Precipitation Forecast for Thi-Qar Province of Iraq Utilizing Machine Learning Approaches

Anwar Alnawas[1], Nassir Jabir Al-Khafaji[2], Hayder Hussein Azeez[3]
{anwar.alnawas@stu.edu.iq[1], nassir.farhan@stu.edu.iq[2], hayder.hussein@stu.edu.iq[3]}

Nassriyah Technical Institute, Southern Technical University, Iraq[1,2,3]

Abstract. Rainfall is considered a main to provide water in rivers along with Iraqi territory. The unpredictable amount of rainfall due to climate change can cause either overflow or dry in the rivers. Although, there are a lot of electronic devices that have harnessed the prediction of precipitation using weather conditions such as humidity pressure, and temperature. Regrettably, these classical methods cannot work efficiently, so exploiting machine learning techniques can predict accurate outcomes. Therefore, predictions of data-based models using deep learning algorithms are promising for these purposes. This empirical study seeks to build a precipitation prediction model using a deep learning mechanism through utilizing historical weather data. Deep learning outperformed other classifiers based on the findings collected. The current study's experiment yielded accurate findings of up to 91.59% when testing the model with actual weather data within the specified period.

Keywords: Machine Learning, Data Mining, RapidMiner, Rainfall Rate, Thi-Qar

1 Introduction

Precipitation prediction is one of the important topics of interest to researchers in the world. Where, the rainfall is a major influence in the economic and social aspects of human life [1], such as health, agriculture, transportation, and others. For this purpose, our study found the need to build a model to understand climate indicators predict rainfall. Weather forecasting equipment such as satellites, balloons, aircraft, buoys, ground stations, and radar systems are used to gather and update atmospheric data [2]. Using this historical data provided by these tools, weather, and rainfall forecast models can be designed using a machine learning approach. One of the key steps in a machine learning approach is to extract a specific set of data that is useful for making predictions and correlations between different parameters. Therefore, data mining provides this to analyse data and devise bases for weather forecasts. Data mining techniques are separated into categorization and grouping procedures, with the data being saved and structured in a database. A data extraction method called classification is used to categorize unknown samples [3]. Using classification algorithms, it is simple to forecast precipitation [4]. Aggregation is a method of grouping objects based on information. In Iraq, agriculture is one of the most important areas affected by rain, many crops need to know the times of rain. In Iraq, agriculture is one of the most important areas affected by rain. The Ministry of Agriculture-Iraq has installed many automatic weather stations in various agricultural areas in Iraq to measure many weather factors that influence agriculture such as air temperature, relative humidity of the air, the intensity of solar radiation, atmospheric pressure, and others. Automatic weather stations

IMDC-IST 2021, September 07-09, Sakarya, Turkey

DOI 10.4108/eai.7-9-2021.2314897

provide daily data that can be used to create rain forecast models. In this study, data provided by automatic weather stations were used to construct a rain forecast model using the Deep Learning (DL) technique.
The rest of the study is organized as follows. in Section 2. , related works are explained. In Section 3. , the proposed methodology is described. Section 4. , describes how the experiments are conducted. Section 5. , the results are investigated and discussed. Finally, Section 6. , concludes the paper and remarks on key findings.

