Weather Forecasting over Iraq Using Machine Learning

Israa Jasim Mohammed [*1], Bashar Talib Al-Nuaimi [2], Ther Intisar Baker [3]

[1] College of Science, University of Diyala, Baqubah, Iraq
[2] Computer Science Department, University of Diyala, Diyala 32001, Iraq
[3] College of Science, University of Diyala, Baqubah, Iraq
Emails: scicompms2211@uodiyala.edu.iq; alnuaimi_bashar@uodiyala.edu.iq; dher@uodiyala.edu.iq

Abstract

The weather generally comprises various factors, such as wind speed, precipitation, and rainfall. Environmental weather forecasting is a demanding task for researchers, and in recent years it has attracted much study attention. Our assessment considers a wide range of weather conditions across Iraq utilizing information gathered from NASA's estimate of the world's energy resources for the years 1981 to 2021. Therefore, the correct forecast of meteorological parameters is a difficult challenge due to their changing environmental conditions. Random forest, decision tree, and GBR algorithms are used for weather forecasting. A comparison among used methods is performed and the RF is achieved the best results with accuracy, MAE, MSE, R2 of 92%, 0.5, 2.45, and 0.92, respectively.

Keywords: Weather forecasting; random forest; decision tree; GBR.

1. Introduction

Global concerns have been raised about climate change and its consequences according to [1]. During the past century, temperature of earth has risen by 0.74 °C, and increased greenhouse gas (GHG) concentrations in the atmosphere due to human activities have caused global temperatures to increase by approximately 1°C. Researchers are interested in the effects of climate change because it is an unequivocal phenomenon. The air temperature is increasing due to climate change on a global scale. The term change in climate indicates a variation in the weather statistical distribution patterns due to Radiative Force variation within a certain period. Global climate change is caused by changes in the composition of the earth's atmosphere. The environment, ecosystems, and public health are being threatened by climate change. Agricultural and crop yields are highly dependent on climate elements, and changes in temperature adversely impact them [2][3].

Weather prediction is the act of predicting the weather in a specific area and at a given time. The atmosphere was initially regarded as fluid, so physical equations were used. By solving those equations numerically, the future state of the environment is predicted. Our ability to determine the weather for more than ten days is limited, but science and technology can help us improve this [4].

Instant comparisons between past weather forecasts and observations can be processed using machine learning. With the aid of machine learning, weather models may make predictions that are more accurate by better accounting for prediction errors like overestimated rainfall. The ability to anticipate temperature is crucial in various applications, such as studies relating to the environment, energy, agriculture, medicine, or other fields.[5]

Several studies presented on weather prediction, Al-Ozeer et al. [6] analysed rainfall in Iraq, where three approaches (Isohyet, Thiessen's polygon, and Arithmetic mean) were utilized for mean area rainfall estimation. In this study, eight rain stations from 2000 to 2019 were used (Sinjar, Mosul, Talafar, Zummar, Ba'ashiqah, Hatra, Rabia, and Tal Abtah). According to the Arithmetic, Thiessen Polygon, and Isohyet approaches, the area rainfall amounts were 298.94, 230.12-, and 260.10-mm. Stan et al. [1] developed climate SA system to project both the RCP4.5 and RCP8.5 scenarios for 2018 to 2080, and to evaluate both extreme and average future conditions. The relationship between the

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022

vapour pressure deficit scalar and forest productivity was determined via the MODIS gross primary productivity (GPP) method, and future climate scenarios were used to predict how GPP will change annually and seasonally. In all scenarios, monthly and yearly temperatures rise. Astsatryan et al. [7] proposed a model for air temperature prediction for the Ararat valley. Earth observation data were obtained from different meteorological stations, in addition to a large satellite database at various resolutions and frequencies. Multiple neural networks have been used to predict air temperatures over the Ararat valley for 24 hours. Using the suggested model, you can accurately predict the temperature for the next 3 and 24 hours with an accuracy of 87.31 and 75.57, which is sufficient to replace current state-of-the-art techniques. Zhang et al. [8] developed a model called (CRNN), the model consisted of convolution neural networks (CNN) and recurrent neural networks (RNN). A neural network can be used to learn the temporal and spatial correlation of temperature change based on historical data. Daily temperature data for mainland China from 1952 to 2018 are utilized as training data to assess the proposed CRNN model. The findings demonstrate that the model is capable of making temperature predictions with an inaccuracy of about 0.907°C. In this work, the random forest and gradient boosting regressor are used for Iraq weather forecasting. First, the data is processed and filled missing values then normalized to become ready for weather prediction using random forest and gradient boosting regressor. The best results achieved is 92% accuracy RF algorithm. The aim of the study is to build a model for an accurate weather forecasting based on RF, DCT, and GBR. The structure of the paper is as follows: Section 2 presents the area of study, and section 3 illustrates the tools and methods utilized. Section 4 illustrates Framework. Section 5 discusses the model results, and Conclusion is illustrated in Section 6.

