P-ISSN: 2798-1134 | E-ISSN: 2797-3549 | DOI: 10.52562/injoes.2023.821

Indonesian Journal of Earth Sciences

https://journal.moripublishing.com/index.php/injoes

| Research Article |

The Impact of Preprocessing Approaches on Neural Network Performance: A Case Study on Evaporation in Adana, a Mediterranean Climate

Okan Mert Katipoğlu [1], Muhammet Ali Pekin [2,*], Sercan Akil [3]

[1]Department of Civil Engineering, Erzincan Binali Yıldırım University, Erzincan 24002, Turkiye
[2]12th Regional Directorate of Meteorology, Turkish State Meteorology Service, 25050, Turkiye
[3]Research Department, Turkish State Meteorology Service, 06120, Turkiye
** Correspondence: mapekin@mgm.gov.tr*

Received: 08 September 2023 / Accepted: 02 December 2023 / Published: 29 December 2023

Abstract: The application of artificial intelligence (AI) technologies is quickly expanding in water management. Additionally, the artificial neural network methodology has an advantage over traditional statistical approaches in that it does not need assumptions about the distribution of data and variables. These methods can be used if there is a large enough data collection and criteria relevant to the nature of the problem. Preprocessing data before utilizing it improves the performance of the AI model. Evaporation matters in water management, agriculture processes and soil science. It is critical to ensure proper estimation of evaporation losses for effective water resource planning and management particularly in drought-prone areas such as Adana. Artificial intelligence approaches can be applied successfully in evaporation calculation. In this research, we used the Standard scaler, power transformer, robust scaler quantile transformer (Uniform) and quantile transformer (Normal), and min-max scaler preprocessing techniques to preprocess the multilayer perceptron neural network (MLPNN). We also trained the MLPNN using unprocessed data and compared it to the results of the preprocessed model. In the setup of the model, daily temperature, pressure, wind, sunny hours, and humidity parameters covering the years 2018-2021 were presented as input to the MLPNN model. Consequently, we pinpoint that all preprocessing approaches produce better outcomes than unscaled. Although all models produced statistically high accuracy predictions according to statistical criteria, the MLPNN model established without transformation (test phase: r^2: 0.93, NSE : 0.927, SMAPE: 10.77, RMSE: 1.2, MAE: 0.9) exhibited the lowest accuracy. The evaporation prediction model that was developed using the MLPNN-based standard scalar optimization algorithm exhibited the highest level of accuracy (test phase: r^2: 0.978, NSE: 0.977, SMAPE: 5.93, RMSE: 0.68, MAE: 0.48). Power Transformer (test phase: r^2: 0.978, NSE: 0.977, SMAPE: 5.81, RMSE: 0.67, MAE: 0.49) showed second-degree promising results. It can be concluded from these results that the estimation of meteorological variables requires the scaling and presentation of data in a uniform structure. Therefore, improving efficiency and productivity in water management or agricultural processes can be enhanced by making more accurate evaporation estimates.

Keywords: Evaporation, Artificial intelligence, Multilayer Perceptron Neural Network, Preprocessing, Standart scaler, Minimum Maximum scaler

INTRODUCTION

Studies on the human brain date back thousands of years. Several authors have considered the effects of artificial neural networks (ANN) are non-linear mapping systems with a structure loosely based on biological nervous system principles (Wang, 2013). Detailed examination of ANN by Oğuz & Pekin (2019; 2022) expressed that the success of the artificial neural networks method complies with the structure of the network architecture, the preparation of a sufficient number of training datasets selected from the appropriate parameters, and the optimum selection of hyperparameters. Another contributing factor to model performance in ANN techniques is preprocessing the training dataset with normalization-standardization techniques which have emerge as viable alternatives to traditional statistical modeling techniques in many scientific disciplines. This method is a dominant feature in solving nonlinear and complex problems. Multilayer perceptron is a good illustration of solving complex problems. Additionally, this layer can be classified into input, hidden, and output layers. Previous research has established that multi-lay perceptron are approximators for universal functions which can be used to create mathematical models (Cybenko, 1989).

An ANN is a machine learning technique that draws inspiration from the information processing of biological nervous systems. ANN offers substantial benefits in diverse engineering and natural sciences fields, including regression and classification. It comprises a substantial number of intricately linked neurons that collaborate to overcome diverse challenges. ANNs are based on the idea of using machines to replace the human brain's learning ability. Artificial neuron is a computing unit fundamental to a neural network's functioning. ANN comprises input, hidden, and output layers, with information being transferred and processed across these layers. In order to achieve the utmost success, preliminary data processing is necessary (Olaiya & Adeyemo, 2012). Normalization and standardization are scaling techniques that are utilized to establish relationships between variables and convert them into comparable structures, thereby simplifying operations. These pre-processing techniques are executed to enhance the analyses' dependability, compatibility, and quality. It is necessary when the data features exhibit substantial ranges (Jain et al., 2018; Raju et al., 2020).

Evaporation is a crucial component of the water cycle. A growing body of literature recognizes the importance of evaporation, which is an essential part of the climate system and plays a vital role in sustaining a supply of heat from the atmosphere. Early examples of research into the importance of vaporization include: the formation of evaporation and the conditions necessary for its occurrence (Penman, 1956; Monteith, 1965).

Evaluation of machine learning models can alter working on the models that can be categorized into regression models and classification models, including confusion matrix. As noted by Liu et al. (2017) to enhance the accuracy of machine learning models, more data leads to have reliable results. Implementing hyperparameter tuning and creating ensemble model with multiple trained models, such as linear regression and support vector machines, can increase the accuracy of the machine learning models. Unlike Liu et al. (2017), Podhorányi (2021) argues that big data processing makes data more informative. Abghari et al. (2012) combined wavelet transform and MLP network based on Mexican Hat and polyWOG1 mother wavelet for daily pan evaporation prediction at Lar synoptic station. The analysis showed that the MLP model based on the Mexican Hat is superior in predicting daily pan evaporation. Ghorbani et al. (2018) employed MLP, SVM, and Firefly Algorithm (FFA) to predict the daily pan evaporation in Northern Iran. Thus, the MLP-FFA model is validated with enhanced prediction performance when compared to both the MLP and SVM models. Nourani et al. (2020) conducted a study on the effect of data pre-processing on constructing prediction intervals for the evaporation process at three stations in Iran. Consequently, a replicated smoothed time series pattern can be produced and utilized during training. Ehteram et al. (2022) combined MLP, Multi-objective based salp swarm algorithm (MOSSA), crow algorithm (MOCA) and particle swarm optimization (MOPSO) models to predict daily evaporation. The results of the analysis showed that MLP-MOSSA predicted evaporation most accurately.

