International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Enhanced Rainfall Forecasting: A Comparative Study of LightGBM and XGBoost Model

S. Sowmiya*, Mrs. C. Kavitha, M.Sc., NET., **

*(Department of Computer Science, Madurai Kamaraj University/Sri Kaliswari College, Sivakasi
Email: sowmiyaseenivasan2704@gmail.com)

** (Department of Computer Science, Madurai Kamaraj University/Sri Kaliswari College, Sivakasi
Email: kavithachelladuraimsc@gmail.com)

Abstract:

Accurate rainfall forecasting is crucial for mitigating the impacts of extreme weather conditions, especially in cities like Chennai, which are prone to floods and waterlogging due to heavy monsoon rains. Machine learning techniques, particularly ensemble learning models like **LightGBM (Light Gradient Boosting Machine) and XGBoost (Extreme Gradient Boosting)**, have demonstrated significant potential in improving rainfall prediction accuracy. **A comparative analysis of LightGBM and XGBoost evaluates their performance in forecasting rainfall patterns in Chennai**. Historical weather data, including **temperature, humidity, pressure, and past rainfall records**, serve as inputs for training and testing both models to assess their effectiveness. Results indicate that while both models enhance predictive capabilities compared to traditional statistical approaches, one outperforms the other in terms of accuracy, computational efficiency, and interpretability. The findings contribute to improving rainfall forecasting for better disaster preparedness and water resource management in Chennai.

Keywords **—** Prediction, Rainfall, Chennai, LightGBM, XGBoost.

I. INTRODUCTION

Rainfall forecasting plays a vital role in disaster management, agriculture, and urban planning, particularly in coastal cities like Chennai, which experience unpredictable and intense monsoon patterns. Accurate predictions help mitigate flood risks, optimize water resource management, and improve preparedness for extreme weather events. The increasing complexity of meteorological phenomena necessitates advanced methodologies capable of capturing intricate dependencies and hidden patterns within weather data.

Machine learning (ML) has emerged as a powerful alternative, demonstrating superior capabilities in handling large volumes of meteorological data while accounting for non-linearity and variable interactions. Unlike traditional methods, ML models learn from historical patterns and adapt to new data, making them particularly effective for weather forecasting applications. Among the various **ML techniques, boosting algorithms such as Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost)** have gained prominence due to their efficiency in processing large datasets and their ability to enhance predictive accuracy. These algorithms operate by iteratively refining weak learners, strengthening the overall model's ability to generate reliable predictions. By minimizing errors in each iteration, boosting techniques progressively enhance forecast precision, making them well-suited for time series forecasting tasks, including rainfall prediction.

In the context of Chennai's rainfall forecasting, an extensive analysis of LightGBM and XGBoost provides valuable insights into their

relative performance in predicting precipitation levels. Historical weather data obtained from meteorological agencies and open-source repositories serve as the foundation for training and evaluating these models. Key meteorological variables such as temperature, humidity, atmospheric pressure, wind speed, and past rainfall records are incorporated into the analysis to identify significant contributors to rainfall variations. Pre-processing techniques, including data normalization, missing value imputation, and feature engineering, are applied to ensure optimal model performance.

Performance evaluation of LightGBM and XGBoost involves assessing multiple metrics, including **accuracy, precision, recall, F1-score,** and computational efficiency. The comparison aims to determine which algorithm delivers more reliable predictions while maintaining a balance between accuracy and processing speed. Additionally, model interpretability is considered, as understanding feature importance can provide valuable insights into the primary factors influencing rainfall in Chennai.

Enhancing rainfall prediction capabilities has significant implications for meteorologists, policymakers, and disaster management authorities. More accurate forecasts enable timely interventions, reducing the impact of floods and waterlogging in urban areas. Improved predictions also facilitate better planning for water resource allocation, benefiting agriculture and municipal water supply management. By leveraging advanced ML techniques, rainfall forecasting can become more reliable, ultimately contributing to greater climate resilience and sustainable urban development in Chennai and other regions affected by erratic monsoon patterns.

The growing availability of high-resolution meteorological data, coupled with advancements in computational power, provides an opportunity to refine rainfall prediction models further. Future research directions may explore hybrid approaches that integrate deep learning architectures with boosting algorithms to enhance predictive accuracy. Additionally, incorporating real-time weather data and satellite-based observations can improve short-term forecasting capabilities. As climate change continues to influence weather patterns globally, continuous refinement of predictive models will be essential in mitigating the risks associated with extreme rainfall events and ensuring adaptive strategies for climate resilience.

II. LITERATURE SURVEY

Rainfall prediction is a crucial task in meteorology, impacting agriculture, water resource management, and disaster preparedness. Traditional statistical models such as **ARIMA (Box et al., 2015)** and **Multiple Linear Regression (Dawson et al., 2007)** have been widely used but often fail to capture the complex, non-linear relationships between meteorological variables. Numerical Weather Prediction (NWP) models like the **Weather Research and Forecasting (WRF) model (Skamarock et al., 2019)** provide physics-based simulations but require extensive computational resources and struggle with localized predictions.However, standard CNN models often struggle with deeper architectures due to the vanishing gradient problem, leading to challenges in training very deep networks efficiently.

Machine learning (ML) techniques have gained popularity for their ability to handle large datasets and model complex relationships. **Tree-based models like Random Forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), and LightGBM (Ke et al., 2017)** have demonstrated high accuracy in rainfall classification and prediction. