Advanced Rainfall Classification and Pattern Analysis using Neural Networks

Kasarapu Ramani*[1], Madupuri Rajesh[2], T yasaswini[3],Vennapusa Anju Shaharun[4], Veeravalli Deep Chandu[5],Yuvaraj Duraiswamy[6]

[1]Professor, Department of Data Science, Mohan Babu University(Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
[2,3,4,5] UG Scholar, Department of Computer Science and Systems Engineering, Sree Vidyanikethan Engineering College, Tirupati, India

[6]Professor, Department of Computer Science,Chan University, Duhok, Iraq

*[1]head-ds@mbu.asia,[2]madupurirajesh@gmail.com
[3]yasaswiniyes22@gmail.com,[4]vennapusaanju2003@gmail.com
[5]deepchandu59@gmail.com, [6]d.yuvaraj@duhok.edu.krd

Abstract— Rainfall distribution serves a variety of purposes in meteorology, hydrology, and environmental science, flood forecasting, agriculture, meteorological analysis, and more around. The gathered dataset is collected from meteorological observations which helps to depict the patterns of rainfall for the area and period of study. We applied a variety of Machine Learning and Deep Learning algorithms such as Random Forest, Convolutional Neural Networks (CNN), and Multilayer Perceptron (MLP), to predict and classify rainfall using Austin weather dataset and obtain valuable conclusions from the dataset. The algorithms selected were found suitable as they were able to represent the complex trends and relationships between the temporal and spatial dimensions. From our results, the best performing algorithms in rainfall classification were Random Forest based on RMSE (Root Mean Square Error), and CNN based on classification accuracy and these two algorithms outperformed the other existing algorithms.

Keywords: Rainfall Classification, Multilayer Perceptron, Machine Learning, Random Forest, Convolutional Neural Network, Deep Learning

1 Introduction

The spatial distribution of rainfall is considered to be one important criterion that most people look within scientific disciplines such as climate, hydrology and environmental science. The studying processes in this sphere are motivated by the awareness that rainfall and its impacts on most of our activities encompassed- environment inclusive. As a result, careful endeavours are channeled to the effective gathering of data and accurate detection about alterations in precipitation patterns.

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_34

This requires the development of sound approaches from stringent pre-processing strategies to warrant quality and uniformity in all collected rainfall information. Removing inconsistencies becomes feasible through detailed standardization and cleansing methods making this dataset entirely suitable for further investigation. The environmental science is assisted in a greater degree by the deeper understanding of precipitation effects on ecosystems which gives rise to conservation actions. In agriculture it helps to make informed decisions about cropping patterns and the attendant irrigation based on accurate rainfall distribution statistics. Constant monitoring and real-time analysis are essential, keeping insights current so that they remain not only viable but constantly applicable to such scenarios as making decisions in adverse circumstances or agricultural optimization based on the principle of what kind rainfall produced presently.

2 Literature Review

To demonstrate the importance of this work, a survey was done on different research papers to understand various techniques. A few of them are discussed here.

The authors Smith, J., et al. [1] utilized machine learning approaches to classify rainfall characteristics from remote sensing data archive. Their paper aimed to investigate characteristic features superior from satellite imagery and then apply the supervised learning algorithm for binary classification by distinction rainfall intensity levels. This work lacks fine details concerning the machine learning methods and algorithms in the task of classification. The discussion of the challenges or limitations that could limit the processes of classification is also non-existent.

Chen, Q., et al. [2] put forward a new method which combines the CNN with radar data to classify rainfall types. Their paper does not have an in-depth analysis and interpretation of the Convolutional Neural Network model in connection with radar data for rainfall classification. CNN is highly efficient in predictive performance, it is also explained as "black box" model which make it difficult to understand the underlying decision-making process.