Adana has explicitly been selected as the study area due to its torrid and arid Mediterranean climate, renowned for having the highest evaporation rate in Turkey and occupying a prominent position in the countries' agricultural productivity. The region's agricultural activities are intricately tied to the practice of irrigation. For this reason, accurate evaporation estimation is critical for irrigation planning and water distribution. There are Çukurova Plain and Seyhan Plain in the region where many crops such as cotton, wheat, corn, and citrus fruits are grown. In addition, it is vital to develop forecasting models to analyze the changes in the water of dams and ponds used as irrigation and domestic water in the region.

The following criteria were utilized in this study to forecast the air temperature of Adana province: maximum temperature, lowest temperature, wind, humidity, and sunlight duration between 2016 and 2020. This research aims to discuss the performance and accuracy of artificial neural networks when different normalization algorithms are applied to their inputs (raw data set). Normalizing the data has been proven in the literature to increase the performance of the artificial neural network (Masters, 1993). The application of the appropriate normalization approach varies depending on the data's specific characteristics and the study's objectives. Determining the best preprocessing technique to predict evaporation is a topic limited to the literature. For this, the performance of different normalization techniques must be evaluated using appropriate criteria. The main motivation of this study is to evaluate the effect of MinMax Scaler, Power Transformer, Quantile Transformer, Robust Scaler, Standard Scaler data normalization techniques on evaporation estimation and to reveal the accuracy of various meteorological variables in evaporation estimation.

MATERIALS & METHODS

Study Area

Figure 1 shows the location of the weather observation station in the south of Turkey, which is approximately at sea level in Adana. Latitude and longitude coordinates are 37.0041 and 35.3442, respectively. According to the Köppen classification, Adana has a hot-summer Mediterranean climate, which can be seen in the region. Similarly, the Trewartha classification defines the region as a dry summer subtropical climate. Because of climatic conditions, the amount of evaporation is quite high. It is observed that the temperature does not reach below zero in most of the winter times. Therefore, evaporation can occur throughout the year (Özfidaner et al., 2018).

Figure 1. Location of the Yüreğir Weather Observation Station.

What stands out in Figure 2 is the general pattern of the schematic diagram of the study.

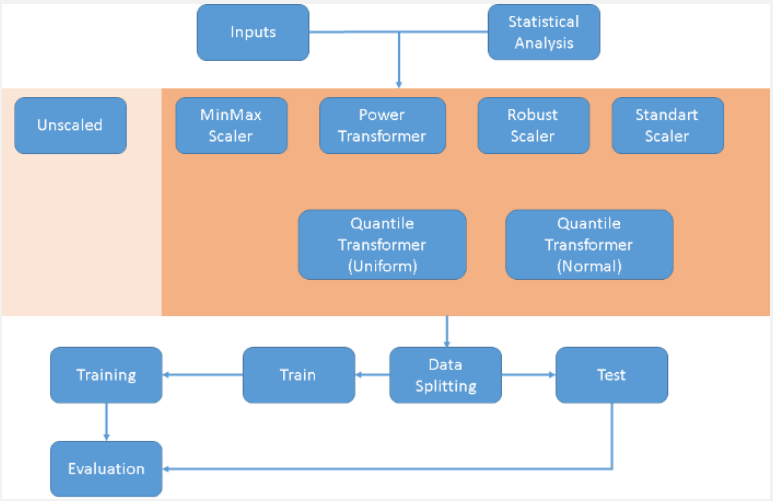

Figure 2. Schematic Diagram of The Study.

Data

In the case of tabular data, a data set relates to one or more database tables, where each row refers to a specific record in the corresponding data set and each column to a single variable. The training data set consisted of four years since it has been noticed that some earlier data are erroneous, and some are missing. 1311 rows in total were used, which was enough for this investigation. Data from four years were used (from 2018 to 2021). Any rows in the series that have a missing row are removed. Data from 2018 through 2020 served as a train, while data from 2021 served as a test. 1311 rows by 6 columns total, 1040 rows by 6 columns for training, and 271 rows by 6 columns for testing.

In selecting the data used in the study, the period with the longest and complete time period was chosen. The data were sourced from the General Directorate of Meteorological Affairs of Turkey. These data pertain to meteorological observations at stations.

The database contains one day's worth of data for October 2020, three days' worth for November 2021, for November 2021, and none for October 2021. Other years' and months' data are largely complete. This paper measured meteorological data were applied on a daily basis: (1) daily pan evaporation in millimeters (PE); (2) daily mean temperature in degrees Celsius (T); (3) relative humidity as a percentage (RH); (4) daily mean wind speed in meters per second (W).

Table 1 provides the descriptive statistics for the data set that was used. According to the study findings, the four-year mean air temperature is 20.3°C. The lowest temperature is 2°C and the maximum is 34.2 °C. The relative humidity ranges from 98.8% to 22.2%. 67% humidity is the four-year average. 1009.7 hPa is the four-year average pressure. For four years, the pressure fluctuated between 995.5 and 1027.5 hPa. The wind speed was measured to be between 0.2 and 3.4 meters per second, with a four-year average of 1.5 meters per second. The highest amount of sunshine in four years was 12.6 hours, while the average was 7.1 hours. A maximum of 18.1 mm/day of evaporation was recorded over a period of four years.

Table 1. Descriptive Statistics of Data Used

	T	RH	SP	W	S	Pan
Mean	20.3	67.0	1009.7	1.5	7.1	7.2
Dev.	7.3	12.8	5.8	0.4	3.5	4.3
Range	32.2	76.6	32.0	3.2	12.6	18.1
Min.	2.0	22.2	995.5	0.2	0.0	0.0
Max.	34.2	98.8	1027.5	3.4	12.6	18.1

Preprocessing Methods

A total of five distinct preprocessing techniques were investigated in this study. The Minimal Maximum Scaler is the first of these approaches. Each value in the data collection is scaled from 0 to 1 using the Minimum Maximum Scaler. It is a straightforward strategy commonly employed in artificial intelligence research (Deepa & Ramesh, 2022).

In the literature, there are two sorts of power transformer methods. These approaches are the methods proposed by (Box & Cox, 1964; Yeo & Johnson, 2000). While the Box-Cox approach can only be used for positive data, Yeo and Johnson's method does not have this limitation. Because temperature may take negative values, Yeo and Johnson's Power Transform approach is more suited for our investigation. As a result, we adopted Yeo and Johnson's Power Transform approach in our research. The approach is used in conjunction with maximum likelihood (λ) estimate. A quantile transformer can be used to lower the weight of outliers. The estimate of the data's cumulative distribution function is employed in this approach. The quantile function (G^(-1)) is then used to scale it to the appropriate amount. Outlier values are thus equalized to the range's boundaries in this manner. It should be noted that this strategy is not linear.

