Weather Forecasting Using Back Propagation Feed Forward Neural Network and Multiple Linear Regression

Doğancan Ulutaş
Department of Computer Engineering
Istanbul Sabahattin Zaim University
0000-0002-5025-1161

Alaa Ali Hameed
Department of Computer Engineering
Istanbul Sabahattin Zaim University
alaa.hameed@izu.edu.tr

Erdal Alimovski
Department of Computer Engineering
Istanbul Sabahattin Zaim University
erdal.alimovski@izu.edu.tr

***Abstract*—Weather forecast is one of the most important research areas in world problems such as meteorology, human civilization, drought, agriculture and dams. We propose Back Propagation Feed Forward Neural Network and Multiple Linear Regression method and models for predicting air precipitation in the project. Proposed Neural Network model were trained with 8 optimization algorithms in order to find proper accuracy. The performance of the models in the project is evaluated with the most appropriate statistical methods. Coefficient of correlation (r), Root Means Square Error (RMSE), Mean Percent Error (MPE), Mean Absolute Percent Error (MAPE), Mean Square Error (MSE) and R-square statistics were used to measure the accuracy of the model proposed in the project. The data sets used in the project were taken from the Istanbul Provincial Meteorology System. Obtained results demonstrated that, It is seen that the model proposed in the project gives better results than the algorithms in other studies. It is seen that the model proposed in the project gives better results than the algorithms, models and techniques in other studies.**

***Keywords*— Neural Network, Multiple Linear Regression, Rainfall, Machine Learning**

I INTRODUCTION

Rainwater, which is the most important part of the world, is a very important issue in hydrology, meteorology and the future of humanity. Rain is the result of the interaction of the weather system and is affected by many environmental factors such as the ecosystem, tree density and terrain. [1]. The impact of rainfall on human civilization and agriculture from past to present is enormous. Rainfall is a difficult and important natural climate event that can be predicted days in advance for human civilization. The most important part of precipitation forecast is required for the advance planning and management of water resources and water use. It is very important that water channels and dams do not overflow. Additionally, rainfall has a strong impact on systems such as traffic, drought, sewage, landslides, floods, tsunamis and other human activities in urban areas. However, due to factors such as the complexity, disorder, continuous change of the ecosystem, the incredible diversity of atmospheric processes that make up both space and time, precipitation is one of the most complex and difficult to predict issues to understand, analyze and model. For these reasons, despite advances in many technological fields such as the increase in weather forecast models and algorithms in recent years, accurate weather forecast is becoming very difficult. Precipitation means agriculture and crops; agriculture and crop mean life. Weather forecast is very important for the agricultural sector, which contributes significantly to the country's economy. Numerous research and projects have been conducted by different researchers around the world to accurately predict weather precipitation using a variety of techniques and models. However, the prediction accuracy obtained with these studies, algorithms and techniques is also below the desired level because the precipitation does not continue in the same way.

The artificial neural network algorithm succeeds in precipitation prediction thanks to its highly nonlinear, flexibility, data-driven learning in modeling without prior knowledge of collection behavior and flow processes. [2]. Artificial neural networks have been highly preferred and successfully used in various aspects of science and engineering in recent years due to their ability to model and analyze both linear and nonlinear systems without making assumptions, as they are more successful in most traditional statistical approaches. ANN technique is frequently preferred and successfully used in all countries of the world for the prediction, recognition and classification of many weather events [3].

In this study, in order to predict the rainfall in a very effective way we propose back propagation feed forward neural network and multiple linear regression models. Proposed Neural Network model were trained with 8 optimization algorithms in order to find proper accuracy. The performance of all models to be applied is evaluated and analyzed using the most appropriate statistical methods for the project. Coefficient of correlation (r), Root Means Square Error (RMSE), Mean Percent Error (MPE), Mean Absolute Percent Error (MAPE), Mean Square Error (MSE) and R-square statistics were used to measure the accuracy of the model proposed in the project.