Kumar, A., et al. [3] proposed a fuzzy logic- based rainfall classification system for agricultural applications. Their system took into consideration aspects of rainfall in terms of intensity, duration and frequency therefore offering on-farm farmers details that could help them come up with operational principles for irrigation scheduling as well crop control. One of the possible shortcomings of the system proposed is that it can be applied only to limited geographical areas, while its generalizability is limited. This, in turn, could turn out in obstacles during adaptation of the system to different regions of the planet or environmental conditions where the degree of the required expert knowledge may differ or be insufficient.

Wang et al. [4] applied Long Short-Term Memory (LSTM) networks for temporal-spatial rainfall pattern analysis. The proposed model was able to simulate both estimated short and long-term temporal dependency spatial relationships in rainfall data that, made reliable predictions as well as classification of precipitation patterns over different scales of time. Their paper shows that LSTM networks are so good in temporal-spatial rainfall pattern analysis, but it lacks of both expandability and detailed methodology description, and it doesn't contain hyper parameters.

Gupta, S., et al. [5] introduced a hybrid rainfall classification model based on integrating a number of machine learning algorithms with ensemble techniques for better results in classifying the information gathered from satellite images into four different types. Through maximizing the capabilities of various classifiers, their model showed enhanced accuracy and reliability in classification of rainfall patterns. The limitation of their paper highlighted is the absence of detailed argument and reasoning for the choice and integration of specific machine learning algorithms and assemble techniques in the hybrid's rainfall classifier.
Lee, K., et al. [6] combined satellite data with numerical weather prediction models in order to classify rainfalls at global scale. Their study sought to develop detailed rainfall classification maps with adequate spatial and temporal resolution which would help in diverse environmental and climate studies. The downside of their paper noted is the obstacle of data fusion and the model fusion when the rainfall classification at a global scale is being done using the satellite data in conjunction with the numerical weather prediction (NWP) models.

Wang et al. [7] proposed a hybrid model built on the combination of Support Vector Machines (SVM) and fuzzy logic for rain classification. Their study combined the spatial and temporal characteristics obtained from weather radar information to categorize rainfall patterns precisely, highlighting that only hybrid techniques might be beneficial for classification. The hybrid model which is a combination of SVMs and fuzzy logic for rainfall classification as it's described in the paper can produces highly complex decision boundaries in high-dimensional space but can also suffer from interpretability.

3 Methodology

3.1 Weather Dataset Overview:

The dataset is collected from Kaggle named Austin weather. The meteorological features which the weather dataset is centred on provide a holistic picture of prevailing weather conditions. The dataset contains 18 attributes and 1320 records. It contains temperature, indicating the degree of warmth or coldness in the air; humidity, measuring the amount of water vapour present; dew point, representing the temperature at which air becomes saturated and condensation occurs; visibility, offering insights into the clarity of the atmosphere for observation; sea-level pressure, reflecting the weight of the air above a particular location; wind speed is one of the factors considered as it gives the intensity and direction of airflow; and precipitation giving the amount of water that falls from the sky in different forms like rain, snow, or hail.

3.2 Convolutional Neural Network Algorithm

CNN learns spatial patterns in rainfall data through convolutional layers, enabling advanced classification of rainfall patterns. The CNN have the capacity to capture spatial relationships in sequential data.

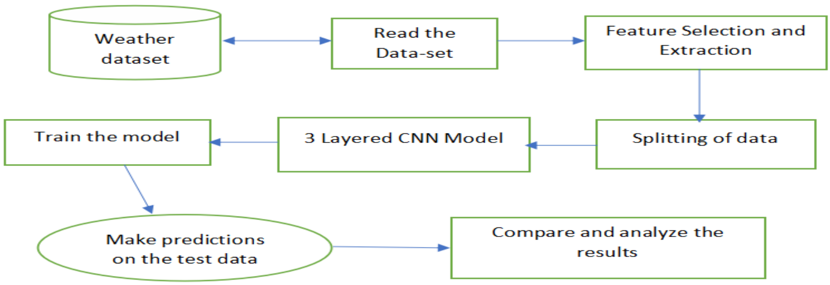

Fig.1: CNN Model Architecture

1) Data Pre-processing: The first step involves cleaning and filtering the weather dataset. We used the Standard scalar library for pre-processing the data set. Split the dataset into features which are known independent variables, and a target variable – rainfall amounts. In our study, we utilize the “Precepitation SumInches” attribute as a fundamental component for determining the accuracy and RMSE metrics. Our dataset is partitioned into training and testing subsets using a test size of 0.2 and a random state of 42.

