PREDICTION OF RAINFALL ANALYSIS USING MACHINE LEARNING TECHNIQUE

[1]Gundu Ramesh, [2]Mr. B. Srinivas Rao

[1]M.Tech Scholar, Department of Computer Science and Engineering, NRI Institute of Technology, Visadala, Guntur, Andhra Pradesh, India

[2]Assistant Professor, Department of Computer Science and Engineering, NRI Institute of Technology, Visadala, Guntur, Andhra Pradesh, India

ABSTRACT: Rainfall prediction plays a crucial role in raising awareness about the potential dangers associated with rain and enabling individuals to take proactive measures for their safety. Now casting plays a crucial role in responding to disasters such as flash floods by predicting rainfall in real time. However, existing now casting models struggle to accurately predict extreme rainfall events, which, although rare, can have devastating impacts. This challenge primarily arises because typical loss functions focus on minimizing average prediction errors rather than emphasizing the importance of extreme events, leading to their underestimation. Therefore, this prediction of rainfall analysis using machine learning technique is proposed. The proposed system uses normalizing for feature values and handles missing observations. This proposed system reduces redundant features by using Principal Component Analysis (PCA). Therefore, this proposed system improves prediction accuracy, reduces computational complexity, and provides reliable rainfall forecasting by using CatBoost classsifer when compared with traditional methods.

KEYWORDS: Rainfall Prediction, Machine Learning, Principal Component Analysis (PCA), Safety, Flash Floods, CatBoost Classsifer

I. INTRODUCTION

Rainfall prediction involves forecasting the amount of rainfall in a specific area[1]. Accurate rainfall predictions offer valuable observations to decision-makers, enabling them to assign resources more efficiently, mitigate risks, and increase socio-economic outcomes [2]. Researchers face significant challenges in attaining effective and precise time series analysis and modelling, primarily due to the complex and volatile nature of time series data, especially when it involves natural processes like rainfall[3]. For hydrologists, accurately predicting daily rainfall has long been of research interest, as it plays a critical role in various areas such as crop planning and agricultural water distribution, particularly in countries like India that heavily depend on agriculture [4].

Accurate and timely rainfall prediction is expected to inject a new intervention phase to the affected sectors accosted with the negative propensities of rainfall extremes [5]. These critical sectors include but are not limited to energy, agriculture, and others, which are greatly affected by rainfall. A plethora of scholarly research has demonstrated that the duration and intensity of rainfall cause major climate-related disasters[6]. The manifestation of the impact of rainfall includes drought, floods, among others and its associated effects [7].

Nowcasting is a highly short-term weather forecasting of up to a few hours. It can provide useful information for short-term early warnings regarding hazardous meteorological conditions to support preparation for and rapid response to emergencies, as well as to mitigate the potential impacts, associated with a climate-related hazard event [8]. For example, flash flooding is a common hazard that can cause severe adverse impacts to human communities [9]. Such phenomena usually occur within hours of mesoscale rainstorm events. An effective nowcasting system for mesoscale heavy rainfall can be used to issue warnings to reduce potential flood exposure of communities that may be caught off guard otherwise [10].

For nowcasting heavy rainfall of a mesoscale convective system (MCS), the numerical weather prediction (NWP) models that are traditionally used for daily weather forecasting have their limitations [11]. NWPs produce simulations of meteorological conditions based on physical equations coupled with measurements of atmosphere. When adopted for applications with high spatial and temporal resolutions, traditional NWP models are usually hampered by the deficiencies in determination of initial and boundary conditions [12], long spin-up time [13], and challenges in physical parameterization [14]. As a result, it is difficult for traditional NWP models to provide useful nowcasting products on MCS rainfall for high-resolution applications such as hydrometeorological modeling and simulation for an urban area or a catchment.

