

Evaluating Machine Learning Models To Enhance Rainfall Prediction

Antony Vigil M S^{1*}, Shwetha R², Swati Singh³, Shruthi R⁴

^{1*,2,3,4}Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

Abstract

Rainfall forecast plays an important role in elevating consciousness about the possible hazards related with it and helping individuals take dynamic measures for their safety. It is used for various purposes like agriculture, water conservation, and seismic reinforcing. Heavy rainfall forecast is a significant problem for the meteorological department. It combines meteorological data, historical weather patterns, and complex computer models. They rely on historical data to identify patterns between rainfall and other agents like temperature, humidity, wind, etc. Statistical techniques are also used to analyze vast datasets and find relationships. The major aim of this study is to recognize the pertinent atmospheric features that create rainfall. This paper investigates the performance of the ML models, namely Decision tree, KNN(K-Nearest Neighbor), and Logistic regression. These model's performances have been calculated through the evaluation metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). The Standard and Min-Max correlation technique was used to select applicable climatic parameters that were used as input for the ML model. The data record was gathered from Kaggle to measure the performance of ML models. The study revealed that KNN surpassed the other models by an accuracy rate of 84.183% compared to Decision Tree with the rate of 83.762%.

Keywords: Machine learning, Logistic regression, Decision Tree, KNN, MAE, MSE, RMSE

I. INTRODUCTION

Rainfall is a vital component of our ecosystem, sustaining life and maintaining ecological balance by replenishing Earth's freshwater resources [1]. Making data driven decisions about agricultural planting, irrigation, and water conservation is essential for sustainable agriculture. ML has progressed into a powerful tool for making highly precise rainfall predictions. Several environmental factors influence rainfall intensity and occurrence, such as temperature, relative humidity, sunshine, pressure, evaporation, etc [2].

Rainfall prediction is a complex challenge due to factors like limited, low-quality, or insufficient data, the intricate and chaotic nature of weather systems, the difficulty in selecting and fine-tuning models, and the increasing unpredictability caused by climate change. To address the challenges posed by rainfall variability, this project focuses on developing a comprehensive machine learning pipeline, encompassing data acquisition, model development, and evaluation.

Rainfall prediction can be characterized based totally on the type of dataset applied, the models used, and the rate of data collection[3]. They can be mainly classified into statistical methods, ML methods and deep learning methods[4]. Traditional weather prediction methods, including empirical models and statistical models have boundaries in identifying hidden styles or non-linear developments in rainfall records. NWP models rely on mathematical modeling and modern-day climatic factors, which might not be immediately well matched with parametric equations[5].

The ML algorithms learn to recognize patterns from historical data between the factors and rainfall. Traditional statistical models are widely used for rainfall forecasting. Many researchers applied data mining techniques[6] to refine the accuracy of short-term, mediumterm, and long-term rainfall forecasts. But many researchers have found that machine learning techniques significantly improve rainfall prediction accuracy across various conditions, including during extreme rainfall events[7]. This research aimed to predict rainfall intensity by identifying key atmospheric factors using machine learning. Raw meteorological data was collected and prepared for analysis. To determine the most influential variables, each feature was correlated with rainfall using Standard and Min-Max scalar correlation. Subsequently, Decision Tree, Logistic Regression, and K-Nearest Neighbor were employed to forecast rainfall.

II. RELATED WORKS

The project aimed to enhance flood and drought management in the Aiyar river basin by creating spatial rainfall maps and forecasting rainfall using ML techniques. The study faced limitations in data availability and model complexity, as indicated by the need for more input data and exploration of hybrid ML algorithms to enhance prediction accuracy[8]. The project aimed to improve rainfall prediction accuracy in Rwanda by comparing LSTM, CNN, and GRU models, with LSTM showing the best performance. The examination highlighted the want for better resolution datasets, including satellite, soil moisture to enhance model accuracy. Additionally, the danger of overfitting when deploying the LSTM model across all stations in Rwanda because of restricted dataset size was identified[9]. The project aimed to predict daily rainfall using a ML model and compared the results to manual calculations. While the model achieved good accuracy, it could be further improved by incorporating additional variables related to rainfall occurrence[10].