The remainder of the article is as follows: Part 2 explains relevant work and provides information on topics. Chapter 3 explains the data set used and details of the proposed models and algorithms. Section 4 presents all the data and results obtained in the project. Chapter 5 contains descriptions of data, estimates, results.

II RELATED WORK

There are many studies and sources for precipitation prediction in past studies. This section is where these studies are mentioned.

In study [4], researchers proposed a precipitation forecast project using artificial neural networks. The proposed model predicts air precipitation with artificial neural networks of the Udupi region in Karnataka state of India. BPNN with feed forward network structure and repeating layer architecture was tested in the project. When we look at the results, it seems that the recurrent network gives more accurate results than BPNN.

In study [5], researches in order to predict the rainfall precipitation proposed LSTM and Convolution Neural Network (CNN) models. Proposed models, Estimated monthly average rainfall data for 10368 Geo Locations worldwide for 39 Months. Obtained results shows that LSTM gets RMSE of 2.25, whereas the RMSE of CNN was 2.44.

In study [6], researches perform C4.5 algorithm in order to predict the rainfall in city Bandung regency in Indonesia. Furthermore, the final sifting method is used to optimize sifting on the model. From the results, it can be seen that the analysis positively affects the model and gives correct results. The average accuracy test result without pruning is 60% and the use of pruning is 93.33%.

In study [7], researches in order to predict the weather in New Delhi, India and Australia applied five different classification algorithms such as Extra Trees, Random Forests, Logistic Regression, Stochastic Gradient Descent (SGD) and Support Vector Machines (SVM). Obtained results shows that Random forest and Extra Trees reached the highest accuracy of 85%. Followed by SVM and Logistic Regression with an accuracy of 83%. SGD reached the lowest accuracy 82%.

In study [8], researcher proposes a hybrid DL approach, a combination of one-dimensional Convolution Neural Network (Conv1D) and Multi-Layer Perception (MLP), All step forward daily precipitation forecast from day 1 to day 5. Next, proposed hybrid model were compared with MLP and SVM. When this model is examined, it is seen that the hybrid model gives more accurate results. Looking at the models in general, the predictive data from the recommended hybrid DL model can be helpful in agricultural irrigation planning and even flooding due to heavy rainfall.

To improve and demonstrate the average prediction accuracy for short-term precipitation, the study [9] proposed a Dynamic Regional Combined short-term precipitation Prediction model (DRCF) using a Multi-Layer Sensor. Experiments were conducted on data sets taken from 56 real space meteorology sites in China. As a result of the experiments, it shows that the proposed model performs better and more successfully than the existing approaches in terms of both threat score (TS) and root mean square error (RMSE).

III MATERIALS AND METHODS

In this section, first about the data set, then about the reconstruction of the missing data and important factors such as data transformation, model selection and accuracy are explained. Secondly it is described the proposed Artificial Neural Network model. Lastly it is described the Linear Regression model.

A. Data set

Data sets used in the project were used from the Istanbul provincial meteorological system and data falling to Istanbul rainfall stations from the FreeMeteo website. In total, 365 data were taken by taking the average values of the years between 2015-2019. Data is divided into 5 branches: temperature, Max. temperature, min. the temperature is classified as water temperature and precipitation. Note that, 70% of the data set were used for training, 15% for Validation and remaining 15% for testing. The reason for separating data in this way is because these values produce values that are the most accurate and the least error value.

