Proceedings of the 8th World Congress on Civil, Structural, and Environmental Engineering (CSEE'23)
Lisbon, Portugal – March 29 – 31, 2023
Paper No. ICGRE 153
DOI: 10.11159/icgre23.153

Forecast Rainfall Density by Utilizing Machine Learning Models

Sung-Chi Hsu[1], Alok Kumar Sharma[2*]
[1]Department of Construction Engineering
Chaoyang University of Technology,
Taichung, Taiwan R.O.C
schsu@mail.cyut.edu.tw
[2]Department of Information Management
Chaoyang University of Technology,
Taichung, Taiwan R.O.C
rbaloksharma@gmail.com

Abstract **-** Organizations can use weather forecasting to help with decision-making when it comes to preventing disasters. Forecasting rain is challenging since weather conditions are always unpredictable in general. The prediction of rainfall uses a variety of methodologies, including statistical, hybrid, and physical approaches. In this research, we have implemented various machine learning models such as Logistic Regression (LR), Random Forest (RF), and Multi-Layer Perceptron (MLP) to predict the density of rainfall. This study has used Taiwan Ruiyan rainfall hourly dataset from 1998 to 2018 which contains five features like Air Pressure, Humidity, Temperature, Windspeed, and Wind Direction to predict the rainfall density such as low, medium, and heavy rainfall. The results data in this study are compared using statistical metrics like AUC, accuracy, recall, precision, and F1-score. The Random Forest, and Multi-Layer Perceptron models, had the highest accuracy scores of 0.71, accurately predicting the results. This study offers a comprehensive overview of several methods and their rainfall density predictions. By comparing these models, we can decide which one is best for predicting rainfall. The suggested work is extensively used in a variety of agriculture and civil applications, including hazard prediction, prevention, operational planning, and many more.

Keywords**:** Rainfall Prediction, Multi-Layer Perceptron, Logistic Regression, Random Forest, machine learning models

1. Introduction

Rainfall is an important meteorological component because of the numerous ways it affects human activities including farming, power generation using water, and managing water supplies. In point of fact, rainfall is one of the most important factors affecting agricultural output in developing nations [1].

Taiwan is situated in a climatic zone characterized by subtropical monsoons, has consistent rainfall throughout the year, is encircled on all sides by water, and has a topography that has significant variations in altitude. Extreme weather systems, like typhoons and land sliding, often create catastrophic flooding in Taiwan [2]. The southern region of Taiwan is considered as subtropical. Typhoon season, which runs from May to October, is the rainiest time in southern Taiwan. The rest of the months get almost no rain. As a consequence, the rainfall that typhoons bring is the primary supply of water in southern Taiwan. Furthermore, the intense rainfall that occurs during the brief length of typhoons not only results in an abundance of water but also creates natural catastrophes like as river water surges, debris flows, and floods farther downstream. For this reason, southern Taiwan is in desperate need of an effective rainfall prediction models to precisely anticipate the real-time rainfall throughout typhoon seasons and to avert the accidents that come from heavy rainfall in local regions.

In recent years, there have been significant advancements made in the field of machine learning (ML). Previous researchers have applied various type of machine learning to predict the rainfall for example Naïve Bayes, support vector machine, artificial neural networks, decision tree, Gradient boosting machine and many mores [3]–[5].

The primary goals of this research are to (1) construct a model for predicting rainfall utilizing Machine learning models and (2) analyse the potential machine learning models that can be useful for rainfall forecasting in southern Taiwan. Both goals will be accomplished during this study. In this research, we have applied three type of machine learning models like as multi-layer perceptron, Logistic Regression and Random Forest to predict the rainfall density for southern part of Taiwan. We will also compare the proposed models-based ROC -Curve, Recall, Precision and f1- score.

The remaining parts of the article are structured as described below. Description of the literature review and classification that can be seen in Section 2: In Section 3, we discussed methodology. The results of the experiments described in Section 4 in addition, the 5th section summarizes our findings from the research.