There are two types of cumulative distribution functions: normal and uniform (Chanal et al., 2021; Seo et al., 2015). Our investigation used the quantile transformer with two distinct cumulative distribution functions (F(x)). Robust Scaler scales data based on 75% and 25% percentiles. Also, does scaling comply with median value. As a result, outlier numbers have less of an impact on the model. It should be noted that there are outlier values in the data, but with a lesser impact (Reddy et al., 2021).

Preprocessing techniques are commonly used in machine learning to transform raw data into a format that can be effectively analyzed and processed by models. These techniques may include data cleaning, scaling, normalization, feature selection, dimensionality reduction, and so on. The choice of preprocessing techniques depends on the specific characteristics of the dataset and the objectives of the analysis. It is common to experiment with different techniques and choose the ones that yield the best results for a particular problem. To apply preprocessing techniques in machine learning, several common steps can be followed: Data Cleaning: this involves handling missing values, outliers, and noise in the dataset; Data Normalization/Scaling: this step ensures that all features have a similar scale, which helps prevent certain features from dominating others during training; Feature Selection/Extraction: it involves choosing relevant features or transforming existing features to improve model performance and reduce dimensionality; Encoding Categorical Variables: categorical variables are often encoded into numerical representations for machine learning algorithms to process them

effectively; Splitting Dataset: the dataset is typically divided into training, validation, and test sets to evaluate model performance.

Several artificial intelligence investigations require dataset standardization because the input numbers are unlikely to follow the conventional normal distribution. This situation may jeopardize the model's success. The typical scaler approach centers and scales the data on the mean and its standard deviation. As a result, the data has a unit variance and a mean of zero (Thara et al., 2019). The preprocessing techniques and formulas tested in this study are given in Table 2.

Table 2. Preprocessing Methods Used

Methods	Formula
MinMax Scaler	$\text{scaled } x_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
Power Transformer	$\text{scaled } x_i^{(\lambda)} = \begin{cases} \frac{[(x_i+1)^2 - 1]}{\lambda} & \text{if } \lambda \neq 0, x \geq 0, \\ \ln(x_i + 1) & \text{if } \lambda = 0, x \geq 0, \\ -\frac{[(-x_i+1)^{2-\lambda}]}{2-\lambda} & \text{if } \lambda \neq 2, x < 0, \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x < 0 \end{cases}$
Quantile Transformer	$\text{scaled } x_i = G^{-1}(F(x))$
Robust Scaler	$\text{scaled } x_i = \frac{x_i - \text{med}(x)}{Q75(x) - Q25(x)}$
Standard Scaler	$\text{scaled } x_i = \frac{x_i - \mu(x)}{\sigma(x)}$

Multilayer Perceptron Neural Network (MLPNN) Structure

Multilayer Perceptron Neural Networks (MLPNNs) are a type of artificial neural network consisting of multiple layers of interconnected nodes (neurons). They are particularly useful in solving complex problems, making predictions, and pattern recognition. MLPNNs have the ability to learn and generalize. There are no set guidelines for choosing the network architecture in investigations of artificial neural networks. Most architecture is developed through a trial-and-error process based on the kind of input parameters and the nature of the problem to be solved (Ma et al., 2023). As can be seen in Figure 3, the structure of the study MLPNN.

Nodes are the locations in MLPNN where data is processed and information is saved. Although there is no hard and fast rule for determining the number of nodes, depending on the nature of the issue, an infinite number of choices can be employed (Chung & Sohn, 2023). In this study 50 nodes (n) are specified. The information is passed from one node to the next using the activation function. In curve fitting investigations, the Rectified Linear Unit (ReLU) is widely recommended as an activation function (Ismail et al., 2023).

As for configuring a Multi-Layer Perceptron Neural Network (MLPNN), there are several key considerations: Architecture: determine the number of layers and the number of neurons in each layer. This choice depends on the complexity of the problem and the amount of available data; Activation Functions: select appropriate activation functions for each layer to introduce non-linearity into the network. Common choices include ReLU, sigmoid, or tanh; Loss Function: choose an appropriate loss function based on the nature of the problem. For classification tasks, cross-entropy loss is commonly used, while mean squared error is often used for regression problems; Optimization Algorithm: select an optimization algorithm, such as stochastic gradient descent (SGD) or Adam, to update the network weights during training; Hyperparameter Tuning: adjust hyperparameters like learning rate, batch size, regularization parameters, and the number of iterations/epochs to optimize the model's performance. We used ReLU in our since its output is based on curve fitting. After the network is built, it needs to be optimized. Adam solver is said to be quicker than other solvers (Ioannou et al., 2022). As a result, in

our investigation, we used the Adam solver as the optimizer. The epoch in the training phase is defined by how many times the inputs in the network are rotated and the weights are trained (Gao & Zhang, 2023). We tried numerous alternative epoch numbers at the start of our work and concluded that 500 epochs were sufficient. After that, working with this epoch number. It should be noted that the major goal of this study is to evaluate preprocessing approaches; the machine learning method is only a testing tool for us. Figure 3 depicts the structure of our MLPNN.

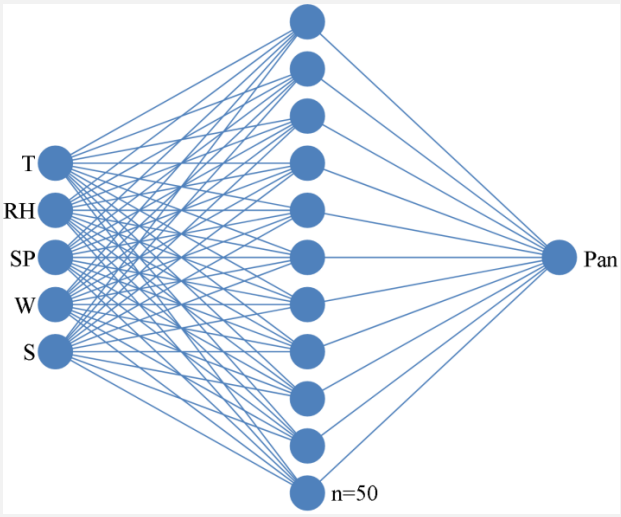

Figure 3. Multilayer Perceptron Neural Network Structure (MLPNN) Topology.