B. Neural Network Model

Artificial neural networks (ANN) have the ability to represent nonlinear processes and many collaborative articles have been published in this area. [10], [11], [12]. An artificial neural network consists of a simple and synchronized substance called neurons, which looks like biological neurons in the human brain. Feed forward ANN can have multiple layers. These neurons are lined up in layers in the network. Neurons in one layer are connected to the next layer. The strength of the connection between two neurons in adjacent layers is called "weight". There are weights between the links, and these weights are changed during the training process in order to produce a result close to the desired value from the input. The training rule is used to adjust the weights for this change. [13]. ANN consists of input, hidden and output layers. In a feed forward network, the weighted links feed forward from only one input layer to the output layer. Each node in a tier receives the weighted input from the previous tier, and jobs then forward its output to nodes in the next tier through links. In forward and backward feed, activation is fed backward. From output layer to hidden or input layer. Our proposed Feed forward back propagation neural network model consists of four layers; one input, two hidden and one output layer with 4,10,10,1 neurons respectively. As activation functions, in both input and hidden layers ReLu, while in output layer soft max (2) were used. Learning rate was set to 0.07. In order to find a proper accuracy, we trained our model with nine different optimization algorithms such as: Lavenberg-Marquardt (1), Bayesian regularization (2), Scaled conjugate gradient (3), Batch training, BFGS(4), Powell-Beale (5), Fletcher-Powell (6), One step scan (7), Resilient (8). Note that, all the mentioned hyper parameters above are constant for each experiment.

$$\Delta w = (J^T J + \mu I)^{-1} J^T e \tag{1}$$

$$\frac{1}{\sigma^2 D}\left[\frac{\sigma^2 D}{2\sigma^2 W}\sum_i w_i^2 + E\right] \tag{2}$$

$$\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T . g_k} \tag{3}$$

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} \tag{4}$$

$$|g_{k-1}^T g_k| \geq 0.2 \|g_k\|^2 \tag{5}$$

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k} \tag{6}$$

$$\tau_n = \tau_{n-1} - f(\tau_{n-1}) \frac{\tau_{n-1} - \tau_{n-2}}{f(\tau_{n-1}) - f(\tau_{n-2})} \tag{7}$$

$$dX = \Delta X . sign(gX) \tag{8}$$

C. Linear Regression

The method proposed in the project is based on multiple linear regression. For estimation, data have been collected from publicly available sources. 70 percent of this data is for education and 30 percent of the data is for testing. The multiple regression method is used to estimate values in the project, and this method is both a mathematical and statistical method. There is a linear relationship between variables and output values. The multiple linear regression equation is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{9}$$

The number of observations is denoted by "n". The dependent variable is y_i and the descriptive variable is x_i. β_0 and β_p are the constant y intercept and slop of descriptive variable respectively. Model error is indicated by $\in$. In the proposed model within the project, more than one meteorological parameter is required to predict the weather. For this estimate to be more accurate, Using multiple linear regression instead of linear regression gives much more successful results. The assumptions which are made by the multiple linear regression are: linear relationship between the both the descriptive and independent variables, the highly correlated variables are independent variables, y_i is calculated randomly and the mean and variance are *0* and σ .

IV Experimental Results

This section shows the results of the ANN and linear regression model proposed in the project. The precipitation data used from 2009 to 2018 consists of 10959 data sets. The data is divided into two parts: 70% for training data and 30% for test data.

In order to evaluate how close are the predictions of proposed neural network model to the eventual outcome and which optimization algorithm effects more positively to the proposed model we perform error metrics. The major error parameters we use are coefficient of correlation (r).

Root Means Square Error (RMSE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE) and R2.

The correlation coefficient is a mathematical, statistical measure of the linear relationship between two variables. The value of R is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(y_i' - \overline{y}')}{\sum_{i=1}^{n}(y_i - \overline{y})\sum_{i=1}^{n}(y_i' - \overline{y}')\ (1)} \tag{10}$$

Where y_i (mm) and y_i' (mm) represent the observed and predicted amount of weather precipitation "t", (mm) and $\overline{y}$ (mm) are the estimated and observed air precipitation, respectively, and "n" indicates the total number of data points. The "R" value ranges from [−1, 1], where -1 represents a perfect negative linear relationship. 0 indicates no linear relationship, and 1 represents a perfect positive linear relationship. The higher the "R" value is in the applied project, the higher the model performance.

RMSE is often used to measure the difference between observed, measured and predicted values. Its value is Always positive and a lower RMSE means a better model performance. RMSE is expressed as:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}} \tag{11}$$

MPE the calculated average of percentage errors when the estimates differ from the actual estimated amount. MPE can be defined as:

$$MPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{a_t - f_t}{a_t} \tag{12}$$

Where a_t "n" dictates that the estimated amount is the true value. f_t is the number of times the variable was predicted and n was the number of different times the variable was predicted.

