Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 4, 3879 doi: 10.2166/ws.2022.017

Evaluation of classification and decision trees in predicting daily precipitation occurrences

S. Samadianfard [a,*], F. Mikaeili[a] and R. Prasad[b]

[a] Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Department of Science, School of Science and Technology, The University of Fiji, Lautoka, Fiji
*Corresponding author. E-mail: s.samadian@tabrizu.ac.ir

SS, 0000-0002-6876-7182

ABSTRACT

Due to the heterogeneous distribution of precipitation, predicting its occurrence is one of the primary and basic strategies to prevent possible disasters and their damages. Hence, this study aims at evaluating the capabilities of Logistic Model Tree (LMT), J48, Random Forest (RF), and PART classification algorithms in precipitation forecasts at Pars Abad station using previous 1–4 days data of meteorological variables. So, five scenarios were considered based on the cross-correlation function and partial autocorrelation function for validation of the studied methods in the period of 2004–2019. In general, by examining the Kappa, root mean squared error (RMSE), mean absolute error (MAE) indicators, scenario number 1 using the input parameters of 1-day lag was determined as the most appropriate scenario to predict daily precipitation. Also, the obtained results showed that the PART had better performance with more than 80% accuracy in precipitation forecasting. Moreover, the most accurate performance of PART was scenario 1 with Kappa = 0.2007, RMSE = 0.3879 and MAE = 0.2856. The conclusive results indicated that by implementing classification algorithms and decision trees and using meteorological data of the previous days, daily precipitation could be predicted accurately.

Key words: lag, machine learning methods, meteorological parameters, rainfall, statistical analysis

HIGHLIGHTS

- Classification algorithms and decision tree models were tested in precipitation occurrence forecasting.
- The capabilities of Logistic Model Tree, J48, Random Forest and PART were examined using kappa and accuracy criteria.
- PART algorithm indicated more than 80% accuracy.
- By implementing classification algorithms and decision trees and using meteorological data of the previous days, daily precipitation can be predicted accurately.

1. INTRODUCTION

Precipitation is one of the most important input data to hydro climatology and hydrology systems. Precipitation studies and measurements are in most cases, necessary for the study of runoff, groundwater, flood, sediment, etc. Due to rapid population growth and increasing problems of access to drinking water in arid and semi-arid regions such as Iran, the importance of proper and providing reliable rainfall models is increasing day by day. Since precipitation depends on many factors such as temperature, humidity, and evapotranspiration, it is complicated to emulate mathematically and predict the precipitation occurrences in the temporal and spatial domain. The influence of tangible and intangible factors on event of rainfall occurrence is so high that it leads to a complex and chaotic structure. In addition to the dynamic and thermodynamic mechanisms for precipitation to occur at a suitable vertical velocity, the presence of sufficient moisture is also essential. Adequate and appropriate humidity is necessary not only during rainfall but also for a certain time. The connection between the precipitation system and humidity source must be established continuously to compensate for the humidity reduction and continues rainfall. This process will strengthen the system again, and this cycle will lead to significant rain with the help of geographical conditions. In precipitation forecasting, various methods have been proposed that these methods are divided into two general categories: dynamic and empirical models. Due to the computational complexity in dynamic methods and finding features and spatial and temporal relationships of historical precipitation data in empirical methods, in recent years intelligent systems based on data mining have been adopted in the modeling of hydrological processes, which has the potential of modeling complex nonlinear processes and can provide more accurate estimates of regional conditions using meteorological and hydrological data (Bhattacharya & Solomatine 2005).

In recent years, several studies have developed and examined different methods for predicting precipitation. Mandal & Choudhury (2014) predicted maximum daily rainfall using a series of probabilistic models and found that the probability of total monthly rainfall of more than 100 mm in summer and winter was about 75–85%. Ramsundram *et al.* (2016) compared the decision tree data-driven model with Artificial Neural Network (ANN) based rainfall prediction model. The results showed that the developed model with weakly correlated input data compared to ANN could be forecast future precipitation circumstances. Bahrami *et al.* (2017), by examining the effect of pre-processing methods on the performance of the ANN, concluded that the minimum-maximum pre-processing method has the best performance in predictions. Nourani *et al.* (2017) used a combination of decision tree and association rules in long-term rainfall forecasting, and the results showed that the confidence index was estimated to be above 60%. Dash *et al.* (2018) used artificial intelligence methods (AI) K-NN, ANN, Extreme Learning Machine (ELM) to predict seasonal rainfall; they proposed that AI approaches for predicting rainfall in both summer monsoon and post-monsoon have good potential. Diop *et al.* (2019), using a hybrid artificial intelligence model, integration of Multilayer Perceptron models (MLP) with whale optimization algorithm (MLP-WOA) predicted annual rainfall. They concluded that the accuracy of standalone MLP using MLP-WOA improved. In addition, other machine learning techniques have also been used in the prediction of precipitation. Balamurugan & Manojkumar (2019) used a machine learning-based approach in the study of short-term rain forecasting. The results implemented those machine learning methods compared with statistical methods had better results. Diez-Sierra & Jesus (2020) used atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods for predicting long-term rainfall; they demonstrated that selected hyperparameters had the most influence on the performance of machine learning methods. According to the studies conducted in precipitation forecasting, the importance of daily precipitation estimating is well understood. Mohammed *et al.* (2020) provided a comparative study among multiple linear regression (MLR), support vector regression (SVR), and lasso regression. However, the key issue is their black-box nature since only their predictions are presented, and the rules remain hidden within the black-box itself. In addition, the MLR, though it provides rules, is unable to capture the non-linearity in a data set, making it not suitable (Mohammed *et al.* 2020). Data mining is a relatively new methodology for finding meaningful relationships between large amounts of data utilizing pattern recognition techniques. Essentially these methods can be categorized in predictive and descriptive modeling, in which predictive modeling may be implemented to estimate particular numerical value. In contrast, descriptive modeling investigates data and focuses on the connections and the hidden relationship between them. Decision trees are the descriptive types of predictive modeling approaches. On the other hand, the decision tree explains its prediction in of clear-cut rules. In addition, in the decision tree, unlike neural networks, nominal data can also use. Consequently, decision tree modeling does provide an apt alternative in the prescriptive data-driven models for precipitation predictions. Decision rules, PART, K-nearest neighbor, J48, M5P, and random forest, etc. are as classification algorithms. PART regards as one of the leading classification algorithms that produce rules. By comparing data to each rule, decision trees create, and the best leaf changes to the new rule. This method combines ripper and C4.5 algorithms and is popular for classification purposes (Hussain *et al.* 2018). A literature review represents that in the field of rainfall predicting, various models such as machine learning methods, artificial neural network and, etc., with the strength, and weakness of each used. However, there are still some state-of-the-art models, such as logistic model tree (LMT), J48, random forest (RF) and PART, which rarely employee for prediction occurrence of precipitation. Therefore they need to be further compared investigated. Hence, this study aims to examine the probability of daily precipitation occurrence by implementing novel classification, and decision trees including LMT, J48, RF and PART due to the introduction of models with a straightforward and understandable structure for decision making. Although these models use very simple techniques, the field of diagnosis and prediction can work with complex methods as well as ANN models. Antecedent input parameters include precipitation (P), average relative humidity (RH_{mean}), minimum relative humidity (RH_{min}), maximum relative humidity (RH_{max}), average air temperature (T_{mean}), minimum temperature (T_{min}), maximum temperature (T_{max}), sunshine hours (−ssh) and average wind speed (S_w) from 2004–2019 used as the inputs. The daily meteorological data from Pars Abad station, as a case study, are utilized. For best feature selections, five scenarios based on cross-correlation function (CCF) and partial autocorrelation function (PACF) are developed to channel the optimal features to the respective models. To the best of the author's knowledge, the considered LMT, J48, RF, and PART models have not been applied previously for forecasting precipitation occurrence. The development, application and evaluation of the classification and decision trees algorithms as the prescriptive data-driven models for precipitation predictions are the key contribution of this study.