2 Related works

In present years, the utilize of machine learning algorithms for precipitation forecasting has attracted considerable attention. Drought prediction is a useful technique for analysing the negative effects of drought occurrences on important water supplies, agriculture, ecosystems, and hydrology. [5]. Therefore, a better rainfall predicting model is crucial for early warning that can reduce hazards to life and property and better manage agricultural farms. Using machine learning we can harness the historical data analysis of precipitation data and can forecasting rainfall in the coming seasons. For a comprehensive look at previous studies related to this topic, several previous studies on machine learning used in rainfall prediction are discussed.
Thirumalai, Harsha [6] discusses the rainfall rate in past various seasons based on the harvest seasons and predicts the rainfall for future years. Different types of real data were measured by linear regression method for two cropping seasons for early prediction. In this study, Kharif and Rabi were taken as the main variable where if one variable was given then the other can be predicted using linear regression. Thirumalai's study also calculated Mean and Standard deviation to predict future crop seasons. The results of the Thirumalai study help farmers to make the right decision to harvest a particular crop according to the crop seasons.
Chatterjee, Datta [7] also conducted an empirical study in India. In this paper, a Hybrid Neural Network (HNN) model is proposed to foretell rainfall rate according to several features for instance temperature, relative humidity, vapor content, and atmospheric pressure. The data has been collected by meteorological stations in West India over the period from 1989 to 1995. The findings of the present paper have suggested that feature selection can reasonably improve the performance of any classifier while predicting rainfall. In the same context, Kar, Thakur [8] have used the fuzzy logic technique to predict the rainfall, given the temperature of that particular geographical location.
Another study conducted by Moon, Kim [9], suggests a method for an effective early warning system (EWS) for very short-term heavy rainfall with machine learning techniques. The tests of EWS were run for 652 locations in South Korea from 2007 to 2012. The empirical results showed that the pre-processing methods improved the prediction quality and logistic regression works well on heavy rainfall nowcasting in terms of F-measure and equitable threat score.
Because agriculture is the backbone of the Indian economy, many studies seek to utilize sophisticated technologies to predict the climate. While agriculture is also more important to many Middle Eastern countries such as Iraq. But unfortunately, few studies take advantage of machine learning models to predict rainfall rate.
Another study related to aspects of precipitation prediction was discussed by Moulana, Roshitha [10]. They discuss a model used to predict long-term precipitation. Machine learning was harnessed for rainfall prediction purposes. The dataset was collected for rainfall prediction from the year 1901-2015 and consecutive for 3 months for each state in India. The results of this study

highlighted that the rate of rainfall is reasonably good in various states of India in the main three months (March, April, and May). Abdel-Kader, Abd-El Salam [11] ware considered as a recent study that proposed a powerful hybrid technology has been implemented to predict rainfall by combining Particle Swarm Optimization (PSO) and Multi-Layer Perceptron (MLP) which is the popular kind used in Feed Forward Neural Network (FFNN). The proposed hybrid technique has two phases, in the first phase is developed neural network by determining the number of neurons for input layer, neurons for hidden layer, and number of neurons for output layer; in the second phase PSO mainly used for automatic generation of optimized weights which used in the first phase for training network.

3 Methodology

Process for analysing weather using data extraction techniques, all standard steps that contain data selection for evaluation is performed. Throughout the process, the dataset is cleaned, formatted, and prepared for mining and interpretation. Fig 1. It shows the main steps in the proposed model for forecasting rain using the DL approach.

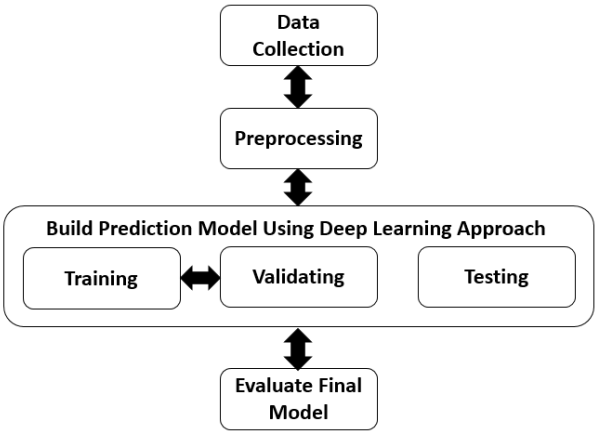

Fig. 1. The proposed model for forecasting rain using the DL approach.

3.1 Data Collection

Machine learning needs a suitable data source to build a correct prediction model. Determining the source of the data is one of the main steps to predict rainfall rate. Therefore, the data provided by the Iraqi Agrometeorological network in the Ministry of Agriculture-Iraq was used as a source of data for this study. Before making any forecasts, it is essential to understand the current weather conditions and their consequences. This process is done by examining a large amount of monitoring data.

3.2 Data Pre-Processing

After collecting the raw data, pre-processing of data aims to improve the quality of this raw data. The data provided contains 2,191 instances and 9 attributes, to predict the occurrence of rainfall. Noise reduces data quality. Therefore, noise must be removed from within the dataset using pre-processing techniques before mining tasks begin [12]. To obtain satisfactory results, a set of filters has been applied appropriately to process the data before further analysis.

3.2 Build Model

The availability of big data and the rapid development of neural network architectures, as well as increasing computational tasks, have contributed to the success of the DL concept. It pushed for its use to replace machine learning more efficiently than conventional networks. In this essential stage of methodology, DL was used to identify exact patterns in the existing dataset. Through, discovering cognitive associations, changes, and important structures from the weather dataset.