2)Feature Selection and Engineering: Perform feature selection that finds relevant features for rainfall prediction. We used feature extraction module from sklearn library for the feature extraction. The methods include the correlation analysis, feature importance ranking, Outlier Removal, Transforming the DataFrame are used. The feature selection is performed on dew point, humidity patterns from observed data for increasing the model output.

3) Model Selection and Architecture Design: The conv1D, MaxPooling1D libraries are imported for the CNN in Fig.1 model. The structure of the CNN model consists of 3 layers namely Convolutional layer, the MaxPooling and Batch Normalization filter. With in the convolutional layer, Rectified Linear Unit (ReLU) serves as a chosen activation function denoted as Eqn. 2 in our paper. In our CNN compilation model,we used the Adam optimizer in conjunction with the mean squared error loss function. Next, it is trained on 10 epochs with 32 batch size over the training data, while 20% of the training data is used for validation.

Convolution Operation:

Convolution operation refers to filtering an input feature map by applying a kernel, also known as the process that generates output features. Mathematically, it can be represented as Eqn.(1).

$$O(i,j) = \sum_{m=0}^{F-1} \sum_{n=0}^{F-1} I(i+m, j+n) \times K(m+n) + b \quad \text{........(1)}$$

Rectified Linear Unit (ReLU):

The ReLU activation function introduces non-linearity to the CNN by applying the function:

$$f(x) = \max(0, x) \quad \text{........... (2)}$$

It zeroes all the negative pixel values thereby allowing the network to master complex patterns.

Max Pooling Operation:

Max pooling is a down sampling operation that dimensional reduction of the input feature map gets done in the spatial domain. It calculates the maximum value of each area within a region along input feature maps. Mathematically, it can be represented as:

$$O(i,j) = \max_{m,n} I(i \times s + m, j \times s + n) \quad \text{...... (3)}$$

4) Data Standardization and Reshaping: Normalize the input features to ensure mean zero and variance one. This step stabilizes the training process and promotes convergence. Reshape the input data according to CNN model's anticipated output shape.

5) Model Training and Evaluation: Its normal to train the CNN model with pre-processed as well reshaped training data. Negative monitor the training process to detect of overfitting or under-fitting. Measure the model's performance with an evaluation set of data by mean squared error (MSE), root mean square error RMSE and a correlation coefficient.

6) Model Interpretation and Analysis: Interpret the trained model to determine what it predicts. Evaluate the learned subset of features and their significance in terms of rainfall prediction.

3.3 Random Forest Algorithm

Random Forest employs an ensemble of decision trees to classify rainfall patterns, leveraging neuro-machinate techniques for robust and accurate classification.

1) Data Collection: The first step involves cleaning and filtering the weather dataset. Split the dataset into features which are known independent variables, and a target variable – rainfall amounts. In our study, we utilize the "PrecepitationSumInches" attribute as a fundamental component for determining the accuracy and RMSE metrics. Our data

set is partitioned into training and testing subsets using a test size of 0.2 and a random state of 42.

2) Feature Selection and Engineering: Perform feature selection that finds relevant features for rainfall prediction. The methods include the correlation analysis, feature importance ranking, Outlier Removal, Transforming the DataFrame are used. The feature selection is performed on dew point, humidity patterns from observed data for increasing the model output.