2. METHODS

2.1. Decision tree (DT)

DT is used for data mining and it is one of the powerful and standard tools for classification and prediction that generates a set of rules in emulating precipitations. During the model training process, the data series is subdivided into homogeneous subsets in predicting or regulating an objective variable culminating into its tree regression structure. The process of creating a tree consists of three steps: dividing the nodes and assigning the nodes to the end classes. A decision tree combines a root, a series of branches and finally leaves. LMT, J48 and RF are famous and practical decision trees.

2.2. Logistic model tree (LMT)

LMT is one of the methods for classification that formed from a combination of logistic regression (LR) and DT learning methods (Landwehr *et al.* 2005). This algorithm uses the LogitBoost algorithm to build the LR model on each node and for growing up the tree. Then CART algorithm prunes the tree wherever necessary (Breiman *et al.* 1984) and then cross-validation is implemented to detect a number of LogitBoost iteration to intercept training data, subsequently reducing overfitting. For each class C, the Logit Boost model employs least-squares fitting additive logistic regression as follows (Doetsch *et al.* 2009):

$$L_C(x) = \beta_0 + \sum_{i=1}^{F} \beta_i x_i \tag{1}$$

where F is the number of traits and β_i represents the coefficient of i^{th} ingredient in the observational vector x. Linear logistic regression can use to estimate posterior probability in leaf nodes (Landwehr *et al.* 2005) as in the following equation:

$$P(C|x) = \frac{exp(L_C(x))}{\sum\limits_{C'=1}^{C} exp(L_{C'}(x))} \tag{2}$$

Here C is the number of classes and in this equation, for the least-squares to change in proportion to $L_C(x)$, it must be $\sum_{C=1}^{C} L_C(x) = 0$. The structure of LMT showed in Figure 1.

2.3. J48

In the J48 decision tree algorithm, the predictions include if-then conditions. Root, branch, and leaf nodes are parts of a decision tree. Each inward node works on a state with several input features, while each branch appoints the results of the situation and each leaf node has a class label (Bhatia 2019). J48 produces the rules for the prediction of purpose and it is an extension of the ID_3 algorithm. J48 has been found effective in calculating missing data, pruning the decision tree, and derivation rules, etc. J48 implement using JAVA with the CART algorithm (Kaur & Chhabra 2014). The steps of this

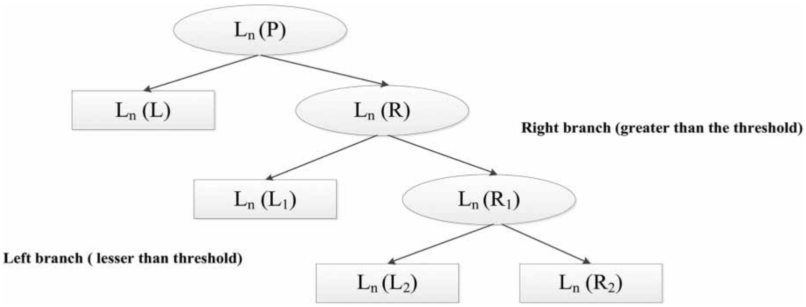

Figure 1 | Logistic model tree structure (Nachiappan *et al.* 2016).

algorithm are as follows: (i) If the examples are related to a class, a leaf is introduced which label with the same classes; (ii) The potential information is calculated. In the next step, the information gain compute; (iii) At last, the best feature is selected and it use for branching. For counting, Gain is used from Entropy and the Entropy of $\vec{y}$ is calculated by

$$Entropy\ (\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} Log\left(\frac{|y_i|}{|\vec{y}|}\right) \tag{3}$$

$$Entropy\ (j|\vec{y}) = \frac{|y_j|}{|\vec{y}|} Log\left(\frac{|y_j|}{|\vec{y}|}\right) \tag{4}$$

And Gain is

$$Gain(\overrightarrow{y,\ j}) = Entropy\ (\vec{y}) - Entropy\ (j|\vec{y}) \tag{5}$$

The aim is to maximize the gain, dividing by overall entropy due to split argument $\vec{y}$ by value j. Finally, the tree is pruned to solve the problem of overfitting.

2.4. Random forest (RF)

RF is a new decision tree method proposed by Breiman 2001 that use for supervised learning, classification, and regression. The RF model performs classification at high speeds for many datasets. Unlike classical models such as regression, which rely on one model, RF uses hundreds of trees to use more information in the data better to infer the variables (Kohestani *et al.* 2015). To create a regression tree, recursive partitioning and multiple regressions used, to prevent the adaptive of different regression trees, RF reduces the diversity of trees by creating training data, which is called bagging. Bagging is a method used to generate train data by random sampling with replacement. In addition, those samples that are not selected in the training of the trees in the bagging process include a subset called out-of-bag, which in the RF method can use to evaluate model's performance. After the formation of the trees, the average of all predicted trees is calculated to obtain the final output. The general process of the RF algorithm is illustrated in Figure 2. The formation of RF-based regression begins with the growth of

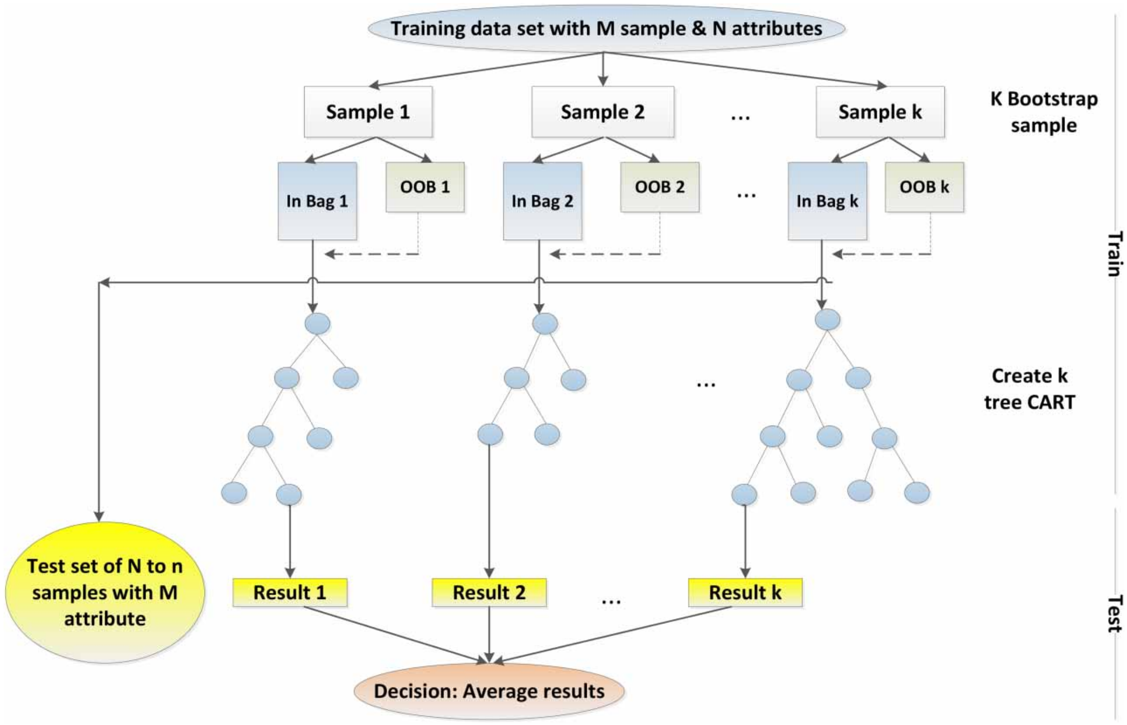

Figure 2 | Random forest structure.