2. Literature Review

2.1. Related work

Since a significant amount of time ago, researcher have relied on three primary types of models to make accurate precipitation forecasts: linear, statistical, and conceptual models. In previous research, many researchers [6]–[8] have used Artificial neural network (ANN) for forecast rainfall. [8] proposed an artificial neural network model to predict rainfall for Columbia summer season. [7] have used Jordan rainfall dataset to predict the yearly rainfall using artificial neural network. [9] have used BP neural network to forecast the rainfall for coastal areas. [4] proposed a method constructed on the Genetic Algorithm that makes use of a dimensionality reduction methodology as well as a Multi-layer Perceptron (MLP) for the purpose of performing an analysis that is both dynamic and efficient using real-time data. [3] suggests using ensemble learning to improve the accuracy of forecasting rainfall. [10] applied various type of machine learning models such as Support vector machine, Decision tree, Random Forest, Naïve bayes, and Neural networks to predict the Malaysia country rainfall. Malaysian stations Selangor provided the dataset. Various pre-processing operations were used to fix missing data and eliminate noise. The analysis outcome shows that Random Forest categorized 1043 of 1581 cases using 10% training data. [5] have used machine learning models to predict the Australian rainfall. In this study they have used ten year of rainfall data from 2007 to 2017. They also developed a neural network method for prediction. This study outcome shows that neural network and traditional method can useful to predict the rainfall with higher accuracy. [11] have proposed an advanced artificial intelligence (AI) model to predict daily rainfall and they also compare their proposed method with other artificial intelligent model such as artificial neural network, support vector machine, and Adaptive network built with fuzzy interference. For the all the experiment they utilized Vietnam daily rainfall data with wind speed, temperature, humidity and solar radiation features. This study also suggested SVM model performance better than with other used models. [12] have applied various machine learning techniques such Bidirectional Multi-Layer Perceptron (MLP) Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Long Short-Term Memory (BLSTM) to predict rainfall using Bhutan rainfall dataset

2.2. Classification model

In this study we have implemented various machine learning models such as Random Forest (RF), Logistic Regression (LR), and Multi-Layer Perceptron (MLP) to predict rainfall density.

2.2.1. Logistic Regression

Logistic regression (LR) [13] is the Supervised Learning. It is among the most common and widely used machine learning algorithms**.** It is employed to predict the categorical dependent variable provided a collection of independent variables. LR is an approach that attempts to tackle classification issues. It is accomplished by the prediction of discrete outcomes, as opposed to the continuous results predicted by linear regression. LR [14] is a method that can be used to categorize observations by using various kinds of data, and it can effectively discover which factors are the most useful when it comes to classifying observations. The logistic regression equation is shown in Eq. (1) where p denotes the probability of 1, e is natural logarithm base and a and b represents model perimeter.

$$P = \frac{1}{1 + e^{-(a+bx)}} \tag{1}$$

2.2.2. Random Forest

Random Forest (RF) [15] is one of many machine learning algorithms used for predictive modelling. It builds a model that is made up of many decision trees, such as a forest with individual trees from different species. These trees are trained to predict a categorical variable as often as possible, given the data and the target property that needs to be predicted. RF has widespread use in the areas of regression and classification analysis. RF model excels in both classification and regression because of its ability to accommodate data sets with both categorical and continuous variables. The Gini index is used in order to determine the relationship between nodes on a decision tree. Gini index equation is shown in Eq. (2). In this equation p_i denotes the relative class frequency that is being examined in the dataset and c is the number of classes.

$$Gini = 1 + \sum_{i=1}^{c} (p_i)^2 \quad (2)$$

2.2.3. Multilayer perceptron

An artificial neural network (ANN) [4], [16] that has a feed-forward architecture and is made up of numerous layers of neurons is called a multilayer perceptron (MLP)**.** A MLP has three or more layers including one input, one output, and one or even more hidden layers**.** MLP uses neurons with a nonlinear activation function, and every layer is fully connected with the next. A sigmoid activation function is used by each and every node that makes up the MLP. Using the sigmoid equation, the sigmoid activation function transforms input real values into a value ranging from 0 and 1. Sigmoid formula is displayed in Eq. (4) and in Eq. (3), z is denoting output, *x* is representing the inputs and *w* is weights.

$$z = \sum_{i=0}^{m} w_i x_i + bias \quad (3)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

3. Methodology

3.1. Dataset & Data Pre-processing

In this study, we have used Ruiyan Taiwan rainfall data to predict rainfall density. The rainfall data provided by the Department of Atmospheric Sciences, Chinese Culture University which is contains 18836 only rainfall records from 1998 to 2018. The rainfall dataset contains six parameters with rainfall such as Air Pressure, Temperature, Humidity, Wind Direction, Windspeed, and rainfall where all attributes have continuous values in the data pre-processing, we have removed all the missing and null values. For this study, first, we converted our rainfall continuous target variable to multiclassification such as Low, Medium, and Heavy rainfall-based rainfall reports based on the MANOBS manual of surface weather observation [17]. Where 2.5 mm equal and below rainfall as Low Rainfall, above 2.5 mm and equals and below 7.5 mm rainfall as medium rainfall, and above 7.5 mm rainfall we consider as heavy rainfall. Before the build our models, we normalize our data 0 to 1 range and changed the target variable from string to number. We assigned 0 as a Low, 1 as a Medium, and 2 as a heavy rainfall status.