II. LITERATURE SURVEY

P. V. S. Krithik, M. T. A. Raghavan, S. P. K and K. Vasanth, et al. [15] novel technique for the Prediction of Monsoon Rainfall using Hybrid Machine Learning Algorithms and their ensembles, operated on Real time Hyderabad Rainfall data is Proposed. Over the years many models have been developed for the prediction of rainfall, An Accurate Prediction of Monsoon Rainfall for farmers is necessary for centuries together. As regression algorithms fail to generalize the outcomes with the mean of rainfall being near 0(no rainfall). We have come up with a Hybrid Technique that uses the classification of rainfall into Binary Data, there upon which applies the Regression Algorithm, which significantly improved the result. Out of the many tests conducted on various algorithms, the Hybrid Model of Random Forest Classifier with Multi-Layer Perceptron Regressor has given the best outcome with a practical Accuracy of 0.92537 and with a Mean Absolute error of 1.063540 tested over a period of 3 months.

N. Tiwari and A. Singh, et al. [16] research has conducted a study on rainfall pattern in Indian states using machine learning techniques and algorithms, which utilizes government provided rainfall data from 1901-2017. This research proposes a study and analysis of rainfall in the Indian states using machine learning algorithms and compared their performances to the standard results.

N. Nallametti, C. Malarvizhi, S. J. Ahmed, R. Sathya, K. H. Vishnu and J. N. Naveen, et al. [17] Intelligent Rainfall Prediction and Flood Alert System using Machine Learning endeavours to revolutionize rainfall prediction and flood management by employing machine learning methods. The dataset contains historical rainfall data from 1901 to 2015 for various subdivisions in India. After filling missing values with column means, the data is visualized to understand annual and monthly rainfall trends. The analysis includes a specific focus on Tamil Nadu’s rainfall patterns. For modelling, the data is prepared by grouping it by ‘SUBDIVISION’ and selecting relevant columns. A pipeline is created for preprocessing, including standardization, and Random Forest Regression is performed with hyperparameter tuning using GridSearchCV Hyperparameters, established prior to training, affect factors such as model complexity, learning rate, and regularization strength. Mean absolute error, Mean squared error, Root mean squared error, and R-squared are used to validate the model's precision.

O. M. S and A. J P, et al. [18] introduces the rainfall prediction model based on the deep learning network, convolutional long short-term memory (convLSTM) system, which promises a prediction based on the spatial-temporal patterns. The weights of the convLSTM are tuned optimally using the proposed Salp-stochastic gradient descent algorithm (S-SGD), which is the integration of Salp swarm algorithm (SSA) in the stochastic gradient descent (SGD) algorithm in order to facilitate the global optimal tuning of the weights and to assure a better prediction accuracy. On the other hand, the proposed deep learning framework is built in the MapReduce framework that enables the effective handling of the big data. The analysis using the rainfall prediction database reveals that the proposed model acquired the minimal mean square error (MSE) and percentage root mean square difference (PRD) of 0.001 and 0.0021.

Y. Chen *et al.*, [19] proposes a hazard prediction model that considers landslide triggering factors, landslide predisposing environment, and the spatial regularity of historical landslides based on multi-modal earth observation data. The proposed model has significantly improved the spatial-temporal hazard prediction performance of rainfall-induced natural terrain landslides in Hong Kong.

F. Tang, X. Wang, Z. Wang, M. Xiao, Y. Ma and Y. Wan, et al. [20] AdaBoost BP neural network displacement prediction model is proposed in this paper. This model can integrate multiple weak BP neural networks into a strong predictor, which alleviates such defects as being easy to fall into the local optimal resolution and low accuracy. Taking Xinpu landslide in Chongqing as an example, firstly, the rainfall data is analyzed, the cumulative rainfall, average rainfall and rainfall days are extracted, and the historical sliding of the monitoring point is considered as the input characteristics of the model. Then, in order to improve the operation efficiency and prediction accuracy of the model, the decision tree algorithm is used to screen out the key features, and the AdaBoost algorithm is used to improve BP neural network. Finally, the mean square error and mean absolute error are used for accuracy verification.