3.2 Evaluation

The evaluation stage of the prediction model is considered one of the important stages [13]. Through the evaluation, the accuracy of the model can be measured. The evaluation was based on a confusion matrix. It provides a set of evaluation factors such as Accuracy (Acc), Precision (P), and Recall (R) to evaluate data mining classifiers.

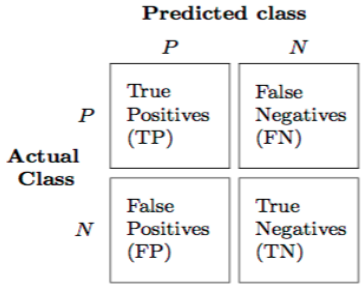

Fig. 2. Confusion matrix model.

For a binary classifier [14], lists of rates that are often computed from a confusion matrix are:

1. **Accuracy**: Indicates the correctness of the model.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. **Recall**: Indicates the sensitivity of the model or how many TPs are returned.

$$R = \frac{TP}{TP+FN} \quad (2)$$

3. **Precision**: Indicates the result of relevancy or how often predicted TP is correct

$$P = \frac{TP}{TP+FP} \tag{3}$$

4 Experimental

RapidMiner was used as the main tool for data processing and analysis. It provides an integrated environment for data collection, formatting and processing, machine learning, DL, and results in presentation. It also provides a graphical interface to facilitate access to all available tools and functions. RapidMiner is used for commercial, educational, training, validation, models improvement, and visualization of results. In this study, RapidMiner studio (V 9.9) Educational Edition was used.

4.1 Data Collection

The official website of the Iraqi Agrometeorological Network provides access to meteorological data. Many stations provide data daily. The source of the data was Souk Al-Shuyoukh station (located at 30.95° N, 46.56° E), for the period from 1/1/2014 to 31/12/2019. While completing the request through the website to obtain the data, the data is submitted in CSV format. Table 1 shows the distribution of the dataset that used in this study, both quantitatively and descriptively.

Table 1: Distribution of the dataset that used in the study.

Attributes	Description	Measuring unit	Instances	Missing data	Data type
Date	Date	Date	2,191	91	date
Rain	Rain	mm	2,191	0	real
AT Avg	Average of temperature	C°	2,191	0	real
AT Max	Maximum of temperature	C°	2,191	0	real
AT Min	Minimum of temperature	C°	2,191	0	real
RH Max	Relative Humidity	%	2,191	0	real
RH Min	Relative Humidity	%	2,191	1	real
SLR Total	Total Solar Radiation	Mj/m2	2,191	0	real
WS Avg	Wind Speed	m/s	2,191	0	real
WS Max	Wind Speed	m/s	2,191	0	real
ET	Evapotranspiration	mm	2,191	0	real

Fig 3 below illustrates the distribution of raw data and visualization of attributes that were collected at the specified periods before pre-processing. The ten major attributes were represented in Fig 3, date attribute was not used since the date scale was not suitable for this study.

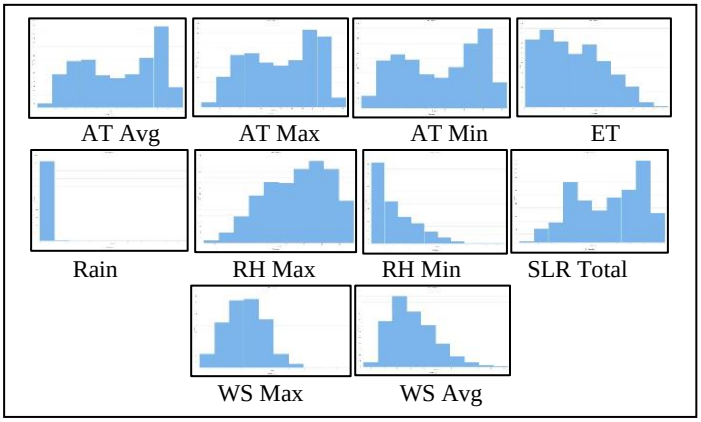

Fig. 3. The representation of attributes before pre-processing.

4.2 Data Pre-Processing

Pre-processing is an important process that reduces data inaccuracy. To obtain the best accuracy of results, the data used must be effective. In this study the following processing was used:

1. **Remove missing values:** Some instances do not have values recorded in several attributes. All instances that contain missing values are ignored.
2. **Data transformation:** Real number values changed to binary for rain attribute. Because it represents the “label” in the classification model. Where the amount of rain greater than 0 mm is represented by "1". The amount of rain equal to zero is represented by “0”.
3. **Remove duplicates:** All attributes with the same values are removed.