MAPE is used to measure the efficiency, consistency and accuracy of the proposed model and algorithm in the project. Its value is calculated by eq. (13)

$$M = \frac{1}{n}\sum_{i=1}^{n}\frac{a_t - f_t}{a_t} \tag{13}$$

Where a_t is the actual value and f_t is the forecast value. MAPE is also presented as a percentage, which is the above equation multiplied by 100, according to different models. The difference between a_t and f_t is divided by the actual value a_t again. The absolute value in this solution is summed up for each point predicted and divided by the number " n". Multiplying by 100% gives the percent error.

The mean square error between observed, predicted, and computed outputs can be defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{14}$$

R-Square measures how much of the variability in the dependent variable will be regulated and explained by the model. It is expressed as:

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y})^2} \tag{15}$$

Figure 1 illustrates the Multi linear regression analysis for rainfall prediction. Blue circles represent the distribution of the normalized data. It is seen that the deviation and distribution of the values in the regression analysis progresses in the plane and the precipitation values diverge according to the months.

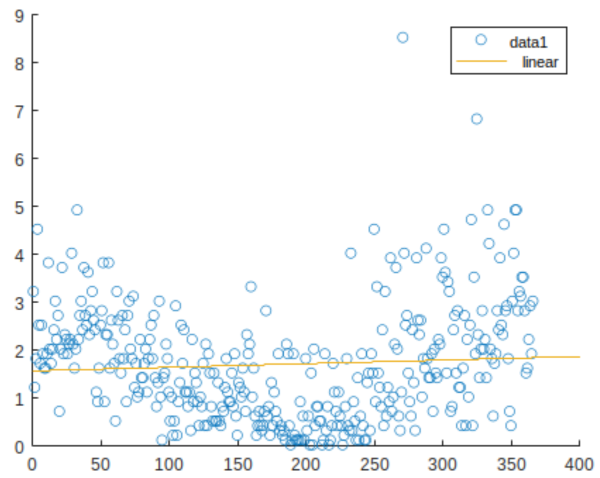

Figure 1: Multi Linear Regression plot.

A. Model Performance

The performance of proposed model with different optimization algorithm in terms of different evaluation metrics is described in Table 1.

First in Table 1. Proposed model is compared among different optimization algorithms. In terms of MPE, FFNN with Scaled Conjugate Gradient performed worst with 2.11, whereas FFNN with Powell-Beale performed the best with value of 1.69. In terms of MAPE, FFNN with BFGS achieved the best value 0.49, whereas FFNN with Powel Fletcher the worst value 0.61. In terms of MSE, FFNN with Powell-Beale performs well with value 0.10, whereas FFNN with One step scan achieved the worst result. In terms of RMSE, FFNN with Powel Fletcher performed the best with 0.18, whereas Scaled Conjugate gradient performed worst with 0.34. In terms of R-square evaluation metric, Lavenberg-Marquardt performed best, whereas Fletcher Powell performed worst. From the Table 1, it can be concluded that Lavenberg-Marquardt and Powell-Bale algorithms achieved the best results.

Secondly, the performance of proposed model with different optimization algorithms in testing phase is demonstrated in Figure 2. In terms of R, Lavenberg-Marquard algorithm gets the best results with 0.38, following by Bayesian Regularization with 0.43.

Next, we select three optimizations to evaluate the error rate in training, validation and testing tasks of the proposed model in the project: Lavenberg-Marquardt, Bayesian Regularization and Scaled Conjugate Gradient. This experiment was performed in Matlab's NF-tool. Note that, In the project, 70% of the data set was used for training, 15% for verification and 15% for testing. It was observed that these rates gave more accurate results. Obtained results in Figure 3 demonstrates that FFNN with Lavenberg-Marquardt (Lavenberg-Marquardt based FFNN) achieved the best result.