3) Model Selection and Training: Training is initiated using the fit() method specifying input_train as the input features and output_train as target variable. Predictions on the training data are made using predict() method, and the results are stored in predictions_forest.

4) Model Evaluation: The performance of a trained model obtained from training set using parameters such as MAE, RMSE and correlation coefficient. The estimation of the model's generalization performance by cross-validation.

3.4 Multilayer Perceptron

Multilayer perceptron utilizes multiple layers of interconnected neurons to learn complex relationships in rainfall data, enabling advanced classification and pattern analysis.

1) Data Collection: The first step involves cleaning and filtering the weather dataset. Split the dataset into features which are known independent variables, and a target variable – rainfall amounts. In our study, we utilize the "Precepitation SumInches" attribute as a fundamental component for determining the accuracy and RMSE metrics. Our dataset is partitioned into traning and testing subsets using a test size of 0.2 and a random state of 42.

2) Feature Selection and Engineering: Perform feature selection that finds relevant features for rainfall prediction. The methods include the correlation analysis, feature importance ranking, Outlier Removal, Transforming the DataFrame are used. The feature selection is performed on dew point, humidity patterns from observed data for increasing the model output.

3) Model Selection and Architecture Design: Select suitable machine learning models for rainfall forecasting. In this specific case, a Multilayer operation on the perceptron regressor is applied.

Create the architecture of MLPR model with 3 different layers corresponding to input, hidden and output layers. The Feedforward operation as per Eqns.(4) and (5). is performed on input, hidden and ouput layers. Activation function is applied to the hidden and output layers as given in Eqn.(6).

Feed forward Operation:

The feedforward operation calculates the network's output for each neuron based on element features. Mathematically, it can be represented as:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j \qquad \text{......... (4)}$$

$$a_j = \emptyset(z_j) \qquad \text{..........(5)}$$

Activation Function:

The activation function brings nonlinearity to the MLP and enables it for learning complex patterns. Common activation functions include:

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad \text{...........(6)}$$

Back propagation Algorithm:

The back propagation is used for the adjustment of weighting factor in order to reduce and get closer as possible to actual output. The weight update rule for each weight wij in the network is given by:

$$w_{ij}^{new} = w_{ij}^{old} - \eta \frac{\partial E}{\partial w_{ij}} \qquad \text{............... (7)}$$

4) Data Standardization and Reshaping: Tune the input features to be zero mean and unit variance. Transform the Input data to fit into the shape expected by an MLPR model.

5) Model Training and Evaluation: Standardize and reshape the train data set before training MLPR model. Assess the performance of the trained model to give an idea about how well it performs on its training set for metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error RMSE.

4 Results and Discussions

The performance measures indicated in the table reveal how many different machine learning algorithms perform well for rainfall prediction. The evaluated metrics of importance are RMSE, the average magnitude of error between predicted and actual rainfall values; Correlation reflecting linearity for strength as well direction in relationships observed among predicted against recorded rainfalls is evaluated; while Accuracy representing percentage classified instances correctly. CNN produced better results with an accuracy level of 99.7666% as shown in Fig1. Random Forest came close after with the same accuracy percentage. Such high accuracy scores show that the CNN as well as Random Forest algorithm efficiently classify rainfall patterns with accurate results.

Table.1 Performance Comparison

Algorithm name	RMSE	Correlation	Accuracy
RNN	0.2556	0.63569	99.7443
LSTM	0.2640	0.6304	99.7359
CNN	0.2333	0.7144	99.7666
XG Boost	0.2640	0.6304	99.7359
Multilayer perceptron	0.2470	0.5594	99.7359
Random Forest	0.2254	0.9629	99.7359

As for the statistical model assessment of all reviewed algorithms, it has been established that Convolutional Neural Network (CNN) demonstrated lowest RMSE value-0.2333 making CNN superior in minimizing prediction errors compared with other evaluated methods and models as shown in Table1. Right after, Random Forest yielded an RMSE of 0.2254 showing its ability to minimize error with a similar level of performance as the first algorithm had done above as well.