F. t. A. Sayyidul Laily, M. M. Amrulloh, D. D. Prasetya and F. K. Siswanto, et al .[21] research utilized the Naive Bayes algorithm for classification, implemented within RapidMiner, using datasets from the Meteorology, Climatology, and Geophysics Agency (BMKG) collected over a one-year period, from January 2021 to February 2022. Several stages were needed to determine changes in rainfall patterns, consisting of data collection, understanding, preparation, modeling, evaluation, and dissemination. The dataset includes four parameters: average temperature, humidity, maximum wind speed, and wind direction. Three training and test data scenarios were implemented to optimize model performance. The study demonstrates that the Naive Bayes algorithm within RapidMiner effectively predicts rainfall categories, including light, medium, or heavy rainfall, with a maximum accuracy of 85.71%, making it a reliable tool for classifying rainfall intensity.

S. Hu and X. Wang, et al. [22] based on variational modal decomposition (VMD), grey correlation analysis (GRA), and Elman neural network (Elman), a group sum prediction model of VMD-GRA-Elman is proposed in this paper. The raw rainfall data were first split using variational modal decomposition (VMD), then the split data were correlated using grey relational analysis, and the data were merged into high-frequency data and low-frequency data. Among them, the prediction accuracy of high-frequency data merging is obviously reduced, while the prediction accuracy of low-frequency data merging is basically unchanged, and merging can save computing time. Finally, this paper uses Elman neural network(Elman) to predict the merged low-frequency data after grey relational analysis(GRA) and the unmerged high-frequency data after variational modal decomposition(VMD), respectively.

Y. Dash, S. K. Mishra and B. K. Panigrahi, et al. [23] presents variable monsoon trends over Kerala in winter, pre-monsoon, summer monsoon and post-monsoon periods. The decreasing trends in seasonal rainfall and post-monsoon rainfall and increasing occurrence of the deficit rainfall years indicates the probable intensification of water scarcity in Kerala. Artificial intelligent methods like Single layer Feed-Forward Neural Network (SLFN) and Extreme Learning Machine (ELM) were used to predict the summer monsoon rainfall as heavy rainfall occurs during this period in the state. The proposed SLFN network prediction model provides 6.3977% of mean absolute error whereas ELM provides 3.8729% of error. ELM is better as compared to SLFN and can be used as a prediction tool for analyzing complex monsoon phenomena.

R. Pandey, M. Upadhya and M. Singh, et al .[24] develop a robust prediction model concerning possible precipitation types Australian territories. Logistic Regression provides a baseline model, while the Random Forest Classifier provides an ensemble based the technique improves accuracy. We look for the content engineering and hyperparameter tuning to improve the model performance. The results of this study are important for decision makers and stakeholders, as accurate rainfall forecasts can provide to strengthen resource management, risk reduction and sustainability planning in different areas. The results of our models is that it will contribute to a better understanding of Australian rainfall Patterns, therefore helps in effective decision making and preparation.

S. K. Sunori *et al*., [25] present paper is the rainfall prediction with knowledge of two input parameters on which the rainfall is strongly connected i.e temperature and humidity. The rainfall forecasting, in India, is a challenging task due to significantly fluctuating nature of weather here. In the present work, artificial intelligence (AI) techniques are used to train a prediction model for the forecasting of the amount of rainfall. The two considered input parameters, for this model, are temperature and humidity. The prediction models have been designed, using MATLAB, using two different AI approaches, one is the subtractive clustering and another is the Levenberg-Marquardt algorithm. Finally, their prediction performance is considered.

III. FRAMEWORK FOR PREDICTION OF RAINFALL ANALYSIS USING MACHINE LEARNING TECHNIQUE

In this section, framework for Prediction of Rainfall Analysis Using Machine Learning Technique is observed in Figure 1. The dataset is collected and then it is pre-processed to remove unwanted data. The data is normalized to prevent features with larger values from dominating the prediction and improves model convergence. Then, proposed system identifies missing values and replaces them using suitable imputation techniques such as Mean value, Median value, most frequent value and Interpolation to prevent information loss. After handles of data, Feature Reduction PCA reduces redundant information reduces dataset dimensionality and removes correlated features. Then, data is splitted into training data and testing data. Finally, catBoost technique is applied to predict the rainfall.