The project aims to enhance rainfall prediction accuracy in the Ziz basin using machine learning techniques, recognizing the limitations of traditional methods[11]. The project idea is to analyze monthly accumulated rainfall data for 100 Indian

cities using deep learning models to understand rainfall patterns and trends. The limitation of the project is the use of monthly accumulated data instead of daily data due to computational constraints, which might affect model accuracy[12]. The project idea is to use machine learning models and factor reduction methods to predict rockfall occurrence. The limitations of the project include the potential for bias in the susceptibility maps derived from factor reduction methods and the need for further research on these maps[13].

The project aims to forecast flooding depth maps using simulated data without physical monitoring stations. While QPF-RIF shows promise, challenges lie in reducing uncertainty from ensemble rainfall data and encompassing real-time monitoring for accurate map corrections[14].

The project focusses to develop a new rainfall prediction technique using FCM and MBOA. While the FCMM-RPS technique shows promising results, future research could focus on incorporating additional environmental features and employing big data analysis for improved prediction accuracy[15]. The project targets to develop a novel stacking ensemble ML model for downsizing rainfall in Lake Urmia and Sefidrood basins under climate change conditions. While the SEML approach shows promise, challenges may arise in selecting optimal meta-algorithms and addressing potential biases in rainfall data[16].

The project aims to predict annual rainfall in Chennai using a hybrid neural network model based on interactive random forests. The model achieves 100% accuracy but faces security vulnerabilities that need to be addressed for future improvement[17]. The project aims to predict rainfall in Jordan using RNN and SVM models, focusing on the influence of meteorological variables on rainfall patterns. However, the project faces challenges in data preprocessing and model selection, as well as limitations in the data's temporal scope and geographic coverage[18]. The project focussed on analyzing the effect of climate variables on rainfall patterns in Java Island, Indonesia. However, the project faced challenges in handling extreme values in the climatological data and the need for more robust statistical methods[19].

The project aimed to forecast monthly rainfall in Benin using an MLP model and compared its performance to LSTM and CF models. However, the study faced limitations in terms of data sources, model architecture, and the ability to predict extreme rainfall events[20]. The project aimed to differentiate the presentation of MLP and LSTM models for monthly rainfall prediction in South Tangerang City. While both models achieved similar results, the LSTM model slightly outperformed MLP in terms of RMSE and MAPE values[21]. The task aimed to evaluate the productiveness of ML models for rainfall prediction in Australia[22].

The project aimed to develop a threshold version for landslide prediction in Badung Regency using satellite data. The first approach using Q1 demonstrated excellent performance, but further research is needed to validate the model's generalizability in other regions[23]. The project aimed to predict time series data using machine learning methods with and without DWT and MODWT. While DWT improved prediction accuracy, the hybrid DWT approach did not meet expectations, while hybrid MODWT models showed significant improvement up to 3 days[24]. The project aimed to assess the decadal prediction skills of Sahel rainfall using CMIP5 and CMIP6 models. The results showed that both initialized and uninitialized simulations from CMIP6 models generally outperform CMIP5 models, highlighting the importance of model resolution and initialization in improving Sahel rainfall predictions[25]. The project aims to develop an automated system for diagnosing pulpitis from dental panoramic radiographs, but it faced challenges in accurately segmenting complex dental structures and handling variations in image quality and patient data[26]. The project aims to visualize and analyze blockchain transactions over a peer-to-peer network to demonstrate the advantages of blockchain technology compared to traditional banking systems[27]. The project aims to use various machine learning algorithms to sequence DNA in a human dataset and compare their performance in accurately identifying genetic variations and mutations[28].

2.1 Machine Learning Algorithms.

To select the best ML algorithms for predicting daily rainfall, existing research was analyzed. Based on these findings, KNN, Decision Trees, and Logistic Regression were chosen for this study. These algorithms were compared to find the most precise model for forecasting daily rainfall using real-time environmental data.

2.2 Logistic Regression (LR)

LR is a statistical technique for binary classification. It models the relationship between this outcome and several factors, like environmental conditions. The logistic function transforms predictions into probabilities. In this study, logistic regression was used to predict the likelihood of rainfall based on various environmental factors. The model is represented by the equation:

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

where $P(Y=1)$ is the probability of rainfall, (β_0) is the intercept, $(\beta_1, \beta_2, \dots, \beta_n)$ are the coefficients for the independent variables $(X_1, X_2, \dots, X_n)$, and (e) is the exponential function. LR is a valuable tool for understanding the factors influencing rainfall and predicting its occurrence.