2. Study Area and Data Source

Iraq is situated between latitudes 290 and 380 in the northeast of the Arab homeland and between latitudes 380 and 480 in the southwest of Asia. Its total size is approximately 437,072 km2. The landscape resembles an alluvial plain basin in Mesopotamia. In the east and north, mountains and plateau chains are the high-altitude regions, and in the southeast and southwest, there is a gradient to low elevation, as shown in figure 1 [9].

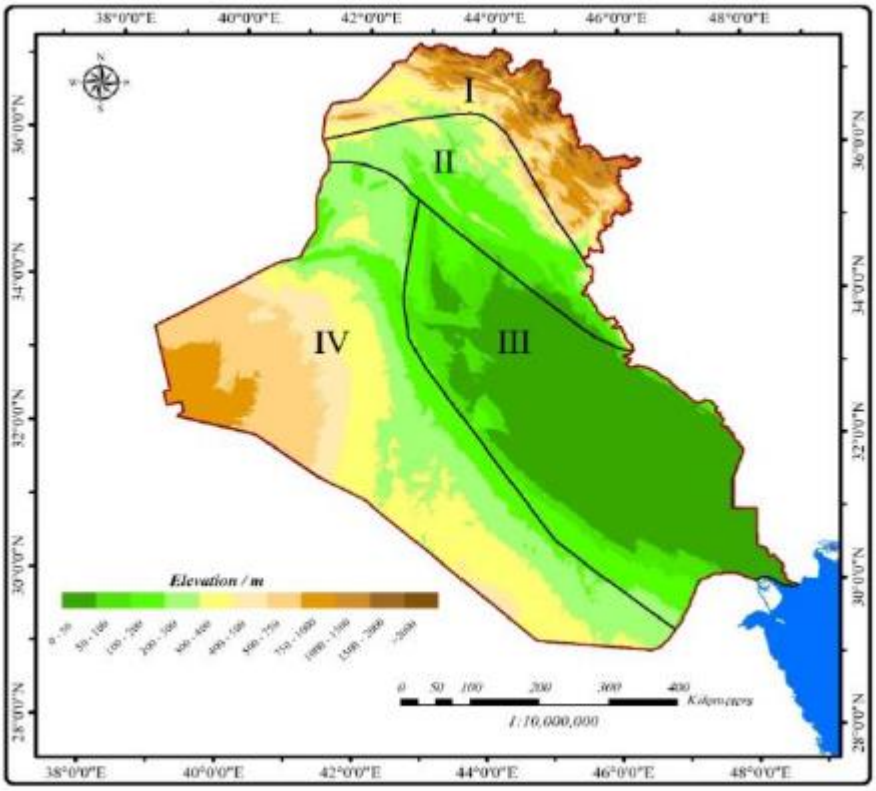

Figure 1: study area and Climate zone

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022

Iraq's geography has a significant impact on its weather. Iraq has four climatic areas, according to [10]. In Zone I, mountain chains in the north and northeast have a Mediterranean climate, which comprises about 21%; Zone II: The steppe climate of the undulating area, which includes about 9.6% of the first zone's south and west; and Zone III: The subtropical semiarid climate, which comprises about 30.2% of the central and southern regions (the Mesopotamian plain). The western continental desert climate, which makes up around 39.2% of Zone IV, is the last climate zone. The four distinct seasons in Iraq are spring (March–May), winter (December–February); autumn (October–November), which is wet and cool; and finally, summer (July–October), which is dry and hot. The average summertime temperature ranges from 27 to 31 to 41 to 45 degrees Celsius. However, in winter, temperatures range from below-freezing in the north to 4-5°C in the centre and south [6].
The observed data of the weather over Iraq has collected from the NASA prediction of Worldwide Energy Resources (https://power.larc.nasa.gov/data-access-viewer/) [11-12]. The data from 1981 to 2021 has been used to illustrate weather patterns and to calibrate projected system output for the area of study.