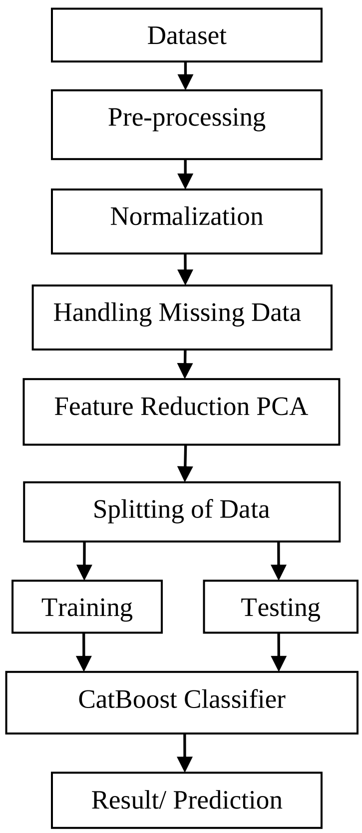

Figure 1. Framework for Prediction of Rainfall Analysis Using Machine Learning Technique

Preprocessing in rainfall prediction is the crucial first step of cleaning, transforming, and organizing raw weather data such as handling missing values, normalization, and feature selection to make it accurate and ready for machine learning models. Normalization in rainfall prediction is a data preparation step that rescales numerical weather variables to a uniform range (such as 0 to 1) using methods like Min-Max scaling or Z-score standardization. This process ensures that large-scale features do not overpower smaller-scale features, helping machine learning models train faster and more accurately.

PCA is commonly used for data preprocessing for use with machine learning algorithms. It can extract the most informative features from large datasets while preserving the most relevant information from the initial dataset. This reduces model complexity as the addition of each new feature negatively impacts model performance, which is also commonly referred to as the "curse of dimensionality." By projecting a high-dimensional dataset into a smaller feature space, PCA also minimizes, or altogether eliminates, common issues such as multicollinearity and overfitting. Multicollinearity occurs when two or more independent variables are highly correlated with one another, which can be problematic for causal modeling. Overfit models will generalize poorly to new data, diminishing their value altogether. PCA is a commonly used approach within regression analysis but it is also leveraged for a variety of use cases, such as pattern recognition, signal processing, image processing, and more.

Rainfall is an integral component of the natural environment and a crucial climatic factor for the survival of people, animals, and plants. Rainfall directly impacts industries, hydropower, home water supplies, and irrigation since it is fundamental to human growth. Rainfall forecasting continues to be of significant concern. Governments, corporations, risk management organizations, and the scientific community have all noticed it. analyzed the scaling behavior of the global tropopause, revealing significant variability and trends that enhance understanding of atmospheric dynamics and climate change impacts. Their study used advanced statistical methods to identify patterns in tropopause height and temperature over time. Rain gauge physical measurement inaccuracies result in lower-quality rainfall forecasts. The hydrological cycle depends significantly on rainfall, which is difficult to estimate using deterministic and stochastic models. Rainfall data-driven and physical models employ two methods. Data-driven models are those used for modelling and analysing a particular system based on historically available data. Data-driven models are frequently used in water research and hydrology to model different meteorological factors since they are simple to use and comprehend. It is based on simulating the relationship between the system's input and output without addressing how the system behaves physically.

CatBoost Classifier is an open-source machine learning algorithm developed by Yandex that uses gradient boosting on decision trees. It is specifically optimized for handling categorical data automatically and efficiently without requiring manual preprocessing like one-hot encoding or label encoding.

IV. RESULT ANALYSIS

In this section, result analysis of Prediction of Rainfall Analysis Using Machine Learning Technique is observed.

In Table 1, performance Comparison between classifiers such as Random Forest, Logistic Regression, XGBoost Classifier is compared with CatBoost Classifier. This CatBoost Classifier shows better results interms of accuracy, precision, recall and F1-Score.

Table 1. Performance Comparison

Parameters	CatBoost Classifier	Random Forest	Logistic Regression	XGBoost Classifier
Accuracy	0.86	0.84	0.77	0.86
Precision	0.89	0.89	0.92	0.88
Recall	0.95	0.91	0.78	0.94
F1-Score	0.92	0.90	0.84	0.91

The Figure 2, shows accuracy performance comparison between CatBoost Classifier with other models such as Random Forest, Logistic Regression, XGBoost Classifier.