Following, Figure 3 and 4 depict how different optimization algorithms effects to the proposed model in testing phase in terms of MSE.

Table 1: Comparison of optimization algorithms in terms of different evaluation metrics.

Optimization Algorithms	Evaluation metrics				
	MPE	MAPE	MSE	RMSE	R2
Lavenberg-Marquardt	1.95	0.56	1.14	0.28	0.3
Bayesian Regularization	1.96	0.52	0.68	0.29	-0.2
Scaled Conjugate Gradient	2.11	0.58	1.15	0.34	-0.1
Batch training	1.99	0.58	0.12	0.31	-0.2
BFGS	1.87	0.49	0.08	0.23	-0.3
Powell–Beale	1.90	0.54	0.10	0.25	0.3
Fletcher-Powell	1.69	0.61	0.11	0.18	-0.7
One step scan	1.81	0.55	0.87	0.23	-0.6
Resilient	1.76	0.57	0.19	0.22	-0.6

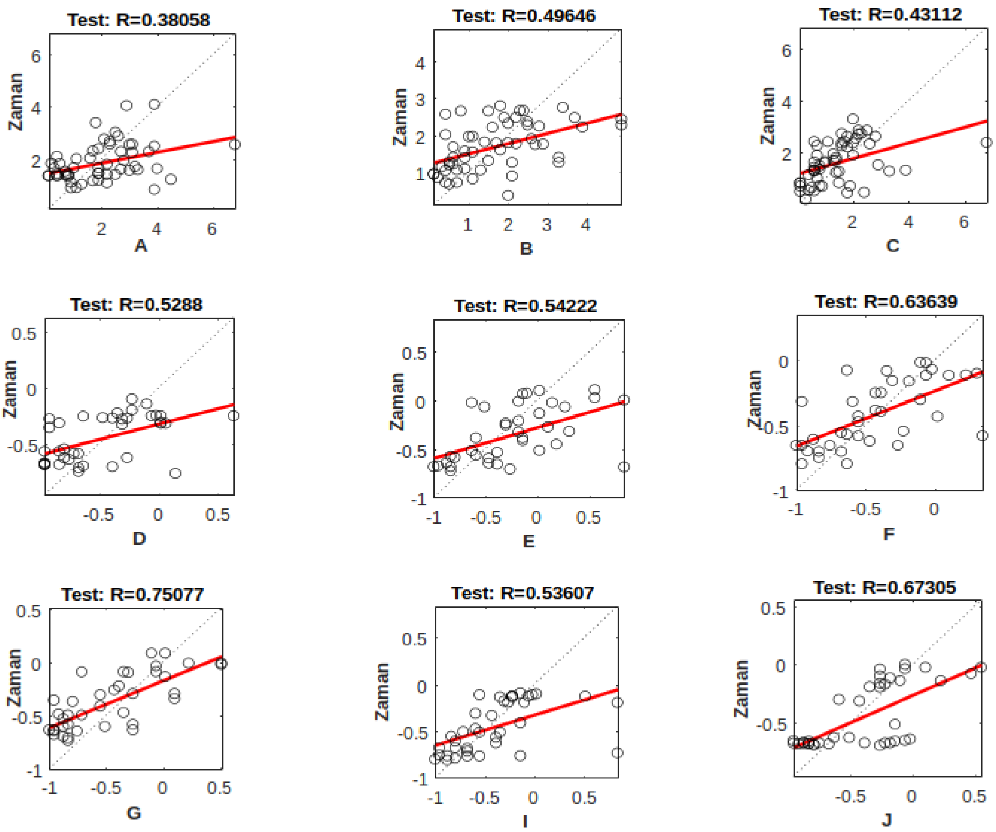

Figure 2: Proposed model - test performances with different optimization algorithms.