2.3 Decision Trees (DT)

DT is a ML method that can predict rainfall intensity based on environmental factors. The model creates a tree-like structure in which each branch represents a decision based totally on specific situations inclusive of temperature or humidity. The process involves splitting data into subsets based on the most informative feature, creating branches for each subset, and making predictions at the final branches. The best feature for splitting is determined using metrics like Gini impurity or information gain. Decision trees are popular due to their simplicity, interpretability, and ability to handle complex relationships in data.

2.4 K-Nearest Neighbors (KNN)

KNN is a ML algorithm that classifies or predicts data based on its similarity to nearby data points. In this study, KNN was used to predict daily rainfall by identifying the closest historical weather patterns. The algorithm's steps involve selecting the number of nearest neighbors, calculating distances, and averaging the rainfall of these neighbors. KNN's reliance on similarity makes it well-suited for complex, non-linear data like weather patterns.

III. PROPOSED WORK

3.1 Data Collection

This module is dedicated to the comprehensive handling and analysis of a substantial dataset encompassing 1,50,000 entries distributed across three distinct datasets. The data is meticulously structured in CSV format, allowing for efficient data manipulation and processing. Each dataset within the module provides a rich array of meteorological information such as location, months, rainfall, windspeed, temperature, cloudad. By capturing temperature and cloudad at multiple time points, the dataset offers a granular perspective on precipitation changes.

3.2 Data Pre-processing

Data preprocessing was conducted to make certain the integrity and suitability of the dataset for rainfall prediction and the usage of system learning algorithms together with K-Nearest Neighbors (KNN) and Decision Trees. Initially, rows containing lacking values had been excluded using the `dropna()` feature to maintain data pleasant. Categorical variables, together with area, wind direction, and evaporation, had been converted into numerical representations through Label Encoding.

The dataset was subsequently partitioned into education and testing subsets to facilitate version assessment. Standardization was implemented to the numerical functions to decorate version, overall performance and mitigate bias. Finally, function choice strategies were hired to identify the most applicable variables, thereby optimizing the predictive accuracy of the models.

3.3 Feature Selection Module

Feature selection was executed to optimize the performance of the rainfall prediction fashions, together with K-Nearest Neighbors (KNN) and Decision Trees.

3.5 Prediction Module

After function choice, the predictive models had been built for the use of the diagnosed key attributes, including block, region, district, network area, X-coordinate, Y-coordinate, range, and longitude. The device getting to know algorithms, which includes K-Nearest Neighbors (KNN) and Decision Trees, had been educated on the usage of these features. A pipeline technique was applied to standardize the input statistics and streamline the modeling technique. The trained fashions were then employed to expect the goal variable, with the prediction procedure iteratively refining the version parameters to limit errors. The performance of the models was judged based on their accuracy in predicting rainfall, with the results compared to perceive the best approach.

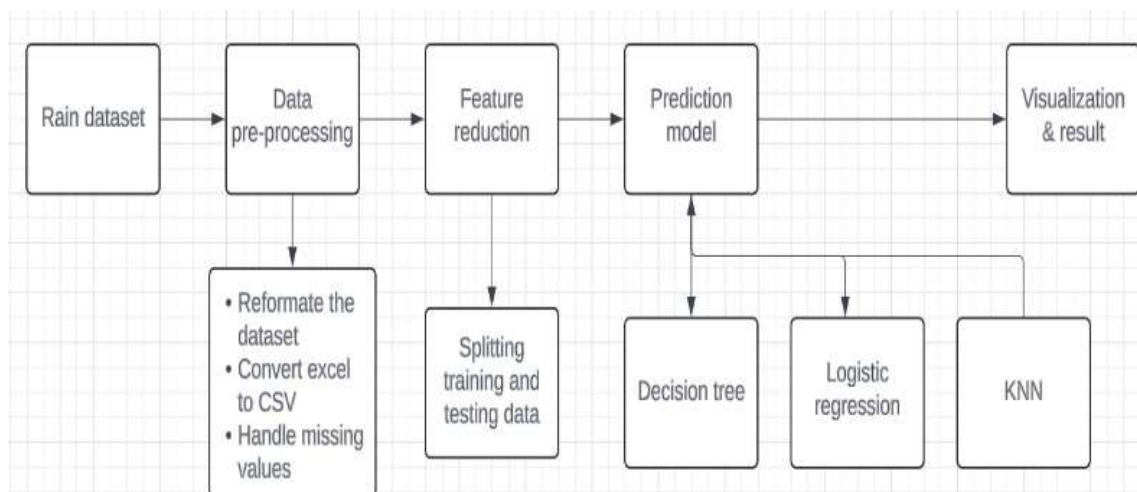


Figure 1: Architectural diagram for Machine Learning

In Fig 1 the analysis centered on figuring out the maximum relevant predictors for rainfall. Using techniques like Recursive Feature Elimination (RFE) and function significance rankings from Decision Trees, key variables have been selected, including MinTemp, MaxTemp, WindGustSpeed, WindSpeed9am, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, and RainTomorrow as given in Table 1. These features were retained to reduce dimensionality and enhance model accuracy, ensuring strong and dependable predictions.

