models in each data group, 3000, 72, 12, and 24 different estimation models were developed with MFFNN, KNN, M5 rules, and SVR algorithm-based LibSVM library, respectively. In all studies, the estimated performance of the models that were developed with the GR2 (month, hour, T, H, and HSD) and the GR5 (month, hour, T, H, WS, and HSD) data groups determined at the end of the feature selection process in Kahramanmaras was more successful. In Isparta however, the models that were developed with the GR3 (year, month, day, hour, P, T, and HSD) data group showed higher performance. The statistical comparisons of the best performing models according to ML algorithms used in SR estimations of both provinces are given in Table 12. Based on the statistical scales that were employed in the study, the MFFNN algorithm estimated SR more accurately in both provinces than the other algorithms. However, similar estimation results were achieved with the KNN and M5 rules algorithm for each province, and the lowest performance values were detected in SVR models that were developed with LibSVM. With the MFFNN algorithm, the SR estimation results achieved in Kahramanmaras and Isparta according to SMAPE were **7.79%** and **7.77%**, respectively; **8.84%** and **8.88%**, respectively, with the KNN algorithm; and **12.14%** and **12.11%,** respectively, with the LibSVM algorithm. According to SMAPE, the fact that the SR estimation results of both provinces selected in the study are very close to each other a level is associated with the similarity of latitude and some geographical characteristics.

In the HAGSR estimation studies, the final performance results of the estimation models that were developed with the GR6 data group by using all the available input data were lower than the final performance results of models that were developed with the GR2, GR3, and GR5 data groups, which

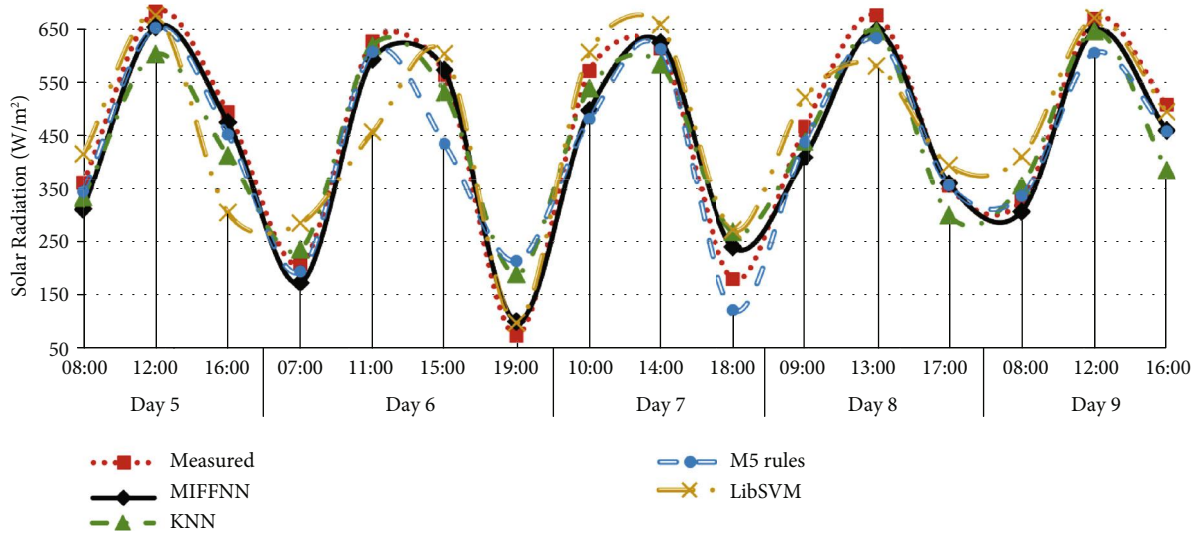


Figure 17: Comparative performance plot of ML algorithm models measured and estimated by using five-day input data from July 2004 in Isparta according to HAGSR results.