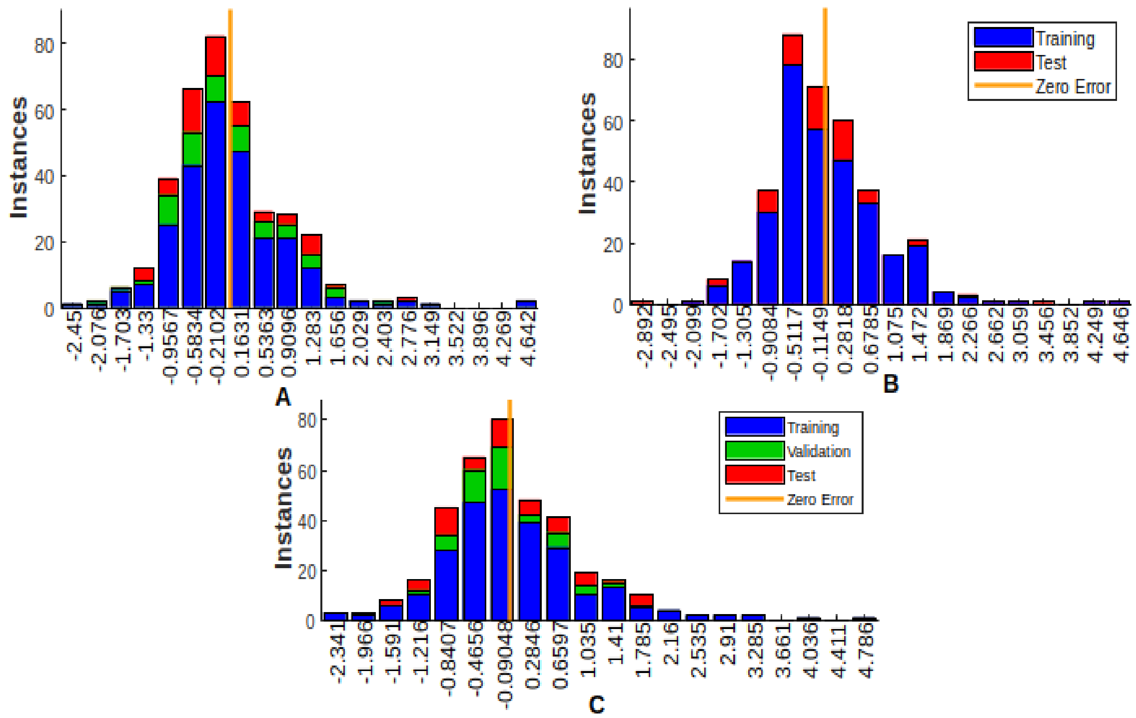

Figure 3: Comparison error performance of proposed model with different optimizer.