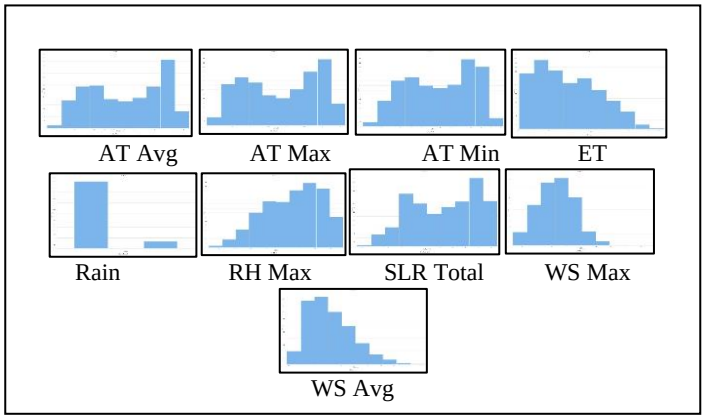

Fig . 4. The representation of attributes before pre-processing.

Fig 4 presents the attributes after pre-processing. The ten attributes have been formatted and normalized. Fig 3 and 4 show the difference in the distribution of the data representation, for example, there is a difference in the representation of rain attribute in Figs 3 and 4. The reason

for this difference is the rain theme contains missing data before processing. Datasets are divided into training and test datasets with 70:30.

4.3 Build a Prediction Model

Through several pre-processing steps, the dataset was formatted for use in building a prediction model. At this stage, DL plays an important role in bridging the gap by representing data to create a prediction model. Due to arithmetic developments and the benefit of various observations, DL has seen continuous improvement in its capabilities. Currently, Three-dimensional (3D-Var) and four-dimensional (4D-Var) variation approaches, as well as ensemble methods, are used in DL systems (commonly Kalman filter). A single deterministic state is estimated using the 3D-Var method by minimizing a cost function that includes the three terms background, observation, and model error. In addition, 4D-Var records variations in observation across time.
RapidMiner provides the ability to model prediction using layouts. The DL operator is based on a multi-layer forward-fed artificial neural network that is trained using random gradient descent by back-propagation. This network contains many hidden layers that are made up of neurons with tanh, rectifier, and maxout activation functions.
An artificial neural network can use the DL operator to select among a variety of activation functions. To be able to construct nearly any form of the artificial neural network today, all of the key training parameters are also available. The following settings have been made:

1. Activation function: Rectifier.
2. Hidden layers: 5 layers hidden by 50 artificial neurons each.
3. Epochs: 10.
4. Loss function: Automatic.
5. Max w2: 0 (maximum value of the sum of the square roots of the inputs of a neuron, set to zero allows any value without limits, by default it is set to the value 10).

5 Results and Discussion

The multiplicity of weather monitoring parameters enriches the analysis results[15]. By going through the confusion matrix, we can display the results obtained using the proposed model. Using variable epoch values to study its effect on the results, Table 2 shows the behaviour of the model using these values.

Table 2. performance results using different Epochs values

Epochs	Acc	R	P
10	0.90	0.72	0.72
20	0.90	0.71	0.71
30	0.90	0.80	0.73
40	0.90	0.66	0.72
50	0.90	0.76	0.73

60	0.91	0.70	0.74
70	0.90	0.78	0.72
80	0.92	0.73	0.77
90	0.92	0.74	0.74
100	0.92	0.73	0.76

The results above show the higher epoch value gives higher performance results for the evaluation factors Acc, R, and P. The performance of two types of classifiers were also compared with the DL approach. Which are Naïve Bayes and Decision Tree. Table 3 shows the result of the comparison.

Table 3. performance results using different classifiers

Classifiers	Acc	R	P
DL	91.59	73.39	76.1
Decision Tree	89.52	62	68.07
Naïve Bayes	67.62	72.56	58.38

Based on the results obtained compared with other classifiers (Decision Tree and Naïve Bayes) in this study, it is possible to rely on the use of DL approaches to design prediction models for weather conditions in general and rain. To ensure the validity of the prediction models, they must be applied to more than one case and different datasets.