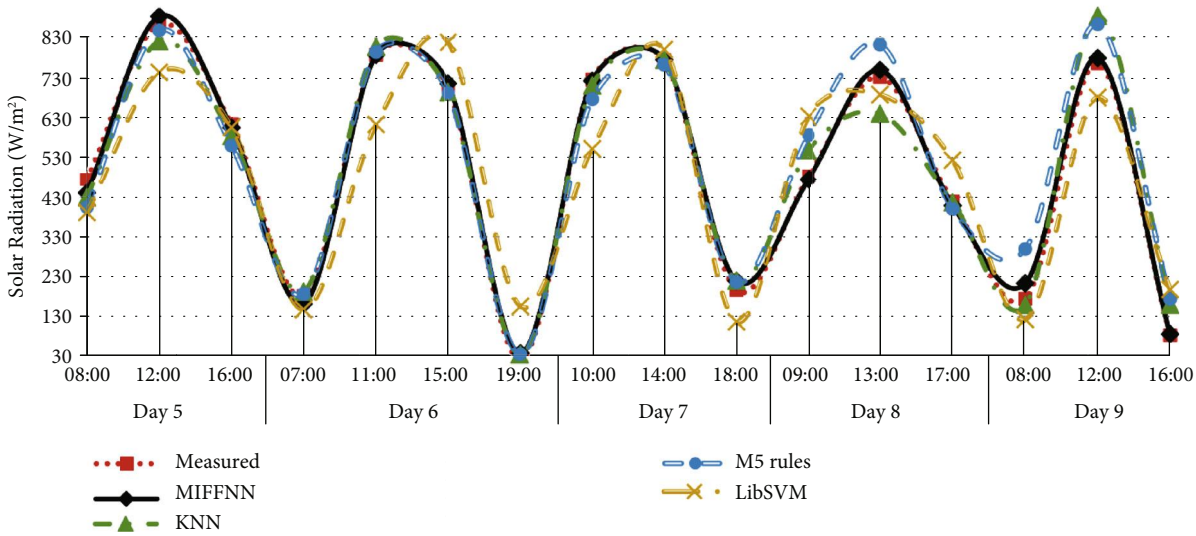


Figure 18: Comparative performance plot of ML algorithm models measured and estimated by using five-day input data from July 2004 in Kahramanmaras according to HAGSR results.

were groups created at the end of the feature selection process. MAE and SMAPE performance plots according to four ML algorithms of all data groups used for the target provinces are given in Figures 15 and 16. It is clearly seen that feature selection processes have positive contributions to the performance of the developed estimation models.

The HAGSR estimation models that were developed for Kahramanmaras and Isparta estimated the solar energy source of target areas quite well in general. However, the studies with the GR1 and GR4 data groups represent the input data groups that have the lowest estimated performance in both provinces. As a result, it was concluded that CfsSE and CorrAE, which are among the feature selection functions, applied to the meteorological and categorical input datasets, were inadequate in determining the best input data. The most successful feature selection functions were ClassAE and WSE for Kahramanmaras and CSE and WSE for Isparta. The comparisons of the SR estimations and real measurement results of the best models that were developed with the four ML algorithm using the 7 input data that were determined with the CSE and WSE feature selection functions for Isparta are given in Figure 17. Similarly, the comparisons of the best models that were developed with 5 inputs for the ClassAE feature selection function and 6 inputs for the WSE feature selection function for Kahramanmaras are given in Figure 18. The five-day hourly input data that were selected randomly from the test data for July 5-9 in 2004 were used for comparisons. It is seen that the HAGSR estimation

Table 13: Comparison of the performance of the present study with various HAGSR estimation models of the previous studies.