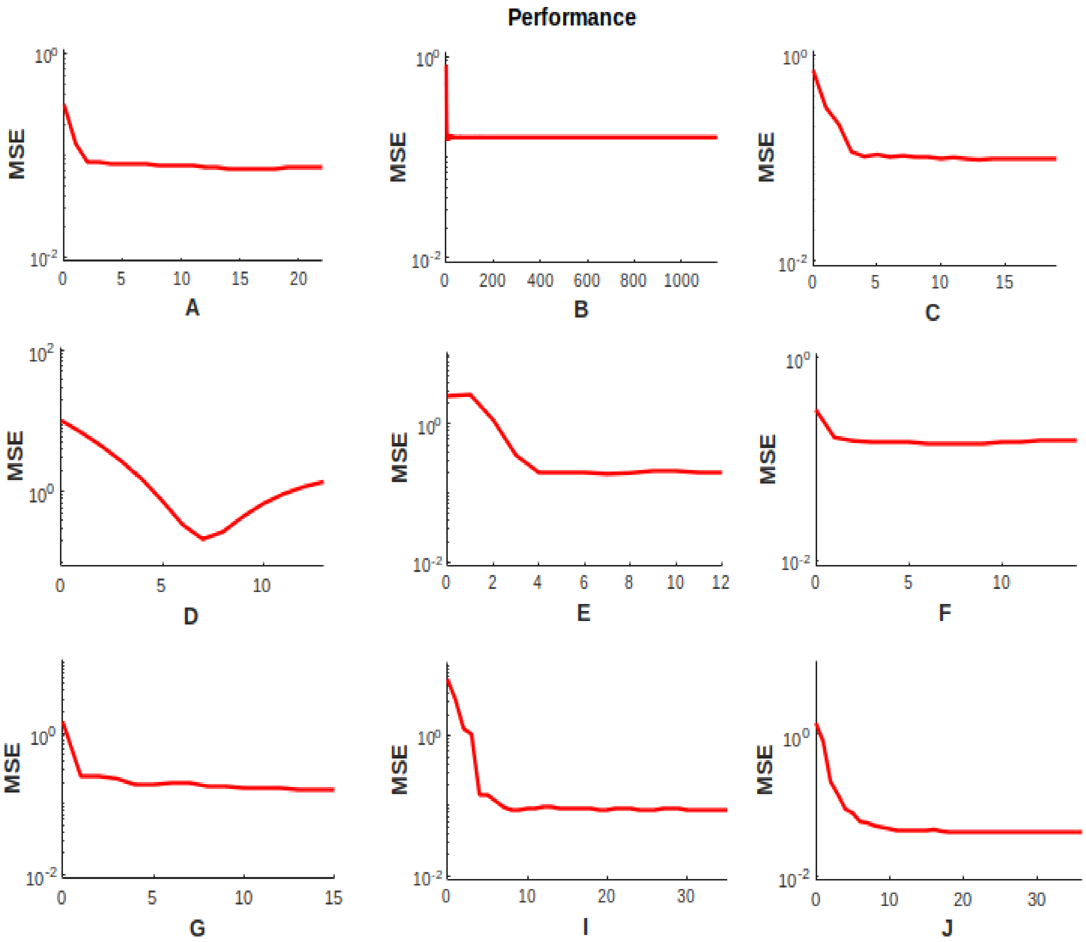

Figure 4: Performance of proposed model in terms of MSE.

A detailed comparison of the method proposed in the project with state-of-the-art techniques is presented in Table 2., in terms of, RMSE, R^2 and. Our proposed model overwhelmed the existing techniques and achieved the highest performance, which are the best results achieved so far over rainfall prediction. In addition, unlike the values in the tables, the MPE value is 1.90 for us. The Correlation value is 0.65 and the MAPE value is 0.54. Other studies: [4] MSE value 0.42. [5] MSE value 2.55, MAPE value 1.68. [7] The MSE value is 0.42. [8] 5.85 RMSE value. [14] RMSE value of 2.55. [17] MSE value 0.25. The R^2 value is 0.9. [19] RMSE value of 9.04. [20] The R^2 value is 0.94. [21] RMSE value 3.94.

Table 2: Comparison of the proposed model with state of the art techniques.

STUDIES	EVALUATION METRICS	
	RMSE	R2
Cavazos [3]	2.36	0.59
Nastos et. [15]	16.4	0.48
Hung et. [16]	1.84	0.41
Moust. et. [18]	12.5	0.24
Azadi & S.[22]	0.28	0.67
Proposed	**0.25**	**0.3**

V CONCLUSION

This article describes a back propagation feed forward neural network and multiple linear regression models to predict air precipitation in Istanbul. Turkey. Proposed neural network based model, were trained with eight different optimization algorithms in order to find best accuracy. The error measurement metrics which we use are: client of correlation (r), Root Means Square Error (RMSE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE) and R-square. Generally, proposed model achieved best accuracy with Lavenberg-Marquardt optimizer. From the obtained results, it can be concluded that proposed models in terms of error metrics achieved better performance compared with the other algorithms in the literature.

REFERENCES

[1] M. D. Crown, “Validation of the NOAA space weather prediction center’s solar flare forecasting look-up table and forecaster-issued probabilities,” Space Weather, vol. 10, no. 6, pp. 1–4, 2012

[2] A. A. Imran Maqsood Muhammad Riaz Khan, "An ensemble of neural networks for weather forecasting," Neural Computing & Applications, vol. 13, no. 2, pp. 112-122,2004.