3.4 Building And Training Module

The analysis identified location and month as key factors influencing the target variable. These features were used to train the predictive model using a ML algorithm from the scikit-learn library. A pipeline was constructed to streamline the model-building process and facilitate the comparison of results across different approaches. In the second model, a StandardScaler was applied to standardize the data, followed by training with a DecisionTreeClassifier. The model iteratively adjusted its parameters to minimize prediction errors, resulting in a refined model optimized for rainfall prediction.

#Fit the training data for minmax scaler

```
Pipe_knn_mm=pipe_knn_mm.fit(X_train, Y_train)
```

3.6 Visualization Module

Visualization techniques had been applied to assess and interpret the version's predictions. After invoking the model(forecast_pred), the ensuing predictions have been plotted against real values to assess the version's accuracy. Scatter plots and line graphs in comparison predicted and observed rainfall values, supplying a visible representation of overall performance. Feature importance became illustrated via bar charts, highlighting attributes which had a huge effect on the predictions. Additionally, confusion matrices and ROC curves assessed class performance, supplying insights into the model's efficacy and figuring out areas for capacity refinement. These visualizations have been critical to carry out a complete evaluation of the version's predictive talents and guiding in addition to optimization efforts.

Table-1 Selected Parameter's utilised for Analysis and for training model's

MinTemp	MaxTemp	WindGustSpeed	WindSpeed9am	Humidity9am	Humidity3pm	Pressure9am
17.5	32.3	41.0	7.0	82.0	33.0	1010.8
15.9	21.7	31.0	15.0	89.0	91.0	1010.5
15.9	18.6	61.0	28.0	76.0	93.0	994.3
14.1	20.9	22.0	11.0	69.0	82.0	1012.2
13.5	22.9	63.0	6.0	80.0	65.0	1005.8

The Figure 2 and Figure 3 are visualized graph's in module:

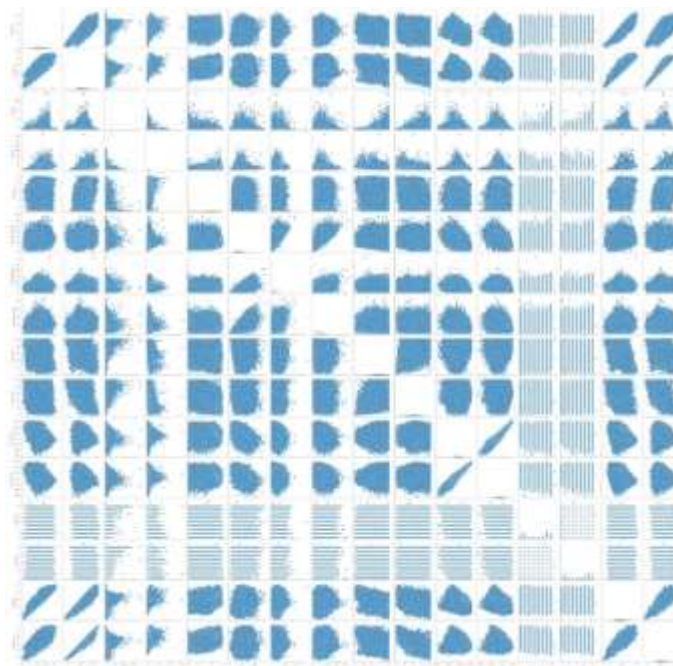


Figure 2: Pairplot of dataset variables

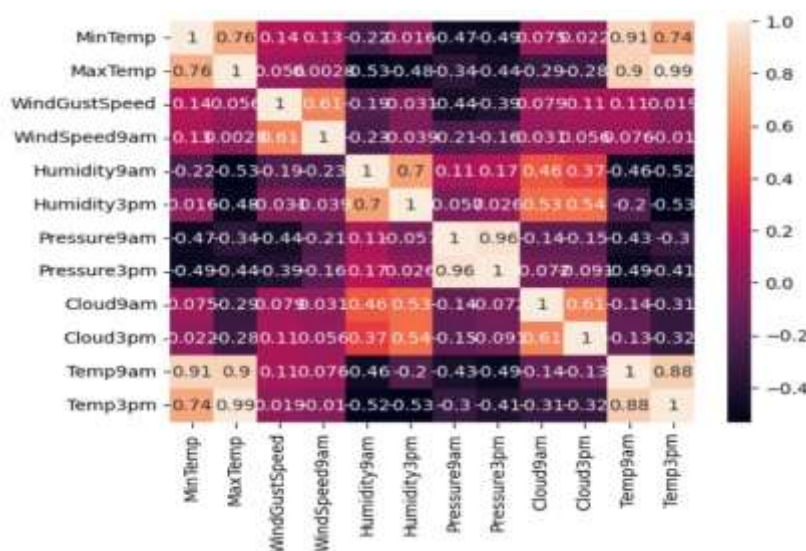


Figure 3: Heatmap of dataset variables

IV. RESULTS AND DISCUSSION

Accurate rainfall forecasting is necessary for productive water resource management and disaster reduction. This study explores the performance of various ML algorithms in predicting daily rainfall, focusing on KNN, Decision Trees, and Logistic Regression. Meteorological data was analysed, with relevant environmental variables selected using Standard and Min-Max scalar correlation.

Results indicated that KNN outperformed the other algorithms, achieving an accuracy rate of 84.183% compared to Decision Tree's 83.762%.

While Logistic Regression provided valuable insights, it did not surpass KNN or Decision Trees in this context. The slight advantage of KNN suggests its ability to better capture the complex relationships within environmental data as seen in Fig 4.

Although this study did not include sensor data, its integration could potentially enhance prediction accuracy. Future research should explore additional environmental variables and leverage big data analytics to develop more refined and reliable rainfall forecasting models.

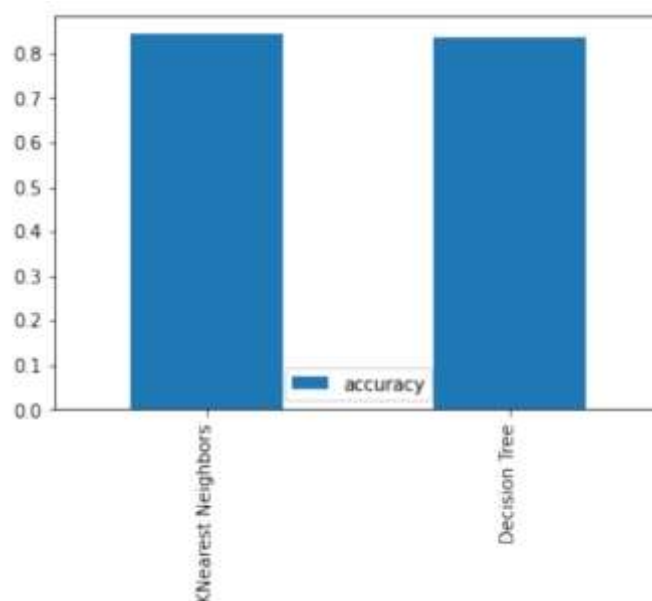


Figure 4: Efficiency rate comparison of algorithms against each other

V. CONCLUSION

Accurate rainfall prediction is crucial for numerous applications, including agricultural planning and flood risk mitigation. This study assessed the effectiveness of various ML algorithms—namely KNN, DT, and LR in forecasting daily rainfall using historical meteorological data. To ensure the relevance of the input data, environmental variables were selected based on the Standard and Min-Max scaler correlation coefficients, optimizing the model's predictive capabilities. Among the algorithms tested, KNN demonstrated superior performance, significantly outperforming both Decision Trees and Logistic Regression in predicting rainfall. This superiority is largely attributed to KNN's proficiency in managing the complexities and nuances inherent in environmental data, making it a more robust choice for rainfall prediction in this context. While Logistic Regression provided valuable insights into the relationships between variables, its predictive accuracy was comparatively lower than that of KNN and DT. To further enhance rainfall prediction accuracy, future research should inspect the incorporation of additional data sources, such as real-time sensor data, which could complement existing meteorological datasets

and enhance model precision. Additionally, examining a broader range of environmental variables and employing advanced statistical analysis techniques may contribute to more refined and reliable rainfall forecasts. In conclusion, leveraging big data and integrating diverse data sources will be crucial for advancing rainfall prediction models, ultimately improving their practical applications in water resource management.