the trees based on training data and random vector θ, and the result is a set of trees, k, equal to $\{h_1(x), h_2(x),\ldots, h_k(x)\}$ which here are represented as $h_k(x) = h(x,\theta_k)$, $x = \{x_1, x_{2,}\ldots,x_p\}$. These input p-dimension vectors form a forest. The group 'k' generated outputs for each tree are equal to $Y_1 = h_1(x)$, $Y_2 = h_2(x)\ldots$ $Y_K = h_k(x)$ where Y is the output of the k^{th} tree, and to obtain the final outputs, the average of all tree predictions calculated. The average square of the generalization error of each predictor h(x) is as follows:

$$E_{x\,,\,y} = (Y - h(x))^2 \tag{6}$$

As the number of forest trees increases, the error is estimated as follows:

$$E_{x\,,\,y} = (Y_{avk}h(x\ ,\ \ \theta_k))^2 \rightarrow E_{x\,,\,y}(y - E_\theta h(x\ ,\ \ \theta))^2 \tag{7}$$

2.5. PART

Essentially, the main advantage of the PART method over other methods is that combining the two set patterns, it creates rules that do not require global optimization. PART is distinct and conquers classifier algorithm. The PART algorithm produces decision lists that are used as a set of rules in the list. As new data is added, it is compared to the existing rule and the clause is transferred if the matching rule does not exist. The partial decision tree formed by part results from the combination of C4.5 and repeated incremental pruning to produce an error reduction (RIPPER) algorithm. In this method, the data set is divided back into a partial tree then the test selected and divided to subsets. The subsets are developed based on the average entropy. This process continues until a subset expands and reaches the leaf and in the next steps, other subsets that have not yet risen are selected (Frank & Witten 1998). The best leaf is introduced as a rule (Shawkat & Smith 2006). The tree-building algorithm showed in Figure 3.

2.6. Study area and data

Pars Abad is the second-largest city in Ardabil province in north-western Iran which is on the border of Iran and the Republic of Azerbaijan. Pars Abad is located in a flat plain with a warm climate. The meteorological data for this study were from Pars

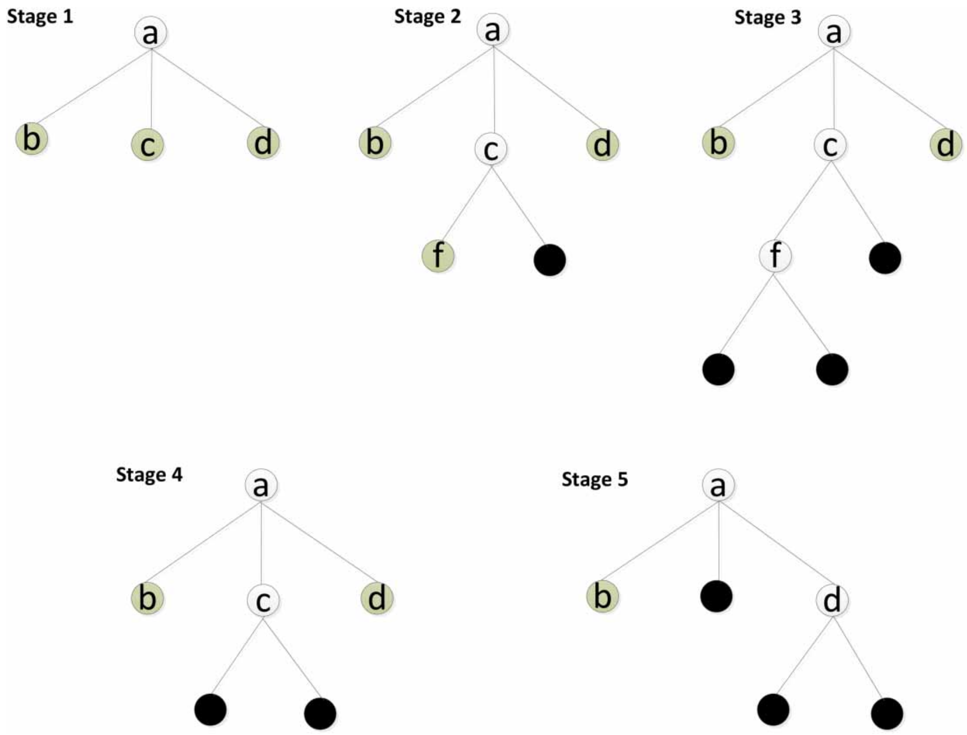

Figure 3 | Structure of PART (Frank & Witten 1998).

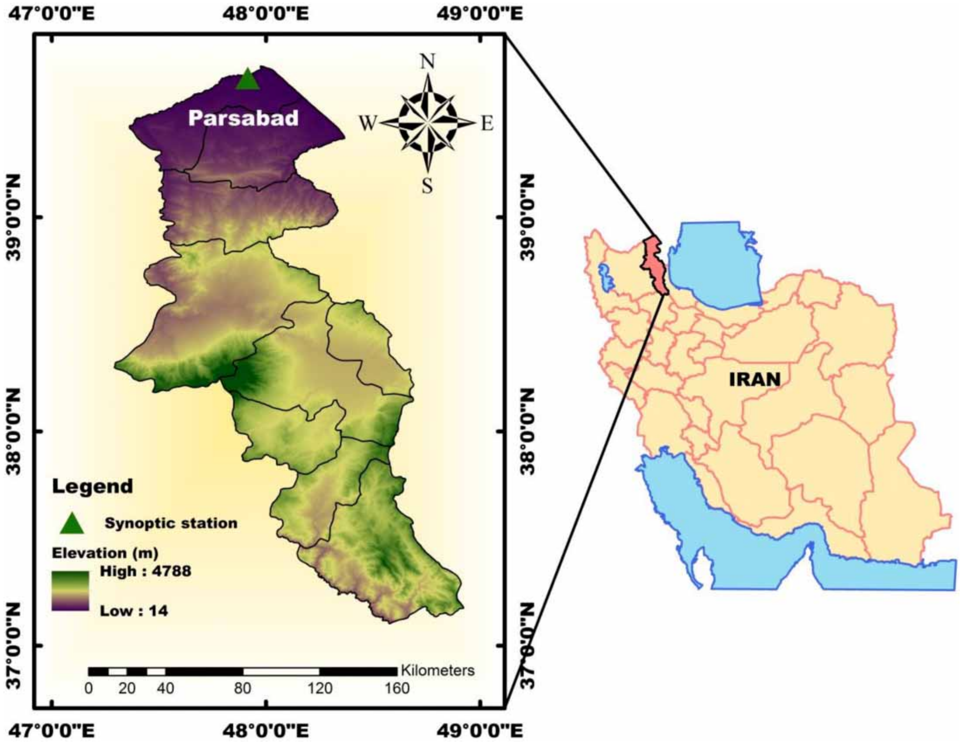

Figure 4 | Location of the study area (PARSABAD).