3. Methodology

The tools and algorithms utilized in this study is demonstrated in this section.

3.1 Gradient Boosting Regression (GBR)

A Boosting is an effective method for merging different base classifiers to create a committee whose effectiveness can be notably better than any basis classifiers. The fundamental goal of boosting is to add new models to the ensemble sequentially. The generalization of gradient boosting, or GBR, consists of three components: a weak learner (employed for making predictions), an additive model(to add weak learners to reduce the loss function), and a loss function that require an enhancement[13]. Here is how the GBR formula is expressed mathematically in more depth. Suppose the additive model form[14].

$$F(x) = \sum_{m=1}^{M} Y_m h_m(x) \quad (1)$$

Where *h*m(x) are the basis functions, which are usually called weak learners in the context of boosting. GBR builds the additive model in a forward stage-wise fashion[12].

$$F_m(x) = F_{m-1}(x) + Y_m h_m(x) \quad (2)$$

The decision tree *h*m(x) is selected to reduce the loss function L at each stage, given the current model $F_{m-1}$and its fit$F_{m-1}(x_i)$.

$$F_m(x) = F_{m-1}(x) + \arg min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \quad (3)$$

The gradient descent direction is the negative gradient of the loss function estimated at the current model Fm-1, for any differentiable loss function can be computed as follow:

$$y_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i) - h(x)) \quad (4)$$

Scalability is GBR's biggest drawback; due to boosting's sequential nature, it is difficult to parallelize. GBR also takes a long time to train.

3.2 Random Forest Regression Algorithm

A boosted estimate with improved performance is produced using the RF (Breiman, 2001) ensemble learning methodology, which combines the results of several trained weak learners (in this case, decision trees) with a voting system. This was motivated by the wisdom of the crowd process, which holds that even highly qualified individuals can analyze and solve complicated problems more effectively in a group setting than alone. RF uses random feature selection and bootstrap resampling to ensure that the weak learners are heterogeneous [15].

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022

3.3 Decision Tree Algorithm (DCT)

Decision trees (DTs) employ a binary classification method where the model inputs—such as solar radiation, external air temperature, and occupancy—are utilized in succession to identify operation modes that forecast a particular value of the model output (i.e., the heating load). First, the "trunk" is divided into the "major branches" using the primary modes of operation; if additional conditions are added, the categorization is then improved, similar to how a tree's branches, twigs, and leaves do[16-18].

3.4 Model Performance Evaluation

Mean square error (MSE), R square, accuracy, and mean absolute error (MAE) were utilized in the model evaluation stage. These statistical measures have been widely used to assess the performance of models in a wide range of disciplines, including air quality, climate research, and meteorology [19-22].

$$Mean\ square\ error = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{y_I})^2 \quad (5)$$

$$Mean\ Abslote\ error = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{y_I}| \quad (6)$$

WhereY_i {\displaystyle y_{i}}is the prediction and $\widehat{y_I}$ indicates {\displaystyle x_{i}} the true value.

$$Accuracy = \frac{\sum True\ postive + True\ negative}{\sum True\ postive + True\ negative + False\ postive + False\ Negative} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Y_i - \widehat{y_I})^2}{\sum_{i=1}^{m}(Y_i - Y`)^2} \quad (8)$$

4. Farmwork

Weather forecasting model is performed based on GBR, DCT, and RF over Iraq. First, the data is collected from the NASA prediction of Worldwide Energy Resources for the period from 1981 to 2021. The data is first analyzed and checked for missing values then normalized. The data is divided into training and testing with a percentage of 70% training and 30% for testing. figure 3 illustrates the framework diagram.