3.2. Split Dataset

To build our machine learning models, we have divided our dataset into training dataset and testing datasets where the training dataset have employed to train our machine learning models and the testing dataset for testing our model. In this research, the training dataset contains 13185 datasets which is 80% of our dataset and the training dataset has 5651 that is 20% of our dataset.

3.3. Classification models Build

In this research, we have applied three type of machine learning models to predict the rainfall density such as Random Forest (RF), Logistic Regression (LR), and Multi-Layer Perceptron (MLP).

The strategy for forecasting rainfall density via machine learning models is presented in Algorithm 1.

Algorithm 1: Predict to Rainfall density

Algorithm for Rainfall Dataset
//Purpose: To forecast rainfall density
//Input Data: Ruiyan Taiwan Rainfall
//Output: Predict the rainfall density
Step 1: Cluster the rainfall continues data to multiclassification data (Low, Medium and Heavy Rainfall)
Step 2: Remove the null and missing values from the rainfall dataset
Step 3: Split the dataset into training and testing dataset after data pre-processing
Step 4: Train the machine learning models
Step 5: Predict the rainfall density for test dataset using machine learning models
Step 6: Rainfall density prediction Outcomes.

3.4 Evaluation matrices

Table 1: Evaluation Matrices Formula

Name	Formulas
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2*\left(\frac{precision*recall}{precision+recall}\right)$

- *TP*: TP indicates as True Positive and reflects the number of properly categorized positive cases.
- *TN*: True Negative is denoted by the term "TN" and indicates the total number of negative occurrences that have been correctly categorized.
- *FP*: FP indicates as False Positive value, which refers to the number of real negative samples that were misclassified as positive.
- *FN*: FN refers to a False Negative value, that denotes the number of real positive cases categorized as negative.

4. Experimental setup and Outcomes

All the algorithms were programmed in Python and executed in a Jupiter notebook using Sklearn python package. All the analysis run on windows Intel® core (TM) i7-3770 CPU @ 3.40GHz with 64 GB ram.

The prediction result of LR is showing the classification result Table 2 where precision for low rainfall is showing 71%, Recall is 100% and F1-score is showing 83% for Low rainfall testing dataset. For Medium rainfall precision is 29%, recall 1% and f1-score is 1%. For Heavy rainfall precision is 44%, recall is 6% and 11% for f1-score. LR's accuracy is 70%. In Roc curve Fig. 1 is showing Roc curve values for low, medium, and heavy is 0.62, 0.58 and 0.69 respectively.

Table 2: LR classification results

	Precision	Recall	F1-Score	Support
Low Rainfall	0.71	1.00	0.83	3964
Medium Rainfall	0.29	0.01	0.01	1237
Heavy Rainfall	0.44	0.06	0.11	450
Accuracy			0.70	5651
Macro Avg	0.48	0.35	0.32	5651
Weighted Avg	0.60	0.70	0.59	5651

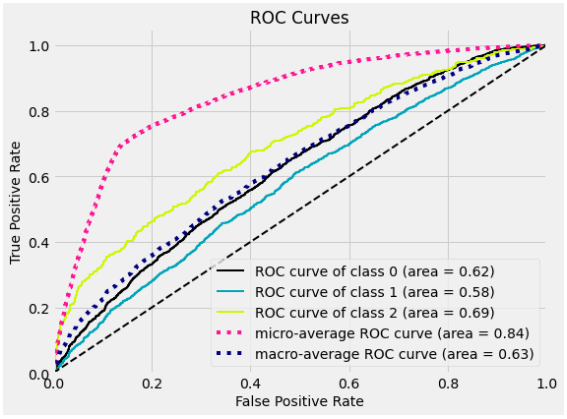

Fig. 1: LR ROC Curve

The classification result Table 3 is showing the prediction result of RF where precision for low rainfall is predicted 73% recall is 97% and f1-score is predicted 83% for Low rainfall testing dataset. For Medium rainfall precision is predicted 42%, recall is 9% and f1-score is predicted 15%. For Heavy rainfall precision is 49%, recall is 12% and 20% for f1-score. RF's accuracy is 71%. In Table 3, Support is showing the predicted value for each class. In Roc curve Fig. 2 is display Roc curve values for low, medium, and heavy is 0.67, 0.62 and 0.74 respectively