Abad station and the geographical location of Pars Abad station is 72.6 meters above sea level at 39° 36′ north latitude and 47° 46′ east longitude (Figure 4). It should be noted that in this study Pars Abad is only a case study. In this study, daily meteorological data of 16 years related to the period of 2004–19 were used. Daily data included precipitation (P), average relative humidity (RH_{mean}), minimum relative humidity (RH_{min}), maximum relative humidity (RH_{max}), average air temperature (T_{mean}), minimum temperature (T_{min}), maximum temperature (T_{max}), sunshine hours (−ssh) and average wind speed (S_W). The statistical characteristics of implemented meteorological parameters in Pars Abad station are shown in Table 1. Following Table 1, P has the greatest skewness. S_w also shows the skewed distribution. Other studied parameters show normal distributions since they have considerably low skewness values. The data set was divided into two parts. In the first part, 70% of the total data were used for training (2004–2015) and the remaining 30% of data were used to test the studied

Table 1 | Statistical characteristic of the meteorological inputs utilized in this study

Variable		Minimum	Maximum	Median	Mean	Coefficient of variation	Standard deviation	Skewness	Kurtosis
P	$mm.day^{-1}$	0	32	0	0.61	4.08	2.49	6.42	50.63
T_{min}	°C	−13.2	27.4	11.4	10.63	0.78	8.25	−0.18	−1.11
T_{max}	°C	−2.8	44	22.6	22.01	0.45	10	−0.16	−1.14
T_{mean}	°C	−8.2	32.2	16.4	15.8	0.57	9.02	−0.13	−1.23
RH_{min}	%	9	98	49	51.31	0.33	16.7	0.45	−0.45
RH_{max}	%	39	104	91	89.51	0.09	7.78	−1.56	3.23
RH_{mean}	%	32.25	99.75	73.13	72.33	0.17	12.16	−0.28	−0.56
−ssh	hour	0	16	7.4	6.55	0.62	4.07	−0.32	−1.18
S_w	$m.s^{-1}$	0	13.38	2.13	2.25	0.67	1.51	1.07	2.4

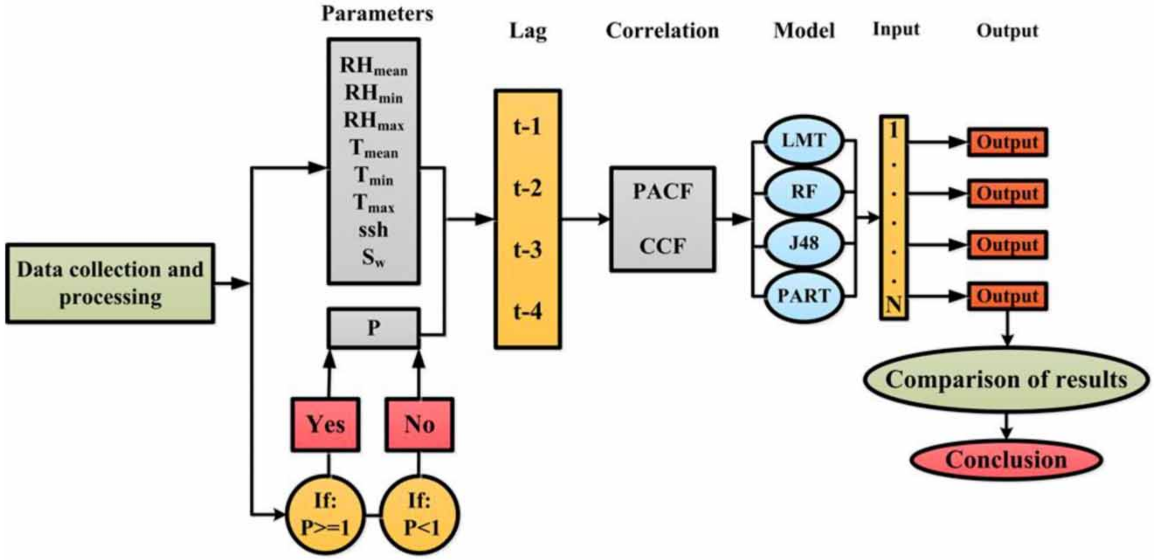

Figure 5 | The proposed method of the present study for the occurrence of daily precipitation.

models (2016–2019). Then, based on rainfall data, the days of the year were divided into two classes: rainfall days and non-rainfall days, and meteorological data were analyzed in the form of 5 series of data with delays of 1–4 days to predict the occurrence and non-occurrence of daily rainfall. The process of the present study illustrated in the Figure 5.

2.7. Model evaluations and error measurement techniques

Statistical performance measurement evaluators are derived from the confusion matrix. In binary classification, positive or negative is as the output. Table 2 shows the performance evaluation criteria for each studied method, and also, the assessment of models was measured with root mean squared error (RMSE) and mean absolute error (MAE), as shown in Equations (8) and (9).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{N} (P_j - \hat{O}_j)^2}{N}} \tag{8}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{j=1}^{N} \left| P_j - \hat{O}_j \right| \tag{9}$$

Here, P_j is the predicted value obtained from models, $\hat{O}_j$ is the observed value and N is the number of data set.

3. RESULTS AND DISCUSSION

3.1. Model development approach

In this study, the feasibility of using meteorological parameters; RH_{mean}, RH_{min}, RH_{max}, T_{mean}, T_{min}, T_{max}, P, −ssh, and S_W in different combinations to predict the occurrence of daily precipitation is tested (Table 1). To this end, we first defined five scenarios by delaying meteorological data from 1–4 days and for each scenario the input parameters were selected based on their correlation coefficient with precipitation. Moreover, for the feasibility of computations, 7–9 input parameters were chosen for each lag. For example, for the first lag, all nine input parameters were selected as scenario one, but in lag 2, instead of using 18 input parameters resulted, i.e., lags 1 & 2 for all the nine variables, based on cross-correlation function (CCF) and partial autocorrelation function (PACF), seven best inputs with higher correlations were selected to have low latency models. This process was continued for all lags and finally, 5-input combinations were established, as mentioned in Table 3. Furthermore, since the correlation of $RH_{max(t-1)}$ was high, it was used in all scenarios. Moreover, the absolute value of the correlation coefficient of meteorological parameters was the criteria used for selecting input parameters. As a

Table 2 | Statistical criteria used in this study

Statistical parameter	Equation	Definition
True Positive Rate (TPR)	$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$	A ratio of the actual positives which are correctly classified as a given class
False Positive Rate (FPR)	$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$	Proportion of wrong positive prediction to quiescent
Precision	$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$	Reflects the percentage of the total set of test records that are correctly classified
Recall	$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$	The efficiency of the algorithm according to the number of events in that class
F-measure	$\mathrm{F - measure} = (2 \times \mathrm{Precision} \times \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall})$	Used in cases where no special importance can be attached to precision and recall in relation to each other. In other words, the harmonic mean is between precision and recall
Matthews Correlation Coefficient (MCC)	$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{FP} - \mathrm{FN} \times \mathrm{FP})/\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}$	A criterion modality of binary classifications and indicate randomness of model
Receiver Operating Characteristic (ROC)	Area under the curve	Plot of TPR vs. FPR
Precision Recall Curve (PRC)	Area under the curve	Plot of precision vs. recall
Accuracy	$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$	The percentage of samples that are correctly classified with respect to all samples
Kappa	$\mathrm{Kappa} = (\mathrm{Total\ accuracy} - \mathrm{Random\ Accuracy})/(1 - \mathrm{Random\ Accuracy})$	a randomly adjusted criterion for matching between classification and correct classes

result, five scenarios were developed based on the CCF and PACF for calibration and validation of the studied methods (Figure 6).

Because there is not any direct method for dividing the training and testing data into data-driven approaches, scholars have used different proportions, such as 67% of data for training (Qasem *et al.* 2019; Kargar *et al.* 2020), or 70% for training (Asadi *et al.* 2020; Samadianfard *et al.* 2020) and as high as 80% for training (Zounemat-Kermani *et al.* 2019) has been used. Therefore, for developing the LMT, J48, RF, and PART algorithms, the data were divided into 70% training data (2004–2015) and 30% testing data (2016–2019). Then, based on rainfall data, the days of the year were divided into two classes: rainfall days and non-rainfall days, and meteorological data were analysed in the form of 5 series with delays of 1–4 days and the combination of parameters in each scenario was done based on the cross-correlation function (CCF) and partial autocorrelation function (PACF) for calibration and validation of the studied methods to predict the occurrence or non-occurrence of daily rainfall (Table 3). The input parameters in Table 3 were selected based on the higher values of correlation coefficients in different lags. In other words, while considering different lags, 7–9 meteorological parameters with higher correlation coefficients were selected as input parameters. Moreover, the computations were completed with Weka software, which contains a large number of machine learning and data mining techniques that will enable the comparison of machine learning techniques and allow users to access features such as visualization and analysis of many data mining algorithms (Witten & Frank 2000).