[3] Cavazos, Tereza. "Downscaling large-scale circulation to local winter rainfall in north–eastern Mexico." International Journal of Climatology: A Journal of the Royal Meteorological Society 17.10 (1997): 1069-1082.

[4] Kumar Abhishek, Abhay Kumar, Rajeev Ranjan, Sarthak Kumar,” A Rainfall Prediction Model using Artificial Neural Network”, 2012 IEEE Control and System Graduate Research Colloquium (ICSGRC 2012), pp. 82-87, 2012.

[5] Aswin, S., P. Geetha, and R. Vinayakumar. "Deep learning models for the prediction of rainfall." 2018 (ICCSP). IEEE, 2018.

[6] Suyatno, Joko Azhari, Fhira Nhita, and Aniq Atiqi Rohmawati. "Rainfall Forecasting in Bandung Regency Using C4. 5 Algorithm." 2018 6th International Conference on Information and Communication Technology (ICoICT). IEEE, 2018.

[7] Quinn, Brandan, and Eman Abdelfattah. "Machine Learning Meteorologist Can Predict Rain." 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2019.

[8] Khan, Mohd Imran, and Rajib Maity. "Hybrid Deep Learning Approach for Multi-Step-Ahead Daily Rainfall Prediction Using GCM Simulations." IEEE Access 8 (2020): 52774-52784.

[9] Zhang, Pengcheng, et al. "Short-term rainfall forecasting using multi-layer perceptron." IEEE Transactions on Big Data (2018).

[10] Minns AW, Hall MJ (1996) Artificial neural networks as rainfall runoff models. Hydrol Sci J 41(3):399–418

[11] Sudheer KP, Gosain AK, Ramasastri KS (2002) A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrol Process 16:1325–1330

[12] Senthil Kumar AR, Sudheer KP, Jain SK, Agarwal PK (2005) Rainfallrunoff modelling using artificial neural networks: comparison of network types. Hydrol Process 19:1277–1291

[13] Ajmera TK, Goyal MK (2012) Development of stage discharge rating curve using model tree and neural networks: an application to peachtree creek in Atlanta. Expert Systems With Applications, Elsevier Ltd. 39(5):5702–5710

[14] Shah, Urmay, et al. "Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques." 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2018.

[15] Nastos, P.T., Paliatsos, A.G., Koukouletsos, K.V., Larissi, I.K. & Moustris, K.P. Artificial neural networks modeling for forecasting the maximum daily total precipitation at Athens, Greece. Atmospheric Research 144, 141 - 150 (2014).

[16] Hung, Nguyen Q., et al. "An artificial neural network model for rainfall forecasting in Bangkok, Thailand." Hydrology & Earth System Sciences 13.8 (2009).

[17] Ilaboya, I . "Performance of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) for the Prediction of Monthly Maximum Rainfall in Benin City, Nigeria" . International Journal of Engineering Science and Application 3 (2019): 21-37

[18] Moustris, K.P., Larissi, I.K., Nastos, P.T. & Paliatsos, A.G. Precipitation Forecast Using Artificial Neural Networks in Specific Regions of Greece. Water Resources Management 25, 1979 - 1993 (2011).

[19] Rajurkar, M. P., U. C. Kothyari, and U. C. Chaube. "Artificial neural networks for daily rainfall—runoff modelling." Hydrological Sciences Journal 47.6 (2002): 865-877.

[20] Riad, Souad, et al. "Rainfall-runoff model usingan artificial neural network approach." Mathematical and Computer Modelling 40.7-8 (2004): 839-846.

[21] Sahai, A. K., M. K. Soman, and V. Satyan. "All India summer monsoon rainfall prediction using an artificial neural network." Climate dynamics 16.4 (2000): 291-302.

[22] Azadi, S., Sepaskhah, A.R. Annual precipitation forecast for west, southwest, and south provinces of Iran using artificial neural networks. Theor Appl Climatol 109, 175–189 (2012).