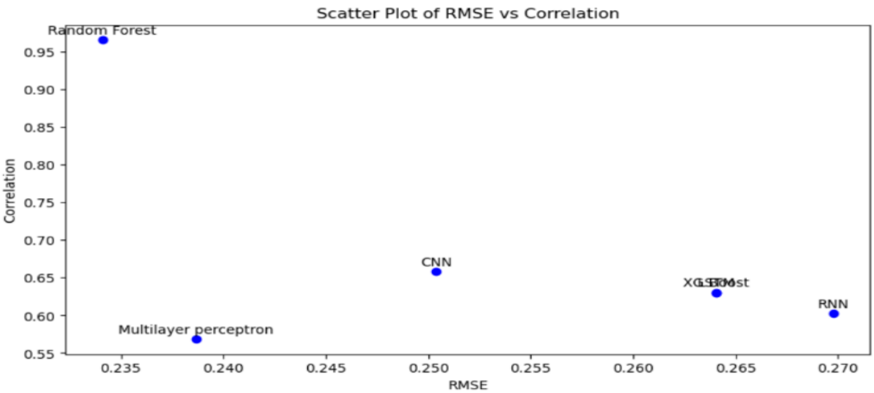

Fig 2: Comparison of the result among different algorithms with respect to the Scatter Plot

This implies that such models as CNN and Random Forest algorithms prove as quite efficient in forecasting rainfall deviating from the observed values to a minimal level. All other category of algorithms is competitive – they include Recurrent Neural Network (RNN), Long Short-Term Memory LSTM, XG Boost for which the RMSE values are relatively larger; more-over all correlation coefficients prove lower.

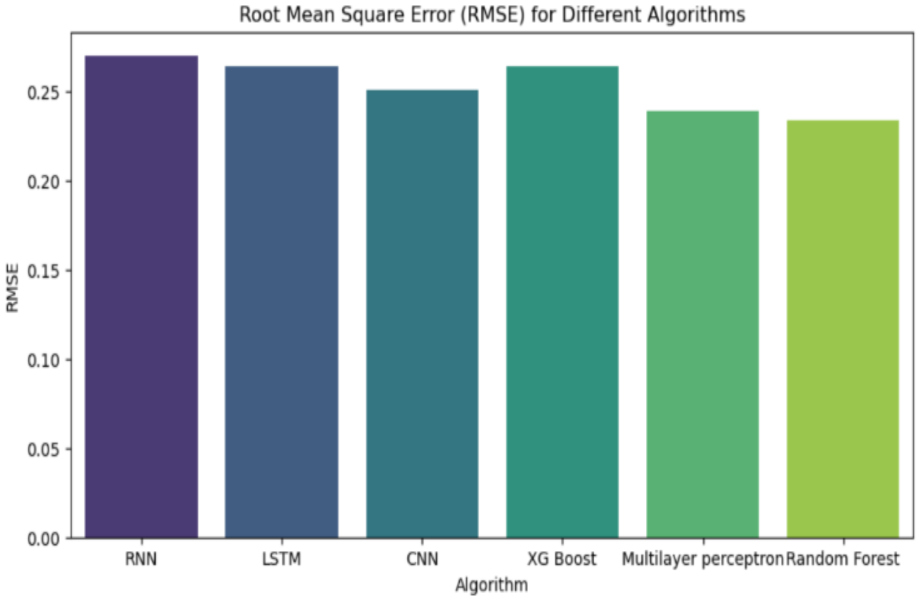

Fig.3: Comparison of the result among different algorithms

5 Conclusion

The results concluded that Convolutional Neural Network (CNN) and Random Forest are shown as most effective methods for rainfall forecasts in all considered indicators. In particular, CNN shows excellent results in terms of accuracy which proves it as a rainfall prediction task has to be one the best options. Having a low RMSE value, random forest demonstrates higher performance in psyching values of prediction errors by implying the effectiveness nature to come up with rainfall predictions that closely match observed. Additionally, the Correlation coefficient for CNN is quite high indicating a fairly linear relationship between predicted versus observed rainfall values. Moreover, CNN and Random Forest reaches the best specificity value showing their ability in rain pattern classification with high accuracy.