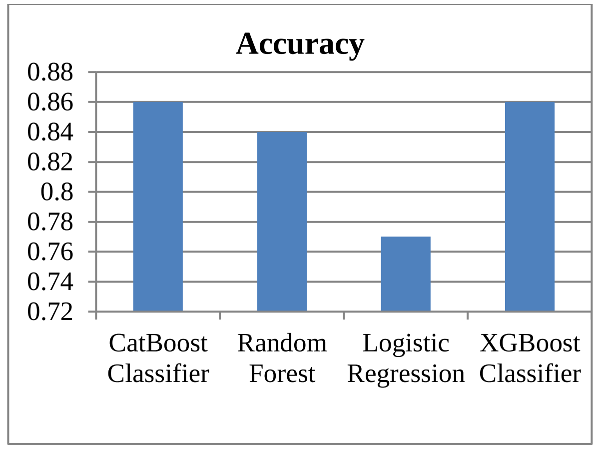

Figure 2. Accuracy Comparison Graph

The Figure 3, shows precision performance comparison between CatBoost Classifier with other models such as Random Forest, Logistic Regression, XGBoost Classifier.

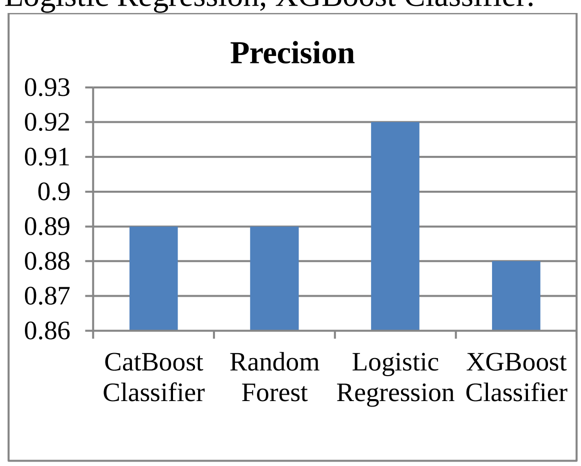

Figure 3. Precision Comparison Graph

The Figure 4, shows recall performance comparison between CatBoost Classifier with other models such as Random Forest, Logistic Regression, XGBoost Classifier. The CatBoost Classifier shows high recall value.

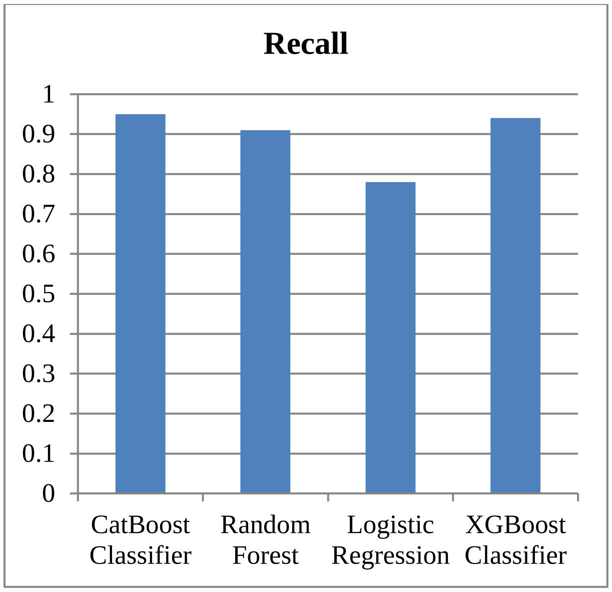

Figure 4. Recall Comparison Graph

The Figure 5, shows F1-Score performance comparison between CatBoost Classifier with other models such as Random Forest, Logistic Regression, XGBoost Classifier. The CatBoost Classifier shows high recall F1-Score.