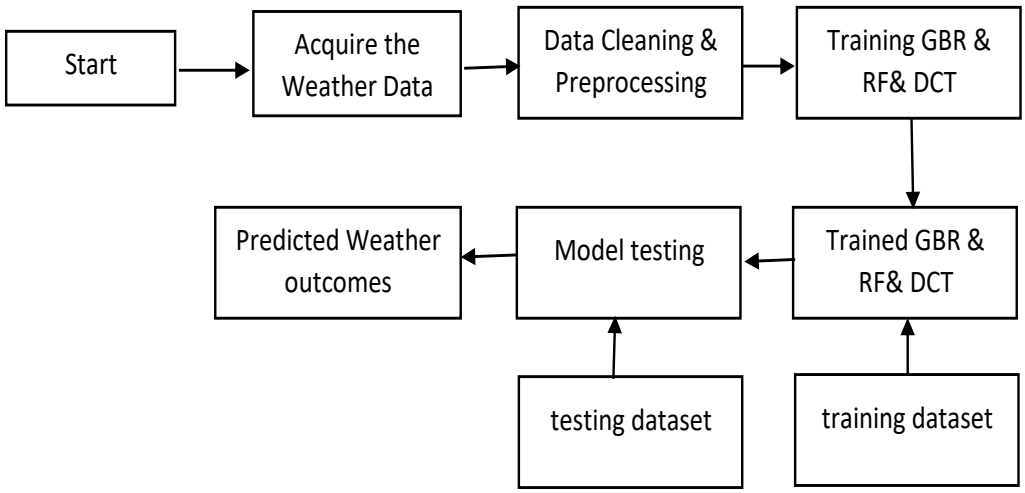

Figure 2: The Framework Diagram

5. Results and Discussion

In this work we explored the use of GBR, DCT, and RF for weather forecasting over Iraq. The data is checked for missing values. Then normalized using an equation below:

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022

$$X` = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (9)$$

The correlation between dataset features is illustrated in figure 3.

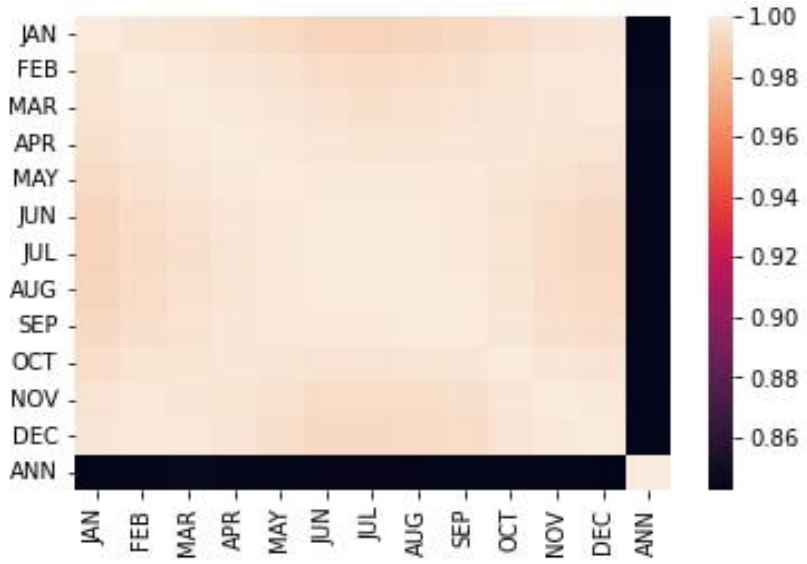

Figure 3: Correlation measure

First the weather is predicted using DCT algorithm. Different values are used for DCT max_depth parameter, the values range (2 to 10). The best results achieved in DCT with max_depth equal to eight with accuracy of 85.98%. The results of DCT algorithm is demonstrated in figure 4.