Evaluation Metrics

There are several assessment criteria for evaluating the outcomes of regression studies. A single criterion is frequently insufficient to give an accurate assessment. It is an appropriate strategy for selecting more than one criterion based on the scope of the investigation. We picked these criteria because the coefficient of determination (r^2), RMSE, and MAE are extensively used in the literature (Chicco et al., 2021). The equations of our evaluation metrics are given in the Table 3 below. The Nash-Sutcliffe Efficiency technique is another relevant criterion. This criterion was created for hydrological research (Apaydin et al., 2020; Nash & Sutcliffe, 1970). SMAPE is a method built on the MAPE method. It gives a better assessment opportunity by scaling the result between 0 and 100 (Flores, 1986; Goodwin & Lawton, 1999).

Table 3. Evaluation Methods

Equation	Scores
$r^2 = \frac{(\sum_{i=1}^{n}(O_i - \overline{O})(S_i - \overline{S}))^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}$	$0 \leq r^2 \leq 1$, Best Score is 1.
$NSE = 1 - \frac{\sum_{t=1}^{n}(S_i - O_i)^2}{\sum_{t=1}^{n}(O_i - \overline{O})^2}$	$-\infty < NSE < 1$, Best Score is 1.
$SMAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\lvert S_i - O_i\rvert}{\lvert S_i\rvert + \lvert O_i\rvert}$	$0 \leq SMAPE \leq 100$, Best score is 0.
$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(S_i - O_i)^2}{n}}$	$0 \leq RMSE < +\infty$, Best score is 0.
$MAE = \frac{\sum_{i=0}^{n}\lvert S_i - O_i\rvert}{n}$	$0 \leq MAE < +\infty$, Best score is 0.

Rank analysis

Determining the best model becomes complex when multiple statistical indicators are used. The statistical indicators employed in this study were assigned separate rank values to determine the most effective model, all done in light of the abovementioned reason. During rank analysis, a rank is

given to every performance parameter. Rankings were arranged from the maximum value equal to the number of models, which was three in our study, to the minimum value equal to one. The third rank is given to the model that performs the best, whereas the first rank is assigned to the model that performs the worst. The model with the highest total rank indicates the best, while the model with the lowest shows the worst (Zhang et al., 2020).

Softwares

As a first step to boosting ML models by HPO, it is important to identify what the key hyper-parameters are that should be tuned to fit the models to specific problems or datasets. As a general rule, ML models can be divided into supervised and unsupervised learning algorithms based on whether the models are designed to model labeled or unlabeled datasets. Machine learning algorithms that use supervised learning are those that map input features to targets based on labeled data, and mainly include. The majority of the work was performed using Python. MLPNN was developed using the scikit learn package, which was also used to prepare the data. For the computation of the evaluation criteria, the HydroErr package was used. For the creation of the charts, QGIS and Microsoft Excel were used. In order to analyze the input data statistically, the Jasp software was used (Buitinck et al., 2013; Roberts et al., 2018).

RESULTS

Statistical Analysis of Data

The results of this study indicate that daily statistics on evaporation for a four-year period. According to Figure 4, evaporation increases to its highest levels in the summer and declines to zero in the winter. Additionally, it is noteworthy that the trend in evaporation from 2018 to 2022 shows a modest increase.

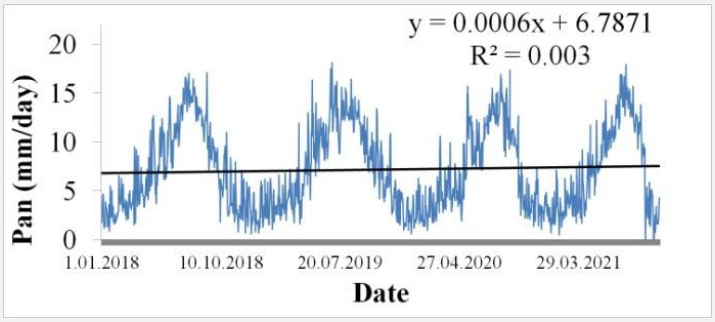

Figure 4. Daily Total Pan Evaporation.

Figure 5 displays the total monthly evaporation data during the previous four years. Although July has the highest monthly evaporation (394.2 mm), and January has the lowest (83.9 mm).

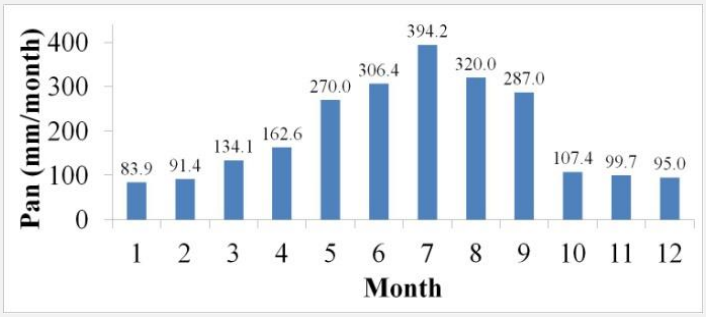

Figure 5. Daily Total Pan Evaporation.

Table 4 indicates the relationships between the data that were used. Temperature had the strongest association with evaporation (CC=0.90), as expected. Wind speed and sunshine length are

strongly positively correlated. Pressure has a significant negative association (CC=-0.73). A negligible but statistically significant association existed between humidity and temperature (CC=-0.16, p.001).

Table 4. Correlations of The Dataset

	T	RH	SP	W	S	Pan
T	—					
RH	-0.12 ***	—				
SP	-0.68 ***	-0.15 ***	—			
W	0.20 ***	-0.19 ***	-0.40 ***	—		
S	0.62 ***	-0.38 ***	-0.26 ***	0.06 *	—	
Pan	0.90 ***	-0.16 ***	-0.73 ***	0.52 ***	0.56 ***	—

* p < .05, ** p < .01, *** p < .001

Evaluation Scores of Preprocessing Methods

Table 5 provides the results for the training and test phases. The bottom line of the table in this example displays the scores of the pan values that were modeled without using any preprocessing techniques as Unscaled. Unscaled claims that, preprocessing, the pan could be modeled with a r^2=0.93 value. Unscaled will be a useful benchmark to illustrate how well preprocessing techniques work. The table shows that the performance of the artificial neural network model was most positively impacted by the standard scaler method. The train and test phases of this approach yield great results. r^2=0.978, NS=0.977, SMAPE=5.93, RMSE=0.68, and MAE=0.48 were the test phase values. The power transformer performed best after the conventional scaler. According to the test phase results, robust scaler, Quantile Transformer (Uniform and normal), and min-max scaler techniques also outperformed the scores obtained without preprocessing. These approaches' scores in the train and test phases were less strong than those of the other methods. According to the table in the below, all preprocessing methods outperformed unscaled. The rank analysis revealed that the standard scalar transformation technique achieved the highest rank value (60), making it the most successful preprocessing technique. The Power Transformer technique closely followed it. The MLP that utilized the unscaled transform also generated the weakest prediction outputs, as showed by its lowest rank value (10).