3.2. Discussion

Our motivation for conducting the present study is to introduce the capability of classification and decision tree models to anticipate the occurrence of daily rainfall with the help of effective meteorological variables in the form of multi-day lags.

Table 3 | Modeling of defined scenarios

	Input variables																																			
	RH_{min}				RH_{mean}				RH_{max}				T_{min}				T_{mean}				T_{max}				P				SS				U			
No.	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4	t – 1	t – 2	t – 3	t – 4
1	■				■				■				■				■				■				■				■				■			
2	■	■			■	■			■																■				■							
3	■	■			■	■			■												■				■				■							
4	■	■			■	■	■		■												■				■				■							
5	■	■			■	■		■	■												■				■				■							

The daily scale data of meteorological station, namely Pars Abad was used to expand LMT, J48, RF, and PART methods. The evaluation of the decision tree-based modeling approaches in predicting daily precipitation levels is presented here. Table 4 explains the details of the criteria for implemented LMT, J48, RF, PART, and methods in predicting precipitation conditions

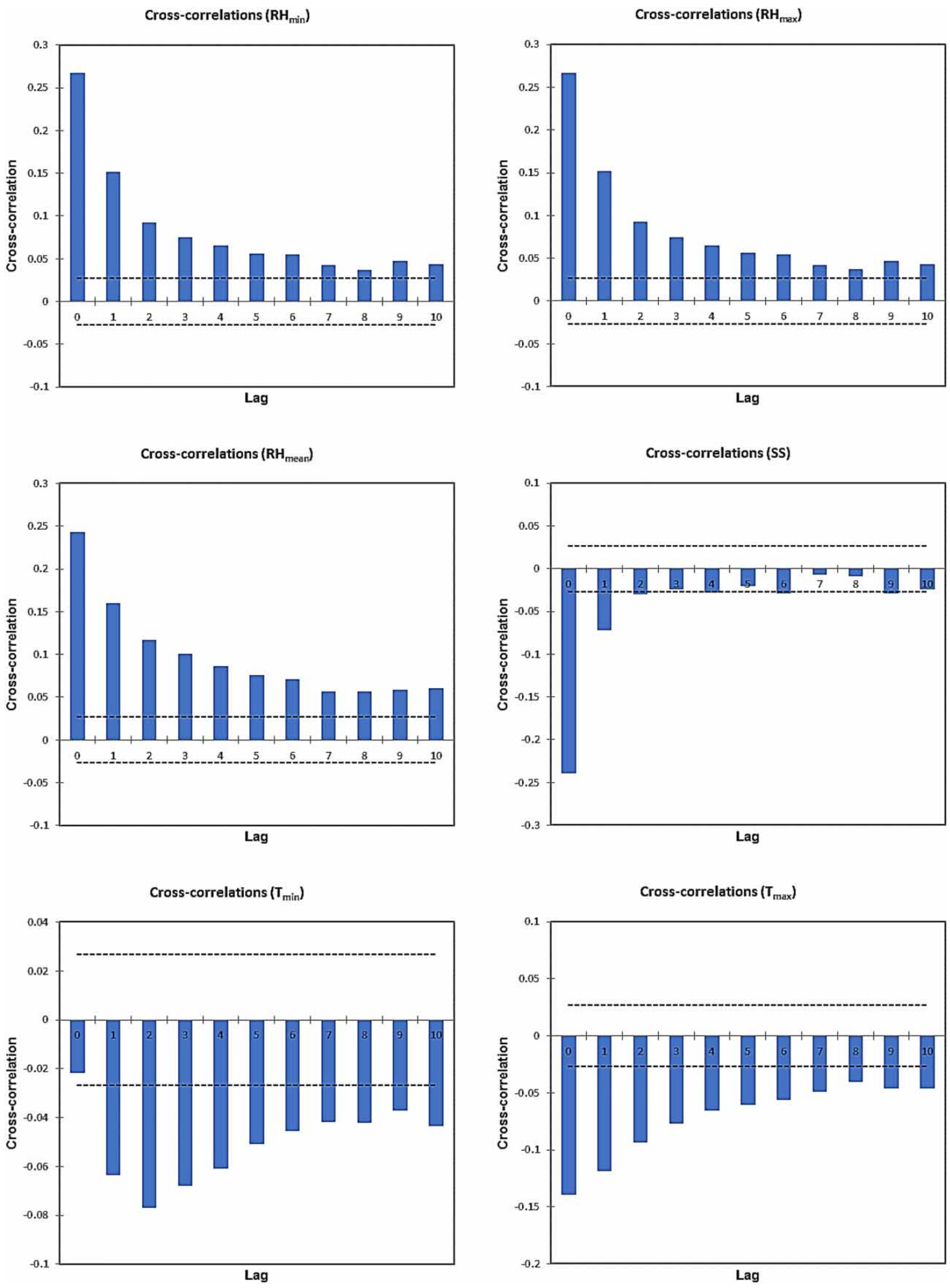

Figure 6 | Partial autocorrelation function (PACF) plots of precipitation and cross-correlation function (CCF) plots between meteorological parameters and corresponding precipitation. (*continued.*).