VI. REFERENCES

- Hassan, M.M., M.A.T. Rony, M.A.R. Khan, et al., "Machine learning-based rainfall prediction: Unveiling insights and forecasting for improved preparedness," IEEE access: practical innovations, open solutions, 11, pp. 132196–132222 (2023).
- Liyew, C.M., and H.A. Melese, "Machine learning Techniques to predict daily rainfall amount," Journal of big data, 8 (1) (2021). <https://doi.org/10.1186/s40537-021-00545-4>
- "Indian summer monsoon rainfall prediction: A comparison of iterative and non-iterative approaches," Applied soft computing, 70, pp. 1122–1134 (2018). <https://doi.org/10.1016/j.asoc.2017.08.055>
- Zoremsanga, C., and J. Hussain, "Particle swarm optimize deep learning models for rainfall prediction: A case study in aizawl, Mizoram," IEEE access: practical innovations, open solutions, 12, pp. 57172–57184 (2024). [10.1109/ACCESS.2024.3390781](https://doi.org/10.1109/ACCESS.2024.3390781)
- Balamurugan, M.S., and R. Manojkumar, "Study of short term rain forecasting using machine learning based approach," Wireless Networks, 27 (8), pp. 5429–5434 (2019). <https://link.springer.com/article/10.1007/s11276-019-02168-3>
- Tharun, V.P., R. Prakash, and S. Renuga Devi, "Prediction of Rainfall Using Data Mining Techniques," <http://dx.doi.org/10.1109/ICICCT.2018.8473177> Zhang, X., S.N. Mohanty, A.K. Parida, S.K. Pani, B. Dong, and
- X. Cheng, "Annual and non-monsoon rainfall prediction modelling using SVR-MLP: An empirical study from odisha," IEEE access: practical innovations, open solutions, 8, pp 30223– 30233 (2020). [10.1109/ACCESS.2020.2972435](https://doi.org/10.1109/ACCESS.2020.2972435)