ML methods	Best methods	Time range	Location	Author [Ref.]	Statistical indicators		
					RMSE	MAE	R^2
ANN	ANN	2001–2007	La Serena (Chile)	Lazzús et al. [18]	—	—	0.9437
ANN	ANN	02 February–31 May 2011	Algeria	Hasni et al. [17]	0.172	2.9971	0.9999
MLP	MLP	2009–2012	Fez (Morocco)	Ihya et al. [62]	—	—	0.8896
MLP-NARX	NARX	2010–2014	Fes (Morocco)	Loutfi et al. [32]	—	—	0.95
AdaBoost, LR, KNN, CART, SVR, ANN, RD regression	RD regression	2013–2015	South Korea	Kim et al. [63]	577.5	—	0.705
ANN	ANN	2006–2010	Ajaccio (Corsica)	Notton et al. [12]	12.43	19.17	0.998
MARS-ANN-LR	ANN	2010–2016	Hong Kong	Li et al. [28]	0.270	0.194	0.918
MFFNN-KNN-M5 rules-LibSVM	MFFNN	2002–2006	Kahramanmaras-Isparta (Turkey)	Present study	0.0341	0.0508	0.9656

models that were developed for Kahramanmaras are slightly more successful in estimating SR compared to Isparta where the test data time zones were selected randomly for each day.

The comparison of the HAGSR estimation models that were developed by using ML algorithms in the literature, and the most successful model developed in this study, is given in Table 13. The most successful models that were developed in previous studies were commonly based on a neural network, as in this study. It is understood that the accuracy of the proposed estimation model is better than, or similar to, previous studies.

4. Conclusion

In the present study, a comparative evaluation was made by developing models based on four different ML (MFFNN, KNN, SVR-based LibSVM, and M5 rules) algorithms to predict the HAGSR of the provinces of Kahramanmaras and Isparta, which are located on the same latitude coordinates of the Mediterranean Region. The most suitable input features were determined for each feature selection function by using meteorological and categorical input data and by developing 29 different applications based on six different feature selection functions with WEKA, and the input data were created in five different selection groups (GR1-GR5). Six different input datasets were determined to be used in modelling by including the GR6 data group in which all input data were collected to this selection group. The most successful estimation models were developed with the MFFNN algorithm in Kahramanmaras and Isparta by using the GR2 and GR3 data groups, respectively. Although month, hour, T, H, and HSD data were the most effective features in Kahramanmaras on estimation models, the variables of year, month, day, hour, P, T, and HSD were the most effective in Isparta. It is clear that HSD is the most effective data on SR in all data groups selected. The results show that the predictive accuracy of models that were developed with the data groups created at the end of the selection process increased, modelling time decreased, and the model is easier to interpret.

According to the data groups, the performance of KNN and M5 rules models was quite similar in each province. The performance of the estimation model that was developed with the KNN algorithm for the GR2 data group in Kahramanmaras was $R^2 = \mathbf{0.9511}$, and $R^2 = \mathbf{0.9506}$ for the M5 rules algorithm. In Isparta, the performance of the estimation model that was developed with the KNN algorithm for the GR3 data group was $R^2 = \mathbf{0.9261}$, and $R^2 = \mathbf{0.9254}$ for the M5 rules algorithm. The lowest performances were received for the GR1 and GR4 data groups in each province.

The best SR estimation performance of the two provinces was achieved with the MFFNN algorithm. When the results were evaluated in statistical terms, very close values were obtained in Kahramanmaras and Isparta. The MAE of the most successful model that was developed in Kahramanmaras for the MFFNN algorithm was found to be **0.0341** and **0.0352** for Isparta. Similarly, the SMAPE of the most successful model that was developed in Kahramanmaras was found to be **7.79%** and **7.77%** in Isparta. Although the statistical evaluation result of the different ML algorithms used in the study was low, similar results were obtained. These results show us that these two cities, which are very far from each other, have similar SR estimation potentials and that the latitude or different geographical characteristics have significant effects on these similarities. As a result of the present study, the HAGSR potential of both cities was estimated successfully and performed better than any other studies conducted in this field. In future studies, different parts of Turkey and the world should be evaluated in terms of performance of various ML algorithms and time intervals.