Table 5. Results of Training and Test Phases.(Bolds indicate the best scores)

	r2		NSE		SMAPE		RMSE		MAE		Total rank
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	
MinMax Scaler	0.958	0.954	0.958	0.954	7.57	9.17	0.87	0.96	0.66	0.67	21
Rank score	3	2	2	3	2	2	2	3	2	2	
Power Transformer	0.985	**0.978**	0.985	0.977	3.72	**5.81**	0.52	**0.67**	0.39	0.49	56
Rank score	6	6	6	6	6	7	6	7	6	6	
Quantile Transformer (Normal)	0.978	0.961	0.978	0.952	4.35	6.9	0.62	0.97	0.47	0.65	32
Rank score	5	3	4	2	4	4	4	2	4	4	
Quantile Transformer (Uniform)	0.968	0.965	0.968	0.962	6.74	8.74	0.75	0.87	0.59	0.67	31
Rank score	4	4	3	4	3	3	3	4	3	3	
Robust Scaler	0.983	0.975	0.983	0.974	3.82	5.92	0.55	0.72	0.41	0.52	41
Rank score		5	5	5	5	6	5	5	5	5	
Standard Scaler	**0.986**	**0.978**	**0.986**	**0.977**	**3.68**	5.93	**0.5**	0.68	**0.38**	**0.48**	**60**
Rank score	7	7	7	7	7	5	7	6	7	7	
Unscaled	0.911	0.93	0.91	0.927	10.19	10.77	1.27	1.2	0.99	0.9	10
Rank score	2	1	1	1	1	1	1	1	1	1	

The scatter values are used to visualize the distribution of both actual and predicted values. The modeled pan and observed pan values in the train and test sections are shown in scatter plots in the Figure 6. The pan data modeled using the standard scaler, power transformer, and robust scaler approaches are seen to fit the observation better in the graph. According to the table and scatter plot, it is feasible to conclude that the power, standard scaler, and robust scaler approaches are more effective than other ways. It is important to note that the standard style conversion process is more active than other conversions, which should be considered. The reason why the standard scalar transformation is considered superior to real values can be inferred from the fact that it exhibits a linear distribution and is centered around the 45-degree regulation line.

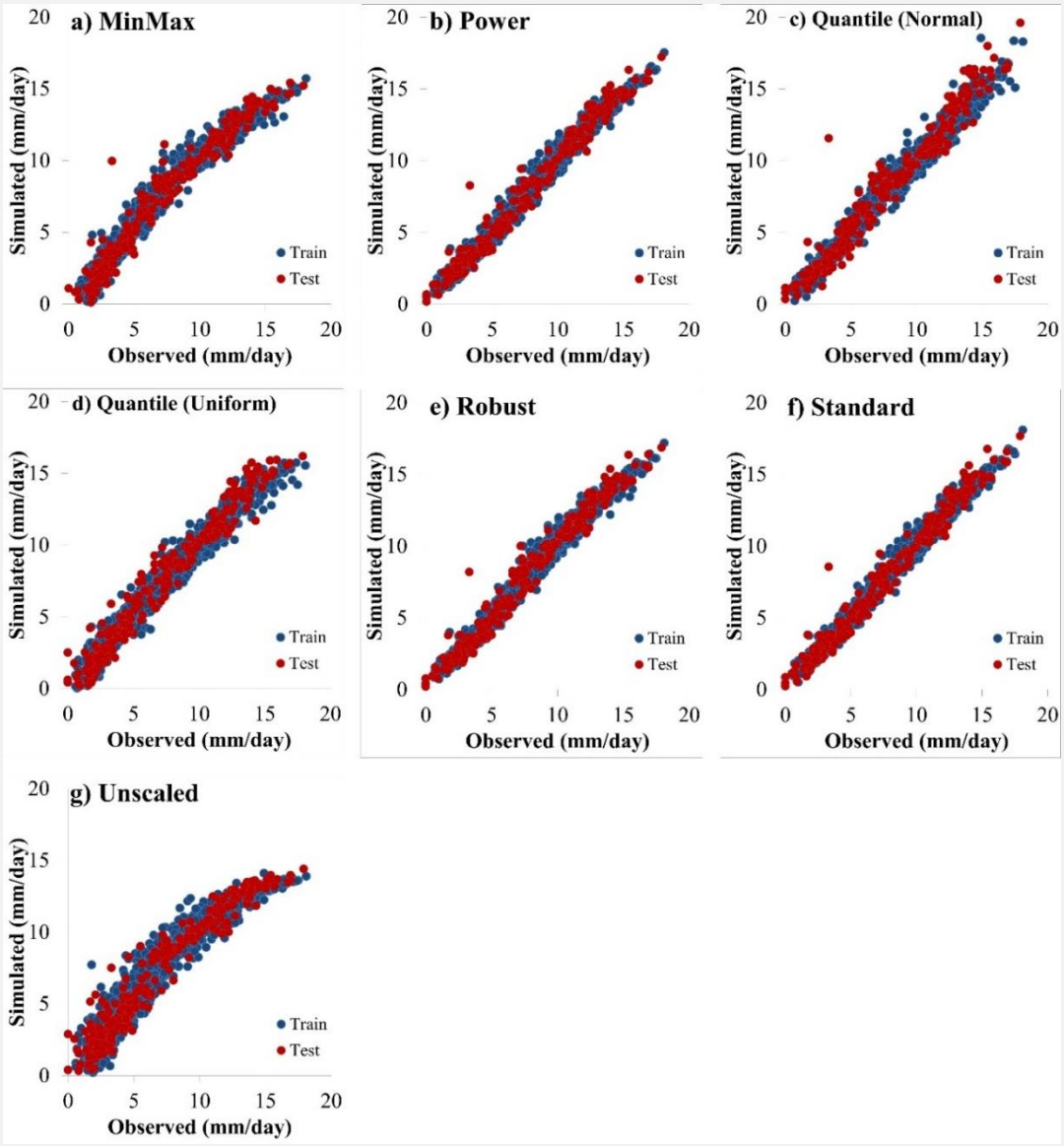

Figure 6. Scatter plots of Observed and Simulated Data.