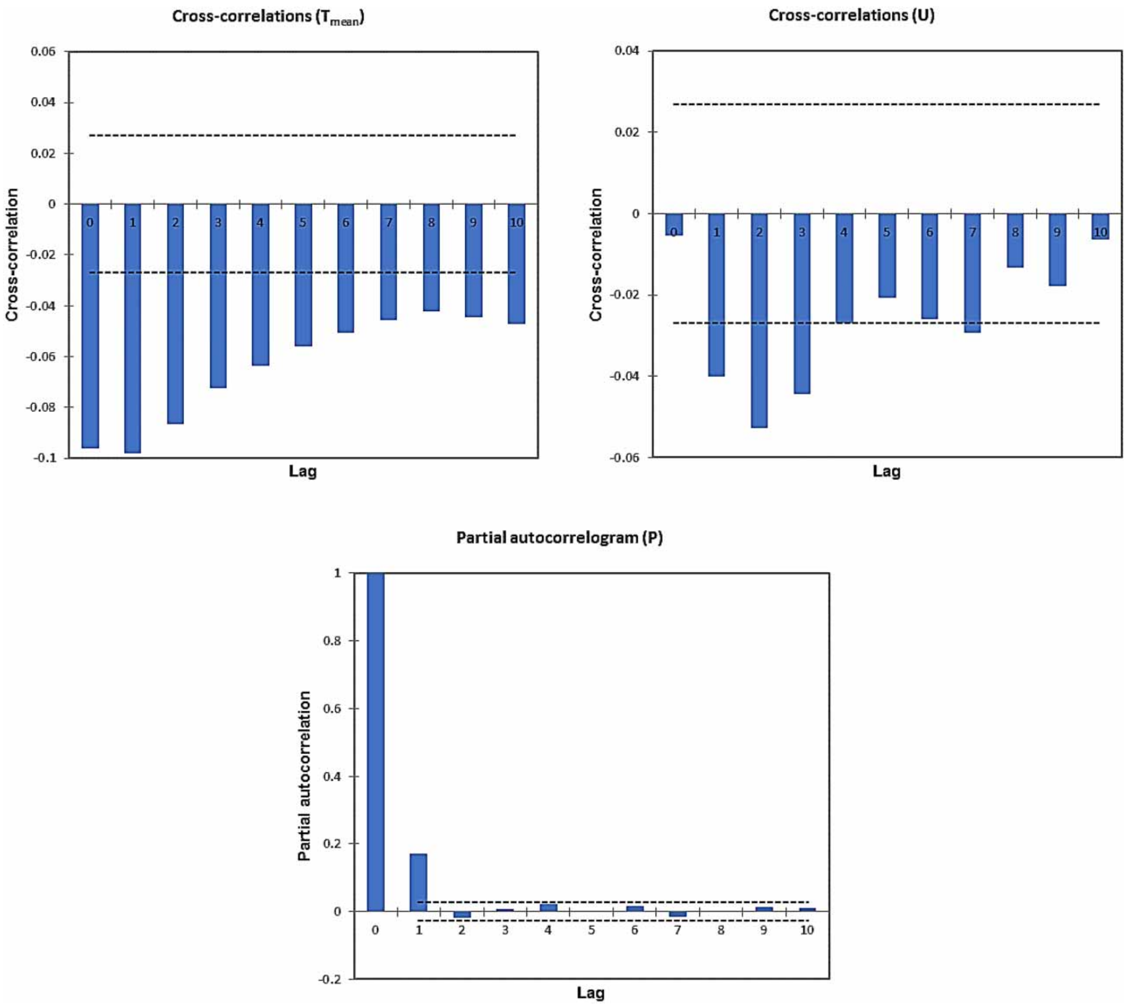

Figure 6 | Continued.

based on a series of daily meteorological data. It reveals that in Scenario 1 with one day lag, PART method with weighted average of precision = 0.770, Recall = 0.806, F-measure = 0.764, MCC = 0.242, AUC = 0.681 and PRC = 0.759 had the best performance compared to the other three methods. In the same scenario, RF with precision = 0.741, Recall = 0.784, F-measure = 0.750, MCC = 0.180, AUC = 0.691 and PRC = 0.773 was recognized as the method with the weakest performance. Then, in Scenario 2 and 3, the PART method had the highest precision and Recall, but in Scenario 4 and 5, the J48 method had the highest precision and again, the PART had the most Recall. In general, the highest and lowest precision was in Scenario 3, the PART method registered precision = 0.788, achieved the highest precision, and LMT with precision = 0.732 had the lowest precision among all defined scenarios. Furthermore, Table 5 shows the results of the utilized indicators, including Accuracy (%), Kappa, RMSE, and MAE for all studied methods. The results of different algorithms used in this study to predict the occurrence or non-occurrence of precipitation showed in Table 5. According to the results in Table 5, the PART method had the highest accuracy of the patterns used in precipitation predictions, while the J48 method had the lowest accuracy. In Scenario 1, the PART method with Accuracy = 80.5503%, Kappa = 0.2007, RMSE = 0.3879, and MAE = 0.2856 was selected as the superior method in predicting precipitation conditions one day later. Moreover, the LMT with Accuracy = 80.4878%, Kappa = 0.1895, RMSE = 0.3845, and MAE = 0.3047 had a better performance for predicting precipitation and J48 and RF methods were in the next position to predict the probability of precipitation in scenario 1, respectively. In Scenario 3 and 5, the RF method was more suitable for predicting rainfall conditions due to its high accuracy, F- measure and MCC compared to other methods. Although daily precipitation was predicted in all methods with more than

Table 4 | Detailed accuracy of studied models

Scenario	Method	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	AUC Area	PRC Area
1	**J48**	Y	0.169	0.041	0.514	0.169	0.255	0.207	0.651	0.322
		N	0.959	0.831	0.819	0.959	0.884	0.207	0.651	0.852
	Weighted Avg.		0.799	0.670	0.757	0.799	0.756	0.207	0.651	0.744
	PART	Y	0.188	0.037	0.565	0.188	0.282	0.242	0.681	0.347
		N	0.963	0.812	0.823	0.963	0.888	0.242	0.681	0.864
	Weighted Avg.		0.806	0.655	0.770	0.806	0.764	0.242	0.681	0.759
	RF	Y	0.194	0.065	0.432	0.194	0.268	0.180	0.691	0.358
		N	0.935	0.806	0.820	0.935	0.873	0.180	0.691	0.879
	Weighted Avg.		0.784	0.656	0.741	0.784	0.750	0.180	0.691	0.773
	LMT	Y	0.175	0.035	0.564	0.175	0.268	0.233	0.703	0.389
		N	0.965	0.825	0.821	0.965	0.887	0.233	0.703	0.888
	Weighted Avg.		0.805	0.664	0.769	0.805	0.761	0.233	0.703	0.787
2	**J48**	Y	0.105	0.022	0.548	0.105	0.176	0.172	0.614	0.292
		N	0.978	0.895	0.811	0.978	0.887	0.172	0.614	0.836
	Weighted Avg.		0.801	0.718	0.757	0.801	0.742	0.172	0.614	0.726
	PART	Y	0.098	0.020	0.561	0.098	0.168	0.171	0.667	0.327
		N	0.980	0.902	0.810	0.980	0.887	0.171	0.667	0.860
	Weighted Avg.		0.801	0.722	0.759	0.801	0.741	0.171	0.667	0.752
	RF	Y	0.135	0.040	0.463	0.135	0.210	0.162	0.657	0.352
		N	0.960	0.865	0.813	0.960	0.880	0.162	0.657	0.863
	Weighted Avg.		0.792	0.697	0.742	0.792	0.744	0.162	0.657	0.759
	LMT	Y	0.040	0.012	0.464	0.040	0.074	0.087	0.687	0.370
		N	0.988	0.960	0.801	0.988	0.885	0.087	0.687	0.885
	Weighted Avg.		0.795	0.767	0.733	0.795	0.720	0.087	0.687	0.780
3	**J48**	Y	0.022	0.002	0.778	0.022	0.042	0.107	0.609	0.293
		N	0.998	0.978	0.800	0.998	0.888	0.107	0.609	0.835
	Weighted Avg.		0.800	0.780	0.759	0.800	0.716	0.107	0.609	0.725
	PART	Y	0.034	0.003	0.733	0.034	0.065	0.128	0.673	0.334
		N	0.997	0.966	0.802	0.997	0.889	0.128	0.673	0.862
	Weighted Avg.		0.801	0.770	0.788	0.801	0.721	0.128	0.673	0.755
	RF	Y	0.154	0.034	0.538	0.154	0.239	0.206	0.666	0.358
		N	0.966	0.846	0.817	0.966	0.886	0.206	0.666	0.874
	Weighted Avg.		0.801	0.681	0.761	0.801	0.754	0.206	0.666	0.769
	LMT	Y	0.037	0.011	0.462	0.037	0.068	0.083	0.689	0.373
		N	0.989	0.963	0.801	0.989	0.885	0.083	0.689	0.885
	Weighted Avg.		0.795	0.770	0.732	0.795	0.719	0.083	0.689	0.781
4	**J48**	Y	0.022	0.002	0.778	0.022	0.042	0.107	0.609	0.293
		N	0.998	0.978	0.800	0.998	0.888	0.107	0.609	0.835
	Weighted Avg.		0.800	0.780	0.795	0.800	0.716	0.107	0.609	0.725
	PART	Y	0.095	0.017	0.585	0.095	0.164	0.176	0.668	0.336
		N	0.983	0.905	0.810	0.983	0.888	0.176	0.668	0.859
	Weighted Avg.		0.802	0.724	0.764	0.802	0.741	0.176	0.668	0.753