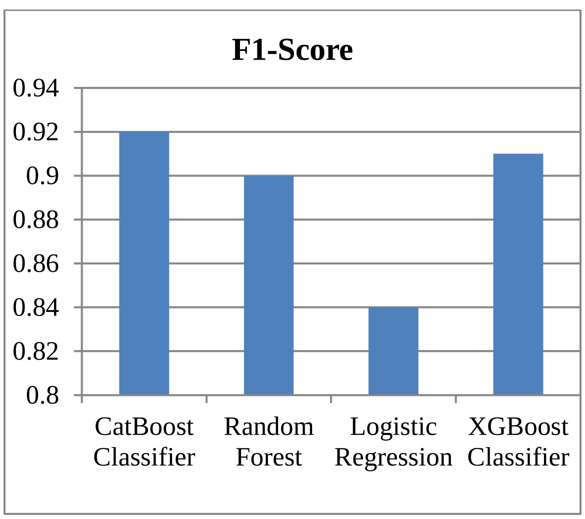

Figure 5. F1-Score Comparison Graph

V. CONCLUSION

Accurate weather prediction is essential for mitigating the impacts of natural disasters, particularly in coastal areas vulnerable to cyclones and intense rainfall. Accurate rainfall prediction is important for disaster mitigation, water management, and agricultural planning. Particularly, in areas with extreme monthly rainfall variations like Malang Regency, future weather prediction is a critical need. Classification of rainfall data could be considered to process information into valuable knowledge for the benefit of many people. This research utilized prediction of rainfall analysis using machine learning technique. This proposed system identifies missing values and replaces them using suitable imputation techniques such as Mean value, Median value, most frequent value and Interpolation to prevent information loss. After handles of data, Feature Reduction PCA reduces redundant information reduces dataset dimensionality and removes correlated features. Then, data is splitted into training data and testing data. Finally, machine learning technique is applied to predict the rainfall.

VI. REFERENCES

[1] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, “Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting,” Mach. Learn. with Appl., vol. 7, Mar. 2022, Art. no. 100204, doi: 10.1016/j.mlwa.2021.100204.

[2] D. A. Ramu, S. A. Rao, P. A. Pillai, M. Pradhan, G. George, D. N. Rao, S. Mahapatra, D. S. Pai, and M. Rajeevan, “Prediction of seasonal summer monsoon rainfall over homogenous regions of India using dynamical prediction system,” J. Hydrol., vol. 546, pp. 103–112, Mar. 2017, doi: 10.1016/j.jhydrol.2017.01.010.

[3] O. S. Ojo and S. T. Ogunjo, “Machine learning models for prediction of rainfall over Nigeria,” Scientific Afr., vol. 16, Jul. 2022, Art. no. e01246, doi: 10.1016/j.sciaf.2022.e01246.

[4] 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, Jun. 2020, Art. no. 104845, doi: 10.1016/j.atmosres.2020.104845.

[5] G. Di Baldassarre, A. Montanari, H. Lins, D. Koutsoyiannis, L. Brandimarte, and G. Blöschl, "Flood fatalities in Africa: From diagnosis to mitigation," Geophys. Res. Lett., vol. 37, no. 22, pp. 529–546, Nov. 2010, doi;10.1029/2010GL045467.

[6] R. C. Deo, S. Salcedo-Sanz, L. Carro-Calvo, and B. Saavedra-Moreno, "Drought prediction with standardized precipitation and evapotranspiration index and support vector regression models," in Integrating Disaster Science and Management. Amsterdam, The Netherlands : Elsevier, 2018, pp. 151–174, doi:10.1016/B978-0-12-812056-9.00010-5

[7] D. T. Bui, P. Tsangaratos, P.-T.-T. Ngo, T. D. Pham, and B. T. Pham, "Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods," Sci. Total Environ., vol. 668, pp. 1038–1054, Jun. 2019, doi: 10.1016/j.scitotenv.2019.02.422

[8] A. J. Schroeder, "The development of a flash flood severity index," J. Hydrol., vol. 541, pp. 523–532, Oct. 2016, 10.1016/j.jhydrol.2016.04.005

[9] F. Silvestro, "Impact-based flash-flood forecasting system: Sensitivity to high resolution numerical weather prediction systems and soil moisture," J. Hydrol., vol. 572, pp. 388–402, May 2019, 10.1016/j.jhydrol.2019.02.055