Table 3: RF Classification Results

	Precision	Recall	F1-Score	Support
Low Rainfall	0.73	0.97	0.83	3964
Medium Rainfall	0.42	0.09	0.15	1237
Heavy Rainfall	0.49	0.12	0.20	450
Accuracy			0.71	5651
Macro Avg	0.55	0.4	0.39	5651
Weighted Avg	0.64	0.71	0.63	5651

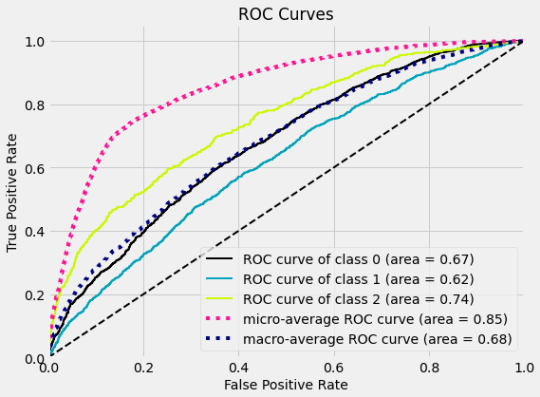

Fig. 2. RF ROC Curve

The classification Result Table 4 is showing the prediction result of MLP where precision for low rainfall is showing 72%, Recall is 99% and F1-score is showing 83% for Low rainfall testing dataset. For Medium rainfall precision is 41%, recall 4% and f1-score is 7%. For Heavy rainfall precision is 51%, recall is 9% and 16% for f1-score. MLP's accuracy is 71%. In Roc curve Fig. 4 is showing Roc curve values for low, medium, and heavy is 0.63, 0.58 and 0.69 respectively.

Table 4: MLP Classification Results

	Precision	Recall	F1-Score	Support
Low Rainfall	0.72	0.99	0.83	3964
Medium Rainfall	0.50	0.02	0.03	1237
Heavy Rainfall	0.47	0.14	0.22	450
Accuracy			0.71	5651
Macro Avg	0.56	0.38	0.36	5651
Weighted Avg	0.65	0.71	0.61	5651

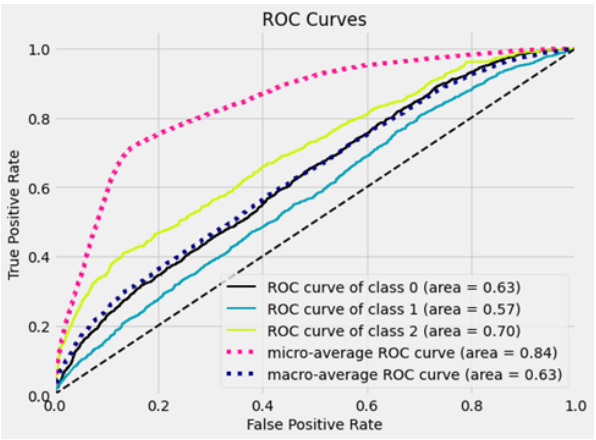

Fig. 4: MLP ROC Curve

This study also compares the used models based on classification results where RF, and MLP have highest accuracy 71%. If we compare with ROC-curve values than RF has highest value for all the classes, where class 0 has value 0.67, class 1 has 0.61, and class 2 has 0.74.

5. Conclusion

Rainfall is an important meteorological component because of the numerous ways it affects human activities including farming, power generation using water, and managing water supplies. Rainfall is one of the most significant aspects affecting agricultural output in developing nations. To solve these kinds of problems in this research, we have employed three types of machine learning such as Logistic Regression, multilayer perceptron (MLP), and Random Forest (RF). We have used the Ruiyan Taiwan dataset to predict the rainfall density for this area. The Chinese Culture University's Department of Atmospheric Sciences is the source of this data on the rainfall. To Predict the rainfall density first, cluster our dataset continuous target variable to multiclassification variable Where 2.5 mm equals and below rainfall as Low Rainfall, above 2.5 mm and equal and below 7.5 mm rainfall as a medium rainfall, and above 7.5 mm rainfall we consider as heavy rainfall. After the build mentioned machine learning models using the training dataset, we predicted the rainfall density on the testing dataset. This study also compares the mentioned models based on classification results where RF and MLP have the highest accuracy 71%. If we compare ROC-curve AUC's values then RF has the highest value for all the classes, where class 0 has a value of 0.67, class 1 has 0.61, and class 2 has 0.74.