Data Availability

The data used to support the findings of this study are available from the corresponding author or Turkey General Directorate of Meteorology Meteorological Data Information Sales and Presentation System (MEVBIS) website upon request; website address: https://mevbis.mgm.gov.tr/mevbis/ui/index.html#/Workspace.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

The authors are grateful to the MEVBIS staff for his assistance during the research.

References

[1] B. Yaniktepe, O. Kara, and C. Ozalp, "The global solar radiation estimation and analysis of solar energy: case study for Osmaniye, Turkey," *International Journal of Green Energy.*, vol. 14, no. 9, pp. 765–773, 2017.

[2] "Renewable energy policy network for the 21st Century (Ren21). Renewables 2018 global status report," March 2020, https://ren21.net/gsr-2018/.

[3] B. Yaniktepe, O. Kara, and C. Ozalp, "Technoeconomic evaluation for an installed small-scale photovoltaic power plant," *International Journal of Photoenergy*, vol. 2017, 7 pages, 2017.

[4] REN21, "Renewables 2019 Global Status Report. Renewables Now, Paris: REN21 Secretariat," 2019, May 2020, https://www.ren21.net/wp-content/uploads/2019/05/gsr_2019_full_report_en.pdf.

[5] O. Kisi, "Modeling solar radiation of Mediterranean region in Turkey by using fuzzy genetic approach," *Energy*, vol. 64, pp. 429–436, 2014.

[6] B. Jahani and B. Mohammadi, "A comparison between the application of empirical and ANN methods for estimation of daily global solar radiation in Iran," *Theoretical and Applied Climatology.*, vol. 137, no. 1-2, pp. 1257–1269, 2019.

[7] A. Teke, H. B. Yıldırım, and Ö. Çelik, "Evaluation and performance comparison of different models for the estimation of solar radiation," *Renewable and Sustainable Energy Reviews.*, vol. 50, pp. 1097–1107, 2015.

[8] J. Almorox, C. Hontoria, and M. Benito, "Models for obtaining daily global solar radiation with measured air temperature data in Madrid (Spain)," *Applied Energy.*, vol. 88, no. 5, pp. 1703–1709, 2011.

[9] C. Voyant, G. Notton, S. Kalogirou et al., "Machine learning methods for solar radiation forecasting: a review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

[10] T. Khatib, A. Mohamed, and K. Sopian, "A review of solar energy modeling techniques," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 2864–2869, 2012.

[11] T. Muneer, E. J. Gago, and S. Etxebarria, "Monthly-averaged hourly solar diffuse radiation models for world-wide locations," *Future Cities and Environment.*, vol. 1, 2017.

[12] G. Notton, C. Paoli, S. Vasileva, M. L. Nivet, J.-L. Canaletti, and C. Cristofari, "Estimation of hourly global solar irradiation on tilted planes from horizontal one using artificial neural networks," *Energy*, vol. 39, no. 1, pp. 166–179, 2012.

[13] J. Zhang, L. Zhao, S. Deng, W. Xu, and Y. Zhang, "A critical review of the models used to estimate solar radiation," *Renewable and Sustainable Energy Reviews.*, vol. 70, pp. 314–329, 2017.

[14] M. Mohandes, S. Rehman, and T. O. Halawani, "Estimation of global solar radiation using artificial neural networks," *Renewable Energy*, vol. 14, no. 1-4, pp. 179–184, 1998.

[15] A. Moosa, H. Sahbir, H. Ali, R. DArwade, and B. Gite, "Predicting solar radiation using machine learning techniques," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1693–1699, Madurai, India, 2018.

[16] F. Baser and H. Demirhan, "A fuzzy regression with support vector machine approach to the estimation of horizontal global solar radiation," *Energy*, vol. 123, pp. 229–240, 2017.