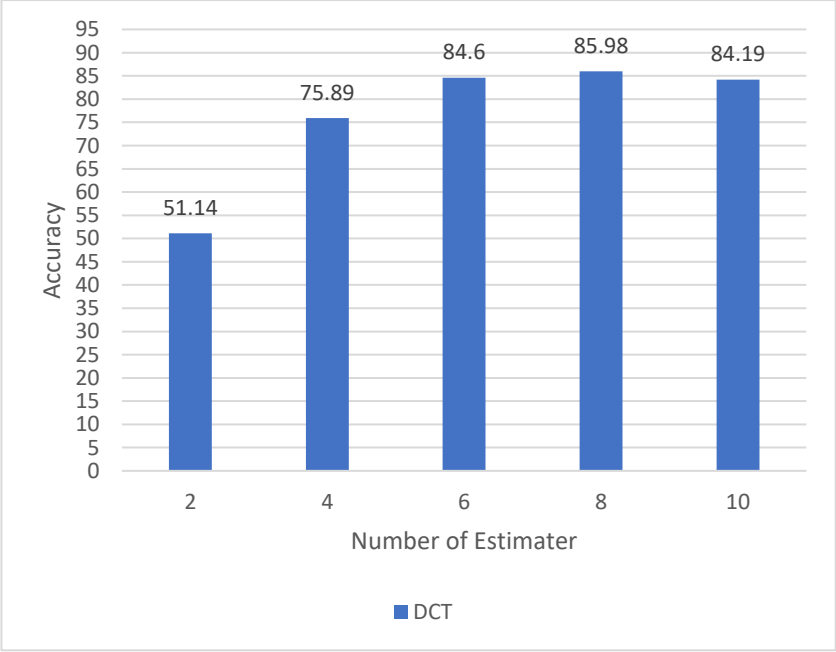

Figure 4: DCT Algorithm Results

RF algorithm with different depth is employed. RF value varies from 1 to 21 and the best result achieved with max_depth equal to 21. Figure 5 demonstrates RF results. Figure 6 demonstrates the results of GBR algorithm with a different number of estimators. The best accuracy achieved is 91.5% with 130 number of estimators.

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022

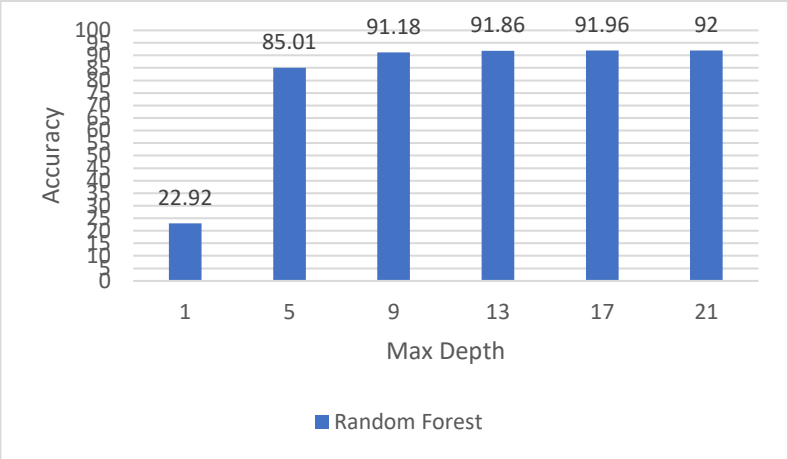

Figure 5: RF Algorithm Results

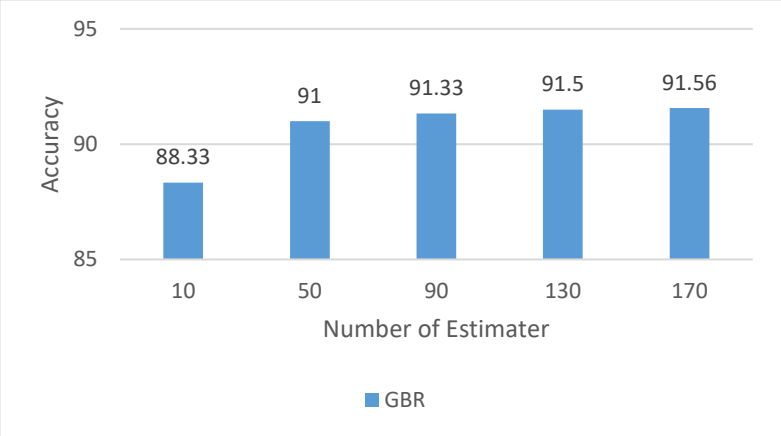

Figure 6: GBR Algorithm Results

A comparison among the GBR, DCT, and RF results are illustrated in Table 1. As illustrated in table2 RF outperformed both DCT and GBR results.

Table 1: Comparison Results of utilized Methods

Method	Accuracy	MAE	MSE	R2
GBR	91.56%	0.63	2.58	0.91
DCT	85.98%	0.49	4.8	0.8435
RF	92%	0.5	2.45	0.92

6. Conclusion

In this paper, three methods are being compared to build a precise approach for predicting weather over Iraq. The first is a Random Forest method, the second is GBR, and the third is DCT. Figures depict the outcomes of quality analysis. The given dataset is cleaned and normalized before the classification process begins in a classification method employed for accurate prediction. The results revealed that the random forest achieved the best results with an accuracy of 92%. For future work optimization algorithms such as genetic algorithm will be used for parameter tuning with DCT, RF, and GBR.

References

[1] K. Stan et al., Climate change scenarios and projected impacts for forest productivity in Guanacaste Province (Costa Rica): lessons for tropical forest regions. Reg. Environ. Chang., 20(14), 1–13, 2020.