[10] J. Ritter, M. Berenguer, C. Corral, S. Park, and D. Sempere-Torres, "ReAFFIRM: Real-time assessment of flash flood impacts—A regional high-resolution method," Environ. Int., vol. 136, Mar. 2020, Art. no. 105375, 10.1016/j.envint.2019.105375

[11] S. Ravuri, "Skilful precipitation nowcasting using deep generative models of radar," Nature, vol. 597, no. 7878, pp. 672–677, Sep. 2021, doi.org/10.1038/s41586-021-03854-z

[12] Z. Sokol and P. Pešice, "Comparing nowcastings of three severe convective events by statistical and NWP models," Atmos. Res., vol. 93, nos. 1–3, pp. 397–407, Jul. 2009, 10.1016/j.atmosres.2008.09.016

[13] C. J. Short and J. Petch, "Reducing the spin-up of a regional NWP system without data assimilation," Quart. J. Roy. Meteorological Soc., vol. 148, no. 745, pp. 1623–1643, Apr. 2022, doi:10.1002/qj.4268

[14] R. McTaggart-Cowan, "Modernization of atmospheric physics parameterization in Canadian NWP," J. Adv. Model. Earth Syst., vol. 11, no. 11, pp. 3593–3635, Nov. 2019, doi:10.1029/2019MS001781

[15] P. V. S. Krithik, M. T. A. Raghavan, S. P. K and K. Vasanth, "Real Time Rainfall Prediction for Hyderabad Region using Machine learning Approach," 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), Kuala Lumpur, Malaysia, 2021, pp. 1-6, doi: 10.1109/i-PACT52855.2021.9697017.

[16] N. Tiwari and A. Singh, "A Novel Study of Rainfall in the Indian States and Predictive Analysis using Machine Learning Algorithms," 2020 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2020, pp. 199-204, doi: 10.1109/ComPE49325.2020.9200091.

[17] N. Nallametti, C. Malarvizhi, S. J. Ahmed, R. Sathya, K. H. Vishnu and J. N. Naveen, "Intelligent Rainfall Prediction and Flood Alert System using ML," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 1778-1782, doi: 10.1109/ICOSEC61587.2024.10722319.

[18] O. M. S and A. J P, "MapReduce and Optimized Deep Network for Rainfall Prediction in Agriculture," in The Computer Journal, vol. 63, no. 1, pp. 900-912, Jan. 2020, doi: 10.1093/comjnl/bxz164.

[19] Y. Chen et al., "Spatial-Temporal Hazard Prediction of Rainfall-Induced Landslides Using Multi-Modal Earth Observation Data," IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 2024, pp. 3761-3764, doi: 10.1109/IGARSS53475.2024.10640781.

[20] F. Tang, X. Wang, Z. Wang, M. Xiao, Y. Ma and Y. Wan, "Displacement prediction of rainfall induced landslide based on AdaBoost BP neural network," 2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS), Zhoushan, China, 2022, pp. 794-799, doi: 10.1109/ICGMRS55602.2022.9849366.

[21] F. t. A. Sayyidul Laily, M. M. Amrulloh, D. D. Prasetya and F. K. Siswanto, "Performance of Naive Bayes Algorithm on Extreme Rainfall Prediction," 2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C), Jember, East Java, Indonesia, 2024, pp. 154-158, doi: 10.1109/BTS-I2C63534.2024.10941710.

[22] S. Hu and X. Wang, "Monthly Rainfall Prediction Based on VMD-GRA-Elman Model," 2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Nanjing, China, 2023, pp. 95-102, doi: 10.1109/ICBASE59196.2023.10303054.

[23] Y. Dash, S. K. Mishra and B. K. Panigrahi, "Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala, India, 2017, pp. 1714-1718, doi: 10.1109/ICICICT1.2017.8342829.

[24] R. Pandey, M. Upadhya and M. Singh, "Rainfall Prediction Using Logistic Regression and Random Forest Algorithm," 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 663-668, doi: 10.1109/IC2PCT60090.2024.10486681.

[25] S. K. Sunori et al., "Rainfall Prediction using Subtractive Clustering and Levenberg-Marquardt Algorithms," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1458-1463, doi: 10.1109/ICOEI51242.2021.9452869.