[17] A. Hasni, A. Sehli, B. Draoui, A. Bassou, and B. Amieur, "Estimating global solar radiation using artificial neural network and climate data in the south-western region of Algeria," *Energy Procedia*, vol. 18, pp. 531–537, 2012.

[18] J. A. Lazzús, A. A. Pérez Ponce, and J. Marín, "Estimation of global solar radiation over the city of La Serena (Chile) using a neural network," *Applied Solar Energy*, vol. 47, no. 1, pp. 66–73, 2011.

[19] M. Benghanem, A. Mellit, and S. N. Alamri, "ANN-based modelling and estimation of daily global solar radiation data: a case study," *Energy Conversion and Management*, vol. 50, no. 7, pp. 1644–1655, 2009.

[20] K. Chiteka and C. C. Enweremadu, "Prediction of global horizontal solar irradiance in Zimbabwe using artificial neural networks," *Journal of Cleaner Production*, vol. 135, pp. 701–711, 2016.

[21] J. Piri, S. Shamshirband, D. Petković, C. W. Tong, and M. H. ur Rehman, "Prediction of the solar radiation on the Earth using support vector regression technique," *Infrared Physics & Technology*, vol. 68, pp. 179–185, 2015.

[22] H. B. Yıldırım, Ö. Çelik, A. Teke, and B. Barutçu, "Estimating daily global solar radiation with graphical user interface in eastern Mediterranean region of Turkey," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1528–1537, 2018.

[23] A. Sözen, E. Arcaklıoğlu, M. Özalp, and N. Çağlar, "Forecasting based on neural network approach of solar potential in Turkey," *Renewable Energy*, vol. 30, no. 7, pp. 1075–1090, 2005.

[24] Ö. Çelik, A. Teke, and H. B. Yıldırım, "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*, vol. 116, pp. 1–12, 2016.

[25] H. A. N. Hejase, M. H. Al-Shamisi, and A. H. Assi, "Modeling of global horizontal irradiance in the United Arab Emirates with artificial neural networks," *Energy*, vol. 77, pp. 542–552, 2014.

[26] D. V. S. K. Rao, M. Premalatha, and C. Naveen, "Analysis of different combinations of meteorological parameters in predicting the horizontal global solar radiation with ANN approach: a case study," *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 248–258, 2018.

[27] Z. Pang, F. Niu, and Z. O'Neill, "Solar radiation prediction using recurrent neural network and artificial neural network: a case study with comparisons," *Renewable Energy*, vol. 156, pp. 279–289, 2020.

[28] D. H. W. Li, W. Chen, S. Li, and S. Lou, "Estimation of hourly global solar radiation using multivariate adaptive regression spline (MARS) – a case study of Hong Kong," *Energy*, vol. 186, article 115857, 2019.

[29] A. Khosravi, R. N. N. Koury, L. Machado, and J. J. G. Pabon, "Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms," *Journal of Cleaner Production.*, vol. 176, pp. 63–75, 2018.

[30] M. Lotfinejad, R. Hafezi, M. Khanali, S. Hosseini, M. Mehrpooya, and S. Shamshirband, "A comparative assessment of predicting daily solar radiation using bat neural network (BNN), generalized regression neural network (GRNN), and neuro-fuzzy (NF) system: a case study," *Energies*, vol. 11, no. 5, p. 1188, 2018.

[31] R. Meenal and A. I. Selvakumar, "Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters," *Renewable Energy*, vol. 121, pp. 324–343, 2018.

[32] H. Loutfi, A. Bernatchou, and R. Tadili, "Generation of horizontal hourly global solar radiation from exogenous variables using an artificial neural network in Fes (Morocco)," *International Journal of Renewable Energy Research*, vol. 7, pp. 1097–1107, 2017.

[33] M. Lazzaroni, S. Ferrari, V. Piuri, A. Salman, L. Cristaldi, and M. Faifer, "Models for solar radiation prediction based on different measurement sites," *Measurement*, vol. 63, pp. 346–363, 2015.