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022

[2] C. Harkness et al., Adverse weather conditions for UK wheat production under climate change. Agric For Meteorol, 15, 282–283, 2020.
[3] Z. Li, Q. Li, J. Wang, Y. Feng, and Q. Shao, Impacts of projected climate change on runoff in upper reach of Heihe River basin using climate elasticity method and GCMs. Sci. Total Environ., 716, 137072, 2020.
[4] B. Bochenek and Z. Ustrnu, Machine Learning in Weather Prediction and Climate Analyses — Applications and Perspectives. Atmosphere (Basel)., 13(180), 1-16, 2022.
[5] G. M. Dharsan, A survey on weather forecasting and their techniques. Int. Res. J. Mod. Eng. Technol. Sci., 4(2), 12–17, 2022.
[6] N. A. Al-hammadi, ESTIMATION OF MEAN AREAL RAINFALL AND MISSING DATA BY USING GIS IN NINEVEH, NORTHERN IRAQ. Iraqi Geol. J., 53, 93–103, 2020.
[7] H. Astsatryan, H. Grigoryan, A. Poghosyan, R. Abrahamyan, and S. Asmaryan, Air temperature forecasting using artificial neural network for Ararat valley. Earth Sci. Inform., 14(2), 711–722, 2021.
[8] Z. Zhang and Y. Dong,Temperature Forecasting via Convolutional Recurrent Neural Networks Based on Time-Series Data.Complexity, 2020.
[9] T. S. Khayyun, I. A. Alwan, and A. M. Hayder, Selection of Suitable Precipitation CMIP-5 Sets of GCMs for Iraq Using a Symmetrical Uncertainty Filter. IOP Conf. Ser. Mater. Sci. Eng., 671(1), 2020.
[10] F. K. Bishay, Towards sustainable agricultural development in Iraq. The transition from relief, rehabilitation and reconstruction to development. FAO, 200, 2003.
[11] Mohamed Saber, A novel design and implementation of FBMC transceiver for low power applications. IJEEI, 8(1), 83-93, 2020.
[12] Mohamed Saber, Efficient phase recovery system. IJEECS, 5(1), 123-129,2017.
[13] C. M. Bishop, Pattern recognition and machine learning, vol. 4, no. 4. Springer New York, 2006.
[14] A. Keprate and R. M. C. Ratnayake, Using Gradient Boosting Regressor to Predict Stress Intensity Factor of a Crack Propagating in Small Bore Piping. in 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 1331-1336, 2017.
[15] D. Wolfensberger, M. Gabella, M. Boscacci, U. Germann, and A. Berne, RainForest: A random forest algorithm for quantitative precipitation estimation over Swizerland. Atmos. Meas. Tech. Discuss., 14, 1–35, 2020.
[16] N. Sirikulviriya and S. Sinthupinyo, Integration of Rules from a Random Forest Int. Conf. Inf. Electron. Eng., 6, 194–198, 2011.
[17] Doaa Sami Khafaga Hussah Nasser AlEisa, El-Sayed M. El-kenawy, Amel Ali Alhussan, Mohamed Saber, Abdelaziz A. Abdelhamid, Transfer learning for chest X-rays diagnosis using dipper throated algorithm, CMC, 73(2), 2371-2387, 2022.
[18] Marwa M. Eid, Fawaz Alassery, Abdelhameed Ibrahim, Mohamed Saber, Metaheuristic optimization algorithm for signal classification of electroencephalography channels. CMC, 71(3), 4627-4641, 2022.
[19] E. Saloux and J. A. Candanedo, Forecasting District Heating Demand using Machine Learning. Energy Procedia, 149, 59–68, 2018.
[20] N. A. Saeed and Z. T. M. Al-Ta'i, Heart Disease Prediction System Using Optimization Technique. in New Trends in Information and Communications Technology Applications, 167–177, 2020.
[21] M. Gholami Rostam, S. J. Sadatinejad, and A. Malekian, "Precipitation forecasting by large-scale climate indices and machine learning techniques," J. Arid Land, vol. 12, no. 5, pp. 854–864, 2020.
[22] D. Chicco, M. J. Warrens, and G. Jurman, "The coef fi cient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., pp. 1–24, 2021.

Doi: https://doi.org/10.54216/JAIM.020204
Received: May 22, 2022 Accepted: November 25, 2022