	RF	Y	0.145	0.034	0.522	0.145	0.227	0.194	0.665	0.366
		N	0.966	0.855	0.816	0.966	0.885	0.194	0.665	0.872
	Weighted Avg.		0.799	0.688	0.756	0.799	0.751	0.194	0.665	0.769
	LMT	Y	0.043	0.011	0.500	0.043	0.079	0.098	0.688	0.373
		N	0.989	0.957	0.802	0.989	0.886	0.098	0.688	0.885
	Weighted Avg.		0.797	0.765	0.741	0.797	0.722	0.098	0.688	0.781
5	**J48**	Y	0.022	0.002	0.700	0.022	0.042	0.098	0.613	0.292
		N	0.998	0.978	0.800	0.998	0.888	0.098	0.613	0.836
	Weighted Avg.		0.799	0.780	0.780	0.799	0.716	0.098	0.613	0.726
	PART	Y	0.095	0.017	0.585	0.095	0.164	0.176	0.667	0.335
		N	0.983	0.905	0.810	0.983	0.888	0.176	0.667	0.859
	Weighted Avg.		0.802	0.724	0.764	0.802	0.741	0.176	0.667	0.752
	RF	Y	0.135	0.030	0.537	0.135	0.216	0.193	0.679	0.361
		N	0.970	0.865	0.815	0.970	0.886	0.193	0.679	0.881
	Weighted Avg.		0.801	0.695	0.758	0.801	0.750	0.193	0.679	0.775
	LMT	Y	0.043	0.011	0.500	0.043	0.079	0.098	0.687	0.371
		N	0.989	0.957	0.802	0.989	0.886	0.098	0.687	0.886
	Weighted Avg.		0.797	0.765	0.741	0.797	0.722	0.098	0.687	0.781

Table 5 | Performance indices for J48, PART, RF, LMT models based on test datasets

		MODELS			
Scenario	**Performance index**	**J48**	**PART**	**RF**	**LMT**
1	Accuracy (%)	79.8624	80.5503	78.424	80.4878
	Kappa	0.1712	0.2007	0.1619	0.1895
	RMSE	0.401	0.3879	0.3919	0.3845
	MAE	0.2831	0.2856	0.3012	0.3047
2	Accuracy (%)	78.3615	80.1126	79.237	79.5497
	Kappa	0.1999	0.1138	0.1295	0.0428
	RMSE	0.4245	0.3884	0.3942	0.3865
	MAE	0.2796	0.2856	0.2867	0.2889
3	Accuracy (%)	79.9875	80.1126	80.1126	79.5497
	Kappa	0.0313	0.0476	0.1636	0.0395
	RMSE	0.3904	0.3857	0.3917	0.3859
	MAE	0.2905	0.2828	0.2844	0.2897
4	Accuracy (%)	79.9875	80.2376	79.925	79.6748
	Kappa	0.0313	0.1135	0.1517	0.0486
	RMSE	0.3904	0.3858	0.3911	0.3859
	MAE	0.2905	0.2827	0.287	0.2896
5	Accuracy (%)	79.925	80.2376	80.05	79.6748
	Kappa	0.03	0.1135	0.1463	0.0486
	RMSE	0.3909	0.386	0.3894	0.3867
	MAE	0.2903	0.2826	0.2857	0.289

78% accuracy, the PART method was recognized as the superior one in all scenarios due to its accuracy of nearly 80%. Additionally, in scenario 2 the J48 method with Accuracy = 78.3615, Kappa = 0.1999, RMSE = 0.4245 and MAE = 0.2796 had the weakest performance comparing with other methods. So, the results indicate that the PART method, rule-based, compared to the other three methods based on tree classification, has a significant ability to estimate the days on which precipitation occurred or not. Therefore, PART, LMT, RF, and finally J48 methods have the first to fourth position, in predicting occurrence or non-occurrence daily rainfall. The results of this study are consistent with the findings of Mahtabi *et al.* (2018), who reported that using meteorological data from 1–5 days ago, daily precipitation was predicted with an accuracy of more than 80% and the performance of DT was the best. Also, the relative humidity and maximum daily temperature had the highest correlation of the occurrence of daily precipitation. The results are also agree with Agnihotri & Mohapatra (2012), which examined the forecast of daily summer monsoon rains in the Karnataka region of India and concluded that the highest correlation of daily rainfall is with relative humidity parameters and minimum daily air temperature.

To evaluate the susceptibility of the models in prediction more clearly, Figure 7 displays a detailed diagram of the weighted average accuracy of the algorithm criteria for all four methods. As shown in Figure 7, the PART method showed higher capability in predicting precipitation in most scenarios due to its accuracy of nearly 80%.

The PART method with more input data, estimates the occurrence of precipitation with higher accuracy, while the performance of this method decreases when the number of input variables is low. For future studies, it is suggested that the methods used in this study be evaluated for different areas (arid or humid) with different lags. In this study, we used only a rule-based model in predicting the occurrence of daily precipitation, which ironically provided the best results among other methods, however in general it cannot be concluded that the performance of all rule-based models will be better than models based on tree classification or not. So, it is recommended that in future studies, the performance of other rule-based models be examined.

4. CONCLUSIONS

In the current study, the daily precipitation occurrences in Pars Abad station was predicted with the application of classification and decision tree models including PART, J48, LMT, and RF using meteorological parameters. To this end, five