[34] H. Long, Z. Zhang, and Y. Su, "Analysis of daily solar power prediction with data-driven approaches," *Applied Energy*, vol. 126, pp. 29–37, 2014.

[35] 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.

[36] A. Moghaddamnia, R. Remesan, M. H. Kashani, M. Mohammadi, D. Han, and J. Piri, "Comparison of LLR, MLP, Elman, NNARX and ANFIS models—with a case study in solar radiation estimation," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 71, no. 8-9, pp. 975–982, 2009.

[37] TMMOB, "Chamber of mechanical engineers, Turkey's energy outlook 2018," February 2020, https://www.mmo.org.tr/sites/default/files/EnerjiGorunumu2018_1.pdf.

[38] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.

[39] C. V. G. Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 2086–2090, Macao, Macao, 2019.

[40] S. Tasdemir, B. Yaniktepe, and A. B. Guher, "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*, vol. 5, pp. 131–139, 2018.

[41] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[42] A. K. Yadav, H. Malik, and S. S. Chandel, "Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 509–519, 2014.

[43] I. H. Witten, E. Frank, L. E. Trigg, M. A. Hall, G. Holmes, and S. J. Cunningham, "Weka: practical machine learning tools and techniques with Java implementations," *Computer Science Working Papers*, vol. 99/11, 1999.

[44] S. Tasdemir, B. Yaniktepe, and A. B. Guher, "Determination of wind potential of a specific region using artificial neural networks," *International Journal of Intelligent Systems and Applications in Engineering.*, vol. 3, no. 5, pp. 158–162, 2017.

[45] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[46] H. K. Elminir, Y. A. Azzam, and F. I. Younes, "Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models," *Energy*, vol. 32, no. 8, pp. 1513–1523, 2007.

[47] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 2000.

[48] Ö. Baydaroğlu and K. Koçak, "SVR-based prediction of evaporation combined with chaotic approach," *Journal of Hydrology*, vol. 508, pp. 356–363, 2014.

[49] C.-N. Ko and C.-M. Lee, "Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter," *Energy*, vol. 49, pp. 413–422, 2013.

[50] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, pp. 113–126, 2004.

[51] C.-C. Chang and C.-J. Lin, "Libsvm," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[52] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[53] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, no. 5-6, pp. 594–621, 2010.

[54] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, 2nd edition, 2009.

[55] J. R. Quinlan, "Learning with continuous classes," in *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence (AI '92)*, pp. 343–348, Singapore, Singapore, 1992.

[56] L. Wang, O. Kisi, M. Zounemat-Kermani et al., "Prediction of solar radiation in China using different adaptive neuro-fuzzy methods and M5 model tree," *International Journal of Climatology*, vol. 37, no. 3, pp. 1141–1155, 2017.

[57] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2nd edition, 2005.

[58] E. Frank and I. H. Written, "Generating accurate rule sets without global optimization," in *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 144–151, Madison, Wisconsin, USA, 1998.

[59] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[60] H. Darhmaoui and D. Lahjouji, "Latitude based model for tilt angle optimization for solar collectors in the Mediterranean region," *Energy Procedia*, vol. 42, pp. 426–435, 2013.

[61] B. Ahlgren, Z. Tian, B. Perers et al., "A simplified model for linear correlation between annual yield and DNI for parabolic trough collectors," *Energy Conversion and Management*, vol. 174, pp. 295–308, 2018.

[62] B. Ihya, A. Mechaqrane, R. Tadili, and M. N. Bargach, "Prediction of hourly and daily diffuse solar fraction in the city of Fez (Morocco)," *Theoretical and Applied Climatology.*, vol. 120, no. 3-4, pp. 737–749, 2015.

[63] S.-G. Kim, J.-Y. Jung, and M. Sim, "A two-step approach to solar power generation prediction based on weather data using machine learning," *Sustainability*, vol. 11, no. 5, p. 1501, 2019.