6 Conclusion

The study of weather has become a magnet in recent years for researchers. This discipline regulates a variety of fields, including agriculture, and governments select the sorts of crops that are dependent on the state of the weather. Therefore, it is important to be aware of the weather forecast for the next few days so that measures may be taken. In this study, the DL approach was used to build a rainfall prediction model. The main steps of building a model were data collection, pre-processing, model building, and model evolution. To understand and implement the workflow, the Rapid Miner program was used to implement the proposed approach. According to the obtained results, DL achieved better results compared to other methods. The performance of the model reached an accuracy of 91.59%. The results obtained from the model help us to understand and predict the times of rain, the model can be used for transportation, agriculture, or any field affected by rain.

Acknowledgments

The authors are deeply grateful to the Iraqi agrometeorological network Ministry of Agriculture-Iraq, to provide weather data and information that used in this study.

References

[1] Anwar MT, Nugrohadi S, Tantriyati V, Windarni VA. Rain prediction using rule-based machine learning approach. Advance Sustainable Science, Engineering and Technology. 2020;2(1):1-6.

[2] Kunjumon C, Nair SS, Suresh P, Preetha S. Survey on weather forecasting using data mining. Conference on Emerging Devices and Smart Systems (ICEDSS); 2-3 March 2018; Tiruchengode, India: IEEE; 2018. p. 262-4.

[3] Thushika N, Premaratne S. A data mining approach for parameter optimization in weather prediction. International Journal of Data Science. 2020;1(1):1-13.

[4] Schultz M, Betancourt C, Gong B, Kleinert F, Langguth M, Leufen L, et al. Can deep learning beat numerical weather prediction? Philosophical Transactions of the Royal Society A. 2021;379(2194):1-22.

[5] Juneja A, Das NN. Big data quality framework: Pre-processing data in weather monitoring application. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon); 14-16 Feb. 2019 Faridabad, India IEEE; 2019. p. 559-63.

[6] Thirumalai C, Harsha KS, Deepak ML, Krishna KC. Heuristic prediction of rainfall using machine learning techniques. International Conference on Trends in Electronics and Informatics (ICEI); 11-12 May 2017; Tirunelveli, India: IEEE; 2017. p. 1114-7.

[7] Chatterjee S, Datta B, Sen S, Dey N, Debnath NC. Rainfall prediction using hybrid neural network approach. 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom); 29-31 Jan. 2018; Ho Chi Minh City, Vietnam IEEE; 2018. p. 67-72.

[8] Kar K, Thakur N, Sanghvi P. Prediction of rainfall using fuzzy dataset. International Journal of Computer Science and Mobile Computing. 2019;8(4):182-6.

[9] Moon S-H, Kim Y-H, Lee YH, Moon B-R. Application of machine learning to an early warning system for very short-term heavy rainfall. Journal of Hydrology. 2019;568:1042-54.

[10] Moulana M, Roshitha K, Niharika G, Sai MS. Prediction of rainfall using machine learning techniques. International Journal of Scientific & Technology Research. 2020;9(1):3236-40.

[11] Abdel-Kader H, Abd-El Salam M, Mohamed M. Hybrid machine learning model for rainfall forecasting. Journal of Intelligent Systems and Internet of Things. 2021;1(1):5-12.

[12] Suryanarayana V, Sathish B, Ranganayakulu A, Ganesan P. Novel weather data analysis using hadoop and mapreduce–a case study. 5th International Conference on Advanced Computing & Communication Systems (ICACCS); 15-16 March 2019; Coimbatore, India: IEEE; 2019. p. 204-7.

[13] Tharun V, Prakash R, Devi SR. Prediction of rainfall using data mining techniques. Second International Conference on Inventive Communication and Computational Technologies (ICICCT); 20-21 April 2018; Coimbatore, India: IEEE; 2018. p. 1507-12.

[14] Yadav SA, Sahoo BM, Sharma S, Das L. An analysis of data mining techniques to analyze the effect of weather on agriculture. International Conference on Intelligent Engineering and Management (ICIEM); 17-19 June 2020 London, UK: IEEE; 2020. p. 29-32.

[15] Yasmin RY, Sakya AE, Merdijanto U. A classification of sequential patterns for numerical and time series multiple source data—a preliminary application on extreme weather prediction. International Conference on Data and Software Engineering (ICoDSE); 1-2 Nov. 2017; Palembang, Indonesia: IEEE; 2017. p. 1-5.