8. Natarajan, S., and A. Jegan Bharath Kumar, "Generating Spatial Distribution and Forecasting the Rainfall by Suitable ML Models- A Case Study of Aiyar River Basin, Tiruchirappalli District," MATEC Web of Conferences, 400, p.02011 (2024). <http://dx.doi.org/10.1051/mateconf/20244000201>
9. Kagabo, J., G.R. Kattel, J. Kazora, C.N. Shangwe, and F. Habiyakare, "Application of machine learning algorithms in predicting extreme rainfall events in Rwanda," Atmosphere, 15 (6), p. 691(2024). <https://doi.org/10.3390/atmos15060691>
10. Sulistiyono, M., B. Satria, A. Sidauruk, and R. Wardhana, "Rainfall prediction using multiple linear regression algorithm," JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 9 (1), pp. 17–22(2023). <http://dx.doi.org/10.33480/jitk.v9i1.4203>
11. Bouziane, S., B. Aghoutane, A. Moumen, A. Sahlaoui, and A. El Ouali, "Enhancing Precipitation Prediction in the Ziz Basin: A Comprehensive Review of Traditional and Machine Learning Approaches," E3S Web of Conferences, 489, p. 04010 (2024). <http://dx.doi.org/10.1051/e3sconf/202448904010>
12. Singh, S., A. Mukherjee, J. Panda, A. Choudhury, and S. Bhattacharyya, "Analysis and forecasting of temporal rainfall variability over hundred Indian cities using deep learning approaches," Earth systems and environment, (2024). <http://dx.doi.org/10.1007/s41748-024-00396-y>
13. Kang, J., B. Wan, Z. Gao, S. Zhou, H. Chen, and H. Shen, "Research on machine learning forecasting and early warning model for rainfall-induced landslides in Yunnan province," Scientific reports, 14 (1), pp. 1–12(2024). <http://dx.doi.org/10.1038/s41598-024-646790>
14. Chang, M.-J., I.-H. Huang, C.-T. Hsu, S.-J. Wu, J.-S. Lai, and G.- F. Lin, "Long-Term Flooding Maps Forecasting System Using Series Machine Learning and Numerical Weather Prediction System," WATER, 15 (20), p.3346(2022). <http://dx.doi.org/10.3390/w14203346>
16. Al Duhayyim, M., H.G. Mohamed, J.S. Alzahrani, et al., "Modeling of Fuzzy Cognitive Maps with a Metaheuristic-Based Rainfall Prediction System," Sustainability: Science Practice and Policy, 15 (1), p. 25(2022). <https://doi.org/10.3390/su15010025>
17. Anaraki, M.V., M. Kadkhodazadeh, A. Morshed-Bozorgdel, and S. Farzin, "Predicting rainfall response to climate change and uncertainty analysis: Introducing a novel downscaling CMIP6 models technique based on the stacking ensemble machine learning," Journal of water and climate change, 14 (10), pp. 3671–3691 (2023). <https://dx.doi.org/10.2166/wcc.2023.477>
18. Asha, P., A. Jesudoss, S.P. Mary, K.V.S. Sandeep, and K.H. Vardhan, "An efficient hybrid Machine Learning classifier for rainfall prediction," Journal of physics. Conference series, 1770 (1), p. 012012 (2021). 10.1088/1742-6596/1770/1/012012
19. Abuhammad, H., "Rainfall prediction using deep learning Algorithms," Journal of electrical systems, 20(7s), pp.1696–1704(2024). <http://dx.doi.org/10.52783/jes.3765>
20. Andriyana, Y., A.N. Falah, B.N. Ruchjana, et al., "Spatial Durbin Model with Expansion Using Casetti's Approach: A Case Study for Rainfall Prediction in Java Island, Indonesia," Science in China, Series A: Mathematics, 12(15), p.2304(2024). <http://dx.doi.org/10.3390/math12152304>
21. Aïzansi, A.N., K.O. Ogunjobi, and F.K. Ogou, "Monthly rainfall prediction using artificial neural network (case study: Republic of Benin)," Environmental Data Science, 3, p. e11 (2024). <http://dx.doi.org/10.1017/eds.2024.10>
22. Gaol, G.A.M.L., M. Syafrullah, and S. Supardi, "Comparison of monthly rainfall prediction using Long Short Term Memory and Multi Layer Perceptron methods in South Tangerang City," Jurnal Sisfokom (Sistem Informasi dan Komputer), 13 (2), pp. 293–297 (2024). <http://dx.doi.org/10.32736/sisfokom.v13i2.2149>
23. Vinta, S.R., and R. Peeriga, "Rainfall Prediction using XGB Model with the Australian Dataset," EAI Endorsed Transactions on Energy Web, 11, (2024). <http://dx.doi.org/10.4108/ew.5386> [23] Aryastana, P., L. Dewi, and P.I. Wahyuni, "A study of rainfall thresholds for landslides in Badung Regency using satellitederived rainfall grid datasets," International Journal of Advances in Engineering Sciences and Applied Mathematics, 13 (2), p.197(2024). 10.11591/ijaas.v13.i2.pp197-208
24. Küllahcı, K., and A. Altunkaynak, "Maximizing daily rainfall prediction accuracy with maximum overlap discrete wavelet transform-based machine learning models," International Journal of Climatology, 44 (10), pp. 3405–3426 (2024). <https://doi.org/10.1002/joc.8530>
25. He, Y., B. Wang, J. Liu, et al., "Assessment of the decadal prediction skill of Sahel rainfall in CMIP5 and CMIP6," Journal of climate, 37 (8), pp. 2471–2490 (2024). <https://doi.org/10.1175/JCLI-D-23-0051.1>
26. Vigil A, Bharathi S. Diagnosis of Pulpitis from Dental Panoramic Radiograph Using Histogram of Gradients with Discrete Wavelet Transform and Multilevel Neural Network Techniques. Traitement du Signal. 2021 Oct 1;38(5). <https://doi.org/10.18280/ts.380532>
27. Vigil A, Pa thak P, Upadhyay S, Singh D, Garg V Blockchain Over Transaction System. In 2018 3rd International Conference on Communication and Electronics Systems (ICCES) 2018 Oct 15 (pp. 1111-1117). IEEE <https://doi.org/10.1109/CESYS.2018.8723962>
28. Vigil MA, Mirutuhula M, Sarvagna S, Supraja R, Reddy GP. DNA Sequencing Using Machine Learning Algorithms. In 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) 2022 Dec 8 (Vol. 1, pp. 1-4). IEEE. <https://doi.org/10.1109/ICDSAAI55433.2022.10028805>