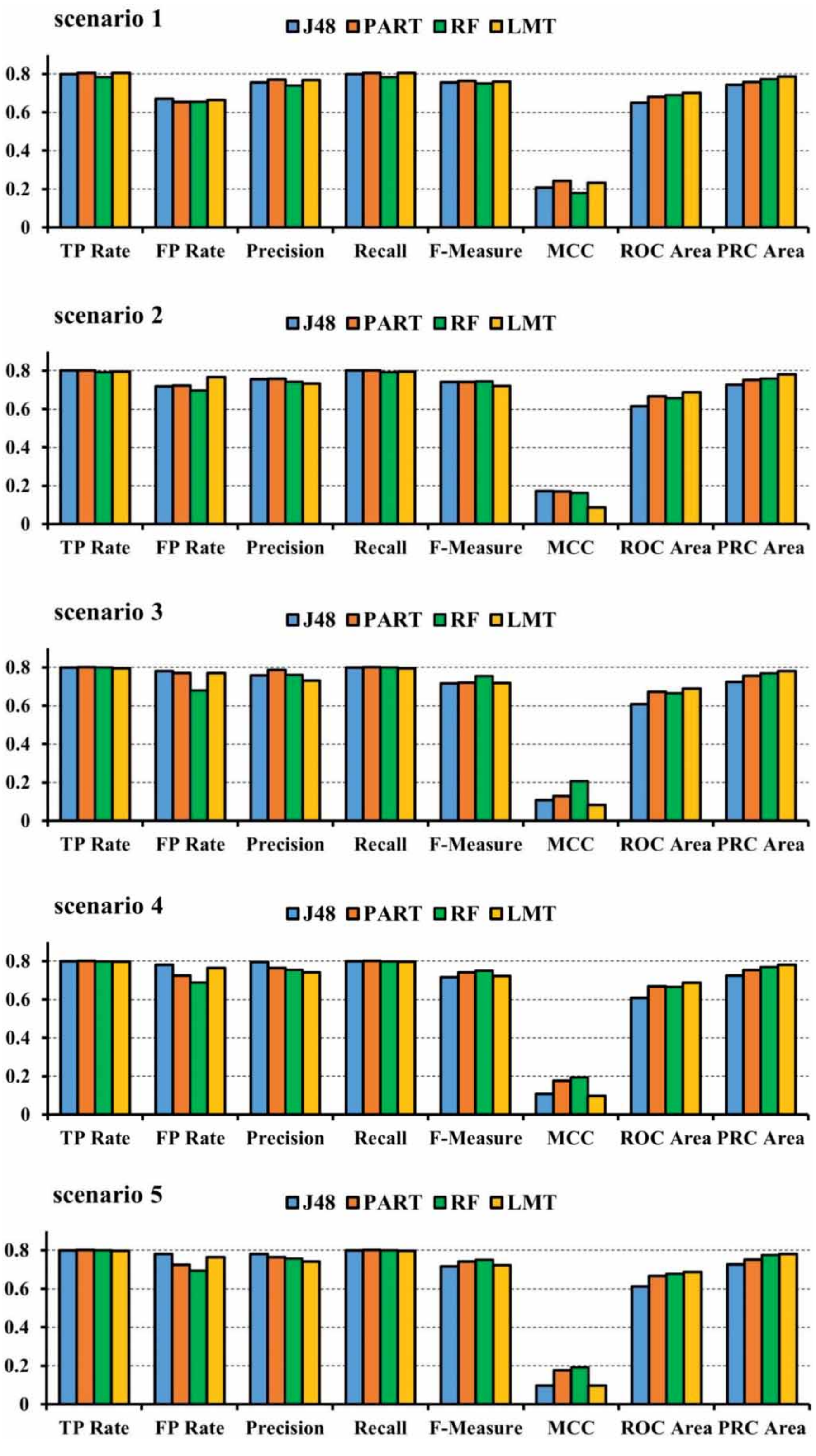

Figure 7 | Bar graphs of the weighted average of the detailed accuracy of studied models.

scenarios were defined by delaying meteorological data from 1–4 days and for each scenario the input parameters were selected based on the CCF and PACF functions. A robust performance evaluation of the models was carried out using accuracy and Kappa. Comparison of the results showed that using daily meteorological data, rainfall up to 4 days can be predicted with more than 78% accuracy. Among five selected scenarios, Scenario 1 with the PART method with accuracy of 80.5% was determined to be the best case. The performance of PART method was more suitable than the other methods, because in all five scenarios, the accuracy of the PART was higher than 80%. In all scenarios, the relative humidity and maximum daily temperature had the most significant influence in predicting daily precipitations. Conclusively, based on the obtained results, the PART method was recommended in predicting precipitation occurrences using the previous meteorological data.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Agnihotri, G. & Mohapatra, M. 2012 Prediction of occurrence of daily summer monsoon precipitation over Karnataka. *Meteorological Applications* **19**, 130–139.

Asadi, E., Isazadeh, M., Samadianfard, S., Ramli, M. F., Mosavi, A., Nabipour, N., Shamshirband, S., Hajnal, E. & Chau, K. W. 2020 Groundwater quality assessment for sustainable drinking and irrigation. *Sustainability* **12** (1), 177.

Bahrami, M., Amiri, M. J., Rezaei, F. & Gafari, K. A. 2017 Determining the effect of data pre-processing on the performance of artificial neural network in order to predict monthly rainfall in Abadeh city. *Echo Hydrology* **4** (1), 29–37.

Balamurugan, M. S. & Manojkumar, R. 2019 Study of short term rain forecasting using machine learning based approach. *Wireless Networks* **27**, 5429–5434.

Bhatia, P. 2019 *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge University Press, United Kingdom.

Bhattacharya, B. & Solomatine, D. P. 2005 Neural networks and M5 model trees in modelling water level-discharge relationship. *Neuro Computing* **63**, 381–396.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 *Classification and Regression Trees*. Belmont Wadsworth International Group, Boca Raton, CA.

Dash, Y., Saroj, K., Mishra, S., Bijaya, K. & Panigrahi, B. 2018 Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Computers and Electrical Engineering* **70**, 66–73.

Diez-Sierra, J. & Jesus, M. 2020 Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. *Journal of Hydrology* **586**, 124789.

Diop, L., Samadianfard, S., Bodian, A., Yaseen, Z., Ghorbani, M. A. & Salimi, H. 2019 Annual rainfall forecasting using hybrid artificial intelligence model: integration of multilayer perceptron with whale optimization algorithm. *Water Resources Management* **34**, 733–746.

Doetsch, P., Buck, C., Golik, P., Hoppe, N., Kramp, M., Laudenberg, J., Oberdörfer, C., Steingrube, P., Forster, J. & Mauser, A. 2009 Logistic model trees with AUC split criterion for the KDD Cup 2009 small challenge. *Proceedings of Machine Learning Research* **7**, 77–88.

Frank, E. & Witten, L. H. 1998 Generating accurate rule sets without global optimization. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. pp. 144–151.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M. & Ribata, N. 2018 Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science* **9** (2), 447–459.

Kargar, K., Samadianfard, S., Parsa, J., Nabipour, N., Shamshirband, S., Mosavi, A. & Chau, K. W. 2020 Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Engineering Applications of Computational Fluid* **14** (1), 311–322.

Kaur, G. & Chhabra, A. 2014 Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications* **98** (22), 13–17.

Kohestani, V. R., Hassanlourad, M. & Ardakani, A. 2015 Evaluation of liquefaction potential based on CPT data random forest. *Naturals Hazards* **79** (2), 1079–1089.

Landwehr, N., Hall, M. & Frank, E. 2005 Logistic model trees. *Machine Learning* **59** (1), 161–205.

Mahtabi, G. H., Taran, F. & Mozafari, S. 2018 Prediction of precipitation using meteorological data from previous days: case study of Isfahan. *Natural Geography Quarterly* **39** (11), 99–114.

Mandal, S. & Choudhury, B. U. 2014 Estimation and prediction of maximum daily rainfall at Sagar Island using best fit probability models. *Theoretical Applied Climatology* **121** (1–2), 1–11.

Mohammed, M., Kolapalli, R., Golla, N. & Sai Maturi, S. 2020 Prediction of rainfall using machine learning techniques. *International Journal of Scientific and Technology Research* **9** (1), 3236–3240.

Nachiappan, M. R., Sugumaran, V. & Elangovan, M. 2016 Performance of logistic model tree classifier using statistical features for fault diagnosis of single point cutting tool. *Indian Journal of Science and Technology* **9** (47), 1–8.

Nourani, V., Sattari, M. & Molajou, A. 2017 The combined method of decision tree and association rules in long- term precipitation forecasting. *Water and Irrigation Management* **6** (2), 331–346.

Qasem, S. N., Samadianfard, S., Sadri Nahand, H., Mosavi, A., Shamshirband, S. & Chau, K. W. 2019 Estimating daily dew point temperature using machine learning algorithms. *Water* **11** (3), 582.

Ramsundram, N., Sathya, S. & Karthikeyan, S. 2016 Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables. *Irrigation and Drainage Systems Engineering* **5**, 175.

Samadianfard, S., Hashemi, S., Kargar, K., Izadyar, M., Mostafaeipour, A., Mosavi, A., Nabipour, N. & Shamshirband, S. 2020 Wind speed prediction using a hybrid model of the multi-layer perceptron and whale optimization algorithm. *Energy Reports* **6**, 1147–1159.

Shawkat, A. & Smith, K. A. 2006 On learning algorithm selection for classification. *Applied Soft Computing* **6** (2), 119–138.

Written, L. & Frank, E. 2000 *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann Publishers, Burlington, MA.

Zounemat-Kermani, M., Seo, Y., Kim, S., Ghorbani, M. A., Samadianfard, S., Naghshara, S., Kim, N. W. & Singh, V. P. 2019 Can decomposition approaches always enhance soft computing models predicting the dissolved oxygen concentration in the St. Johns River, Florida. *Applied Sciences* **9** (12), 2534.

First received 5 October 2021; accepted in revised form 28 December 2021. Available online 19 January 2022