Figure 7 shows the time series graphs of the pan values modeled and observed using the power, robust, and standard scaler approaches. All three approaches are seen to be in good agreement with the findings in this graph. The graph also shows that the extreme values (a, b, c, and d) can be successfully predicted. When examining Figure 7, it is worth noting that the models constructed with dimensions exceeding 12 mm exhibited slight deviations from the actual values. In Figure 7b and c, it can be observed that the difference between the predicted values and the observed values exhibits an increase below the threshold of 1 mm. In addition, it can be deduced that the standard scaler transform can generate more realistic outputs compared to other preprocessing techniques.

DISCUSSION

A strong relationship between conventional scaler method and preprocessing techniques has been reported in the literature. Regarding the first research question, it was found that the conventional scaler method is the most effective preprocessing technique. Choosing one of the standard scaler, power, or robust scaler approaches depends on the nature of the problem to be handled.

Comparison of the findings with those of other studies confirms Air temperature, humidity, wind, speed, air pressure and sunshine duration are among the most significant meteorological

parameters for evaporation. It also reported by Abed et al. (2023). A statistically significant high correlation with temperature, wind speed, and sunlight duration in the study. Whereas humidity and air pressure had negative associations.

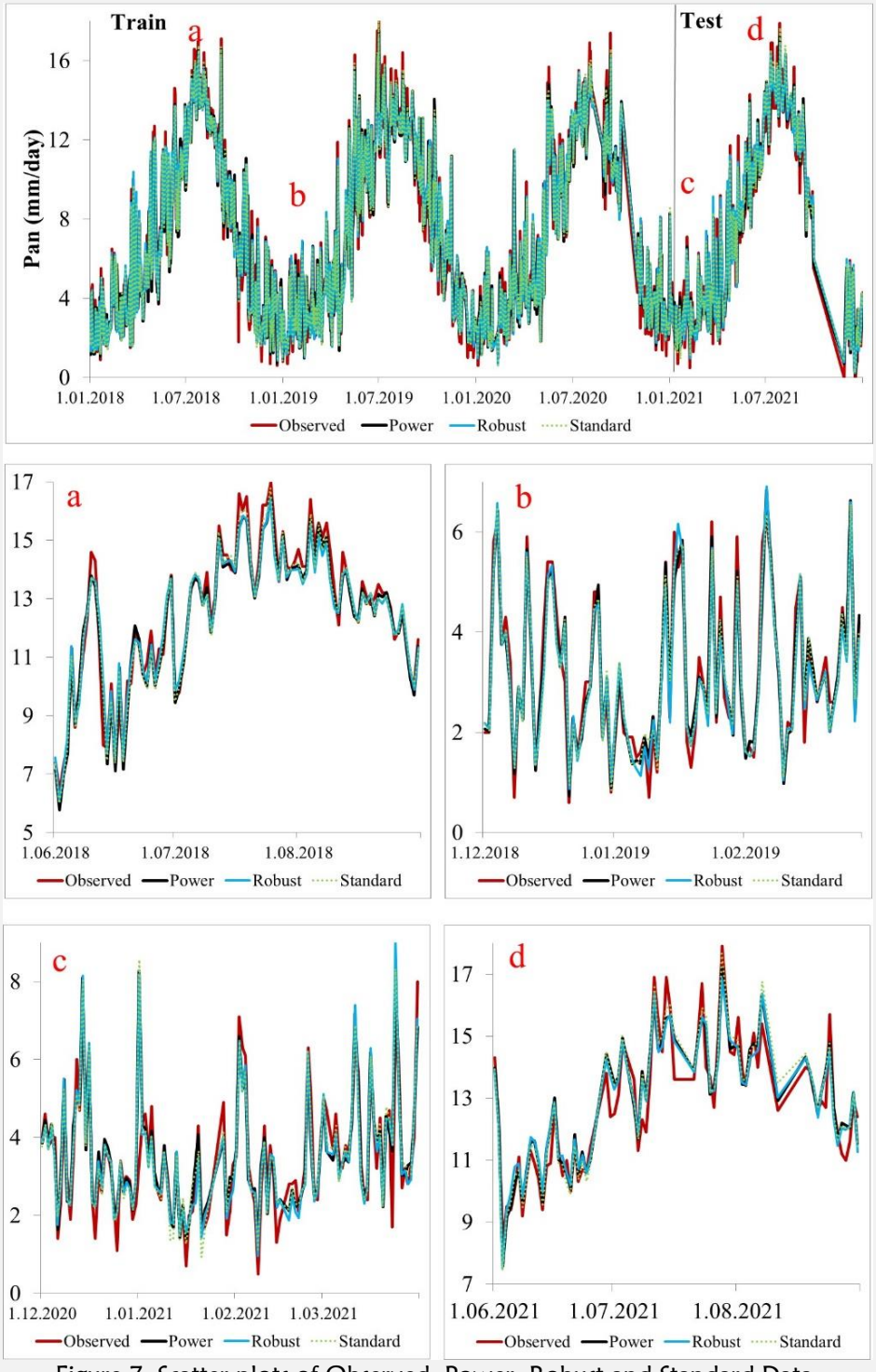

Figure 7. Scatter plots of Observed, Power, Robust and Standard Data.

The most obvious finding from the analysis is that the preprocessing approaches are compared to the unscaled experiment. The most successful outcomes are Standard Scaler and Power transformers. Data normalization is among the preprocessing techniques and can improve the accuracy of artificial intelligence models by scaling the data and reducing its variance. With data normalization,

meteorological variables of various scales were scaled and expressed in the same range. Thus, by increasing the generalization ability of the AI model, the learning process was improved and the model accuracy was increased. Therefore, this approach guarantees that the established model will acquire knowledge in a balanced and consistent way. Data normalization increased generalization ability and forecast performance by reducing the noise and unnecessary variance in meteorological data. The results obtained coincide with the studies of Singh et al. (2009) and Singh & Singh (2020).

Kisi & Zounemat-Kermani (2014) used ANFIS with grid partition (GP) and subtractive clustering (SC) method to predict daily reference evapotranspiration (ET0). ANFIS models performed better than the corresponding empirical equations in modeling ET0. Gümüş et al. (2016) used the climatic various parameters to model Adana's monthly average evaporation. For this, ANFIS, ANN and GEP models were applied. It was obtained that the evaporation estimation of all methods used gave good results. But, the combination of 6 inputs indicated the best outputs in the ANFIS. Sarıgöl & Katipoğlu (2023) estimated the evaporation values of the Southeast Anatolia Project (GAP) area in Turkey using gradient boosting machines (GBM) and Mode Decomposition techniques. Precipitation, average, minimum and maximum air temperature, wind speed, actual air pressure, relative humidity, and solar time variables were first divided into subcomponents and then presented as input to the GMB model. As a result of the analysis, it was concluded that the data preprocessing based GBM model produced the best output in evaporation prediction. Kisi & Zounemat-Kermani (2014); Gumus et al. (2016); Sarıgöl & Katipoğlu (2023) studies confirm the study in terms of the input combinations used in evaporation prediction and the successful prediction ability of AI mathematicians.