In future studies, we will use some other machine learning models and try to improve the model accuracy for rainfall prediction, and, we will also try to use this method in different datasets.

Acknowledgements

The authors would like to express our gratitude to the Ministry of Science and Technology in Taiwan for providing necessary funding for this study (MOST 110-2625-M-324-001-). In addition, we would like to express our gratitude to Chinese Culture University's Department of Atmospheric Sciences for providing the rainfall data.

References

[1] H. N. Nguyen, T.A. Nguyen , H. B. Ly , V. Q. Tran, L. K. Nguyen , M. V. Nguyen , and C. T. Ngo, "Prediction of daily and monthly rainfall using a backpropagation neural Network," *J. Appl. Sci. Eng.*, vol. 24, no. 3, pp. 367–379, 2021, doi: 10.6180/jase.202106_24(3).0012.

[2] C. C. Wei and T. H. Huang, "Modular neural networks with fully convolutional networks for typhoon-induced short-term rainfall predictions," *Sensors*, vol. 21, no. 12, pp. 1–19, 2021, doi: 10.3390/s21124200.

[3] N. S. Sani, A. H. A. Rahman, A. Adam, I. Shlash, and M. Aliff, "Ensemble Learning for Rainfall Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 153–162, 2020, doi: 10.14569/IJACSA.2020.0111120.

[4] P. Satish, S. Srinivasulu, and R. Swathi, "A hybrid genetic algorithm based rainfall prediction model using deep neural network," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 5370–5373, 2019, doi: 10.35940/ijitee.L3777.1081219.

[5] M. Raval, P. Sivashanmugam, V. Pham, H. Gohel, A. Kaushik, and Y. Wan, "Automated predictive analytics tool for rainfall forecasting," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021, doi: 10.1038/s41598-021-95735-8.

[6] Berlilana, W. M. Baihaqi, and Sarmini, "Artificial neural network for rainfall prediction base on historical rainfall data by day," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.5 Special Issue, pp. 275–280, 2019, doi: 10.30534/ijatcse/2019/4881.52019.

[7] S. Sharadqah, S. M. Perez, A. M. Mansour, M. A. Obeidat, and R. Marbello, "Nonlinear Rainfall Yearly Prediction based on Autoregressive Artificial Neural Networks Model in Central Jordan using Data Records: 1938-2018," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 2, pp. 240–247, 2021, doi: 10.14569/IJACSA.2021.0120231.

[8] A. Nair, G. Singh, and U. C. Mohanty, "Prediction of Monthly Summer Monsoon Rainfall Using Global Climate Models Through Artificial Neural Network Technique," *Pure Appl. Geophys.*, vol. 175, no. 1, pp. 403–419, 2018, doi: 10.1007/s00024-017-1652-5.

[9] J. Wu and X. Xing, "Rainfall forecast and computer data algorithm optimization in coastal areas based on improved neural network," *Arab. J. Geosci.*, vol. 14, no. 15, 2021, doi: 10.1007/s12517-021-07579-1.

[10] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for malaysian rainfall prediction," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1148–1153, 2016, doi: 10.18517/ijaseit.6.6.1487.

[11] B. T. Pham, L. M. Le, T. T. Le, K. T. T. Bui, V. M. Le, H. B. Ly, and I. Prakash, "Development of advanced artificial intelligence models for daily rainfall prediction," *Atmos. Res.*, vol. 237, no. November 2019, p. 104845, 2020, doi: 10.1016/j.atmosres.2020.104845.

[12] M. Chhetri, S. Kumar, P. P. Roy, and B. G. Kim, "Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan," *Remote Sens.*, vol. 12, no. 19, pp. 1–13, 2020, doi: 10.3390/rs12193174.

[13] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.

[14] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[16] A. K. Sharma, L.-H. Li, and R. Ahmad, "Identifying and predicting default borrowers in P2P lending platform: A machine learning approach," in *2021 IEEE International Conference on Social Sciences and Intelligent Management (SSIM)*, Aug. 2021, pp. 1–5, doi: 10.1109/SSIM49526.2021.9555200.

[17] Atmospheric Environment Service, *Manual of Surface Weather Observations 7th Edition*, no. January. 2013.