6 References

1. Su, B. et al. Estimated changes in diferent forms of precipitation (snow, sleet, and rain) across China: 1961–2016. Atmos. Res. 270, 106078 (2022).

2. P. P. Singhal, P. Singh and A. Vidyarthi (2020) Interpretation and localization of Thorax diseases using DCNN in Chest X-Ray. Journal of Informatics Electrical and Electronics Engineering,1(1), 1, 1-7
3. M. Vinny, P. Singh (2020) Review on the Artificial Brain Technology: BlueBrain. Journal of Informatics Electrical and Electronics Engineering,1(1), 3, 1-11
4. Aparna, Y., Somasekhar, G., Bhaskar, N., ...Divya, G., Reddy Madhavi, K..Analytical Approach for Soil and Land Classification Using Image Processing with Deep Learning, 2023 2nd International Conference for Innovation in Technology (INOCON),Bangalore, India. Mar 3-5, 2023. IEEE -| DOI: 10.1109/INOCON57975.2023.10101169.
5. Pala, V.C.R., Kamatagi, S., Jangiti, S., Reddy.Madhavi, K. , Kumar, G.N. Yoga Pose Recognition with Real time Correction using Deep Learning",2nd International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2023 – IEEE Proceedings, 2023, pp. 387–393
6. K. Chane, F.M. Gebru, B. Khan (2021) Short Term Load Forecasting of Distribution Feeder Using Artificial Neural Network Technique. Journal of Informatics Electrical and Electronics Engineering, Vol. 02, Iss. 01,S. No. 002, pp. 1-22, 2021.
7. P. Xie, R. Joyce, S. Wu, S. Yoo, Y. Yarosh, F. Sun, and R. Lin, "Re processed, bias-corrected CMORPH global high-resolution precipitation estimates from 1998," J. Hydrometeor., vol. 18, pp. 1617–1641, 2017.
8. Lachapelle, M.; Thériault, J.M. Characteristics of Precipitation Particles and Microphysical Processes during the 11–12 January 2020 Ice Pellet Storm in the Montréal Area, Québec, Canada. Mon. Weather Rev. 2022, 150, 1043–1059.
9. Au, Q.; Herbinger, J.; Stachl, C.; Bischl, B.; Casalicchio, G. Grouped feature importance and combined features effect plot. Data Min. Knowl. Disc. 2022, 36, 1401–1450.
10. Molla, T., Khan, B., & Singh, P. (2018). A comprehensive analysis of smart home energy management system optimization techniques. Journal of Autonomous Intelligence, 1(1), 15-21.
11. Łupikasza, B. E. & Małarzewski, Ł. Trends in the indices of precipitation phases under current warming in Poland, 1966–2020. Adv. Clim. Change Res. 14, 97–115 (2023).
12. Leach, J. A., Kelleher, C., Kurylyk, B. L., Moore, R. D. & Neilson, B. T. A primer on stream temperature processes. Wiley Interdiscip. Rev.: Water 10, e1643 (2023).
13. Aqnouy, M. et al. Comparison of hydrological platforms in assessing rainfall-runof behavior in a mediterranean watershed of northern Morocco. Water 15(3), 447 (2023).
14. Madhavi, K. Reddy, S. Viswanadha Raju, and J. Avanija. "Data Labeling and Concept Drift Detection using Rough Entropy For Clustering Categorical Attributes." HELIX 7, no. 5 (2017): 2077-2085.
15. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci. 2021;2(6):1–20.