By combining artificial intelligence techniques with data normalization techniques, hydrological and meteorological forecasts can experience an improvement in model prediction accuracy. Implementing this can greatly enhance efficiency and productivity in water management or agriculture processes, enabling decision makers to take quicker action and implement more reliable resource and risk management strategies.

The main limitations of this study are that it only used 4 years of data, only used weather-based variables as input, and only the output was modeled according to the MLPNN algorithm. In addition, it is expected that future research can achieve more accurate evaporation repair results by combining various machine learning algorithms and hybrid algorithms. We also suggest that future studies could improve evapotranspiration estimates by including additional parameters such as vegetation cover, groundwater levels, semi-arid conditions, and topographic factors. Considering these variables, datasets on temperature, humidity, pressure, wind speed, and sunshine length will be beneficial in pan-modeling research. The high-layer artificial neural network (MLPNN) approach is advised because it is effective. Future studies on the current topic are therefore recommended.

CONCLUSION

In pan evaporation estimation research, input parameters: temperature, humidity, pressure, wind speed and sunlight duration are adequate in terms of evaluation of the results. More broadly, research is also needed to determine meteorological components strongly related to Pan. The research may employ the MLPNN approach in operational pan calculations since reaching reliable results. The analysis led to the conclusion that the evaporation prediction model, which was constructed using the MLPNN-based standard scalar optimization algorithm, exhibited the highest level of accuracy (test phase: r^2: 0.978, NSE: 0.977, SMAPE: 5.93, RMSE: 0.68, MAE:0.48).

This investigation aimed to assess the success of the preprocessing methods on machine learning performance. The findings reported here shed new light on multiple machine learning techniques that can be carried with preprocessing methods. As a result, greater efforts are needed to ensure preprocessing phase should not be skipped in machine learning performance. Standard scaler, power transformer, robust scaler, quantile transformer (Uniform) and quantile transformer (Normal), and min-max scaler are all successful methods respectively as test stage. Other preparation approaches are described in the literature review. This would be a fruitful area for further work.

An improvement in model prediction accuracy can be achieved for hydrological and meteorological forecasts by implementing the MLPNN model and data normalization techniques. The implementation of this can have a significant impact on enhancing efficiency and productivity in managing crop water requirements and irrigation. As a result, it enables decision makers to take quicker action and implement more reliable strategies for water resource management, climate change adaptation, and risk management.

The study highlighted the importance of using data preprocessing techniques, specifically normalization, in predicting meteorological variables like evaporation. Additionally, it can be implied that artificial intelligence models constructed using unprocessed data fail to reach their full potential because they lack scalability.

The main limitation of this study stems from the fact that the optimization of the parameters used was not performed. It is highly recommended that future studies evaluate the performance of prediction models for evaporation, utilizing a range of machine learning algorithms including M5Tree, Self-Organizing Maps, K-Nearest Neighbors, and XGBoost. In addition, one can also use deep learning algorithms like Convolutional Neural Networks, Recurrent Neural Networks, and Deep Belief Networks. By implementing metaheuristic optimization techniques to optimize their parameters, these models can be taken to the next level, resulting in prediction results that are not only more accurate but also more realistic.

ACKNOWLEDGMENT

Authors are thankful to the Turkish State Meteorological Service for providing the dataset used in this study.

REFERENCES

Abed, M., Imteaz, M. A., Ahmed, A. N., & Huang, Y. F. (2023). A novel application of transformer neural network (TNN) for estimating pan evaporation rate. *Applied Water Science, 13*(2), 31. https://doi.org/10.1007/s13201-022-01834-w

Abghari, H., Ahmadi, H., Besharat, S., & Rezaverdinejad, V. (2012). Prediction of daily pan evaporation using wavelet neural networks. *Water resources management, 26*, 3639-3652. https://doi.org/10.1007/s11269-012-0096-z

Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., & Chau, K. W. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water, 12*(5), 1500. https://doi.org/10.3390/w12051500

Bhatnagar, R. (2018). Machine learning and big data processing: a technological perspective and review. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)* (pp. 468-478). Springer International Publishing. https://doi.org/10.1007/978-3-319-74690-6_46

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 26*(2), 211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*. https://doi.org/10.48550/ARXIV.1309.0238

Chanal, D., Steiner, N. Y., Chamagne, D., and Pera, M. C. (2021, October). Impact of standardization applied to the diagnosis of LT-PEMFC by Fuzzy C-Means clustering. In *2021 IEEE Vehicle Power and Propulsion Conference (VPPC)* (pp. 1-6). IEEE. https://doi.org/10.1109/VPPC53923.2021.9699234

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623. https://doi.org/10.7717/peerj-cs.623

Chung, D., & Sohn, I. (2023). Neural Network Optimization Based on Complex Network Theory: A Survey. *Mathematics, 11*(2), 321. https://doi.org/10.3390/math11020321

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems, 2*(4), 303–314. https://doi.org/10.1007/BF02551274

Deepa, B., & Ramesh, K. (2022). Epileptic seizure detection using deep learning through min max scaler normalization. *International Journal of Health Sciences, 6*(S1), 10981–10996. https://doi.org/10.53730/ijhs.v6nS1.7801

Dongare, A.D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology, 2*(1), 189-194.

Ehteram, M., Panahi, F., Ahmed, A. N., Huang, Y. F., Kumar, P., & Elshafie, A. (2022). Predicting evaporation with optimized artificial neural network using multi-objective salp swarm algorithm. *Environmental Science and Pollution Research*, 1-27. https://doi.org/10.1007/s11356-021-16301-3

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega, 14*(2), 93-98. https://doi.org/10.1016/0305-0483(86)90013-7

Gao, F., & Zhang, B. (2023). Data-aware customization of activation functions reduces neural network error. *arXiv preprint arXiv:2301.06635*. https://doi.org/10.48550/arXiv.2301.06635

Garrido, M. C., Cadenas, J. M., Bueno-Crespo, A., Martínez-España, R., Giménez, J. G., & Cecilia, J. M. (2022). Evaporation forecasting through interpretable data analysis techniques. *Electronics, 11*(4), 536. https://doi.org/10.3390/electronics11040536

Ghorbani, M. A., Deo, R. C., Yaseen, Z. M., H. Kashani, M., & Mohammadi, B. (2018). Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. *Theoretical and applied climatology, 133*, 1119-1131. https://doi.org/10.1007/s00704-017-2244-0

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting, 15*(4), 405-408. https://doi.org/10.1016/S0169-2070(99)00007-2

Gümüş, V., Şimşek, O., Soydan, N. G., Aköz, M. S., & Yenigün, K. (2016). Adana istasyonunda buharlaşmanın farklı yapay zeka yöntemleri ile tahmini. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 7(*2), 309-318.

Ioannou, G., Tagaris, T., & Stafylopatis, A. (2023). AdaLip: An Adaptive Learning Rate Method per Layer for Stochastic Optimization. *Neural Processing Letters, 55*, 1-28. https://doi.org/10.1007/s11063-022-11140-w

Ismail, A. H., Soliman, T. A., Rihan, M., & Dessouky, M. I. (2023). Deep Learning-Based Beamforming for Millimeter-Wave Systems Using Parametric ReLU Activation Function. *Wireless Personal Communications, 129*(2), 825-836. https://doi.org/10.1007/s11277-022-10157-7

Jain, S., Shukla, S., & Wadhvani, R. (2018). Dynamic selection of normalization techniques using data complexity measures. *Expert Systems with Applications, 106*, 252-262. https://doi.org/10.1016/j.eswa.2018.04.008

Kisi, O., & Zounemat-Kermani, M. (2014). Comparison of two different adaptive neuro-fuzzy inference systems in modelling daily reference evapotranspiration. *Water resources management, 28*, 2655-2675. https://doi.org/10.1007/s11269-014-0632-0

Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics, 1(1), 48-56. https://doi.org/10.1016/j.visinf.2017.01.006

Ma, J., Wang, J., Han, Y., Dong, S., Yin, L., & Xiao, Y. (2023). Towards data-driven modeling for complex contact phenomena via self-optimized artificial neural network methodology. *Mechanism and Machine Theory, 182*, 105223. https://doi.org/10.1016/j.mechmachtheory.2022.105223

Masters, T. (1993). *Practical Neural Network Recipes in C++.* London: Academic Press, Inc.

Monteith, J. L. (1965). Evaporation and environment. In *Symposia of the society for experimental biology* (Vol. 19, pp. 205-234). Cambridge University Press (CUP) Cambridge.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology, 10*(3), 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

Nourani, V., Sayyah-Fard, M., Alami, M. T., & Sharghi, E. (2020). Data pre-processing effect on ANN-based prediction intervals construction of the evaporation process at different climate regions in Iran. *Journal of Hydrology, 588*, 125078. https://doi.org/10.1016/j.jhydrol.2020.125078

Oğuz, K. & Pekin, M. A. (2019). Predictability of Fog Visibility with Artificial Neural Network for Esenboga Airport . Avrupa Bilim ve Teknoloji Dergisi (15), 542-551 . https://doi.org/10.31590/ejosat.452598

Oğuz, K. & Pekin, M. A. (2022). Makine Öğrenme Algoritmaları ile PM10 Konsantrasyon Tahmini. *Journal of Advanced Research in Natural and Applied Sciences, 8*(2), 201-213. https://doi.org/10.28979/jarnas.981202

Olaiya, F., & Adeyemo, A. B. (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. *International Journal of Information Engineering and Electronic Business, 4*(1), 51–59. https://doi.org/10.5815/ijieeb.2012.01.07

Özfidaner, M., Şapolyo, D., Topaloğlu, F., & Baydar, A. (2018). Adana İlinde Buharlaşma Serilerinde Gidişlerin Yeni Bir Gidiş Analiz Yöntemi İle Belirlenmesi. *Journal of Agricultural Faculty of Gaziosmanpaşa University (JAFAG), 34*, 59-66.

Penman, H. L. (1956). Estimating evaporation. *Eos, Transactions American Geophysical Union, 37*(1), 43-50. https://doi.org/10.1029/TR037i001p00043

Podhoranyi, M. A. (2021). Comprehensive social media data processing and analytics architecture by using big data platforms: a case study of twitter flood-risk messages. *Earth Science Informatics, 14*, 913–929. https://doi.org/10.1007/s12145-021-00601-w

Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 729-735). IEEE. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Reddy, K. V. A., Ambati, S. R., Reddy, Y. S. R., & Reddy, A. N. (2021). AdaBoost for Parkinson's Disease Detection using Robust Scaler and SFS from Acoustic Features. In *2021 Smart Technologies, Communication and Robotics (STCR)* (pp. 1-6). IEEE. https://doi.org/10.1109/STCR51658.2021.9588906

Roberts, W., Williams, G., Jackson, E., Nelson, E., & Ames, D., (2018). Hydrostats: A Python Package for Characterizing Errors between Observed and Predicted Time Series. *Hydrology, 5*(4), 66. https://doi.org/10.3390/hydrology5040066

Sarıgöl, M., & Katipoğlu, O. M. (2023). Estimation of monthly evaporation values using gradient boosting machines and mode decomposition techniques in the Southeast Anatolia Project (GAP) area in Turkey. *Acta Geophysica*, 1-18. https://doi.org/10.1007/s11600-023-01067-8

Seo, J., Ma, H., & Saha, T. (2015). Probabilistic wavelet transform for partial discharge measurement of transformer. *IEEE Transactions on Dielectrics and Electrical Insulation, 22*(2), 1105-1117. https://doi.org/10.1109/TDEI.2015.7076812

Singh, A., Panda, R. K., & Pramanik, N. (2009). Appropriate data normalization range for daily river flow forecasting using an artificial neural network. *IAHS-AISH publication, 331*, 51-57.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Thara, D. K., PremaSudha, B. G., & Xiong, F. (2019). Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters, 128*, 544-550. https://doi.org/10.1016/j.patrec.2019.10.029

Wang, S. C. (2003). *Interdisciplinary computing in Java programming* (Vol. 743). Springer Science & Business Media.

Yeo, I. K., & Johnson, R. A. (2000) A new family of power transformations to improve normality or symmetry. *Biometrika, 87*(4), 954-959. https://doi.org/10.1093/biomet/87.4.954

Zhang, H., Zhou, J., Jahed Armaghani, D., Tahir, M., Pham, B., & Huynh, V. (2020). A Combination of Feature Selection and Random Forest Techniques to Solve a Problem Related to Blast-Induced Ground Vibration. *Applied Science, 10*(3), 869. https://doi.org/10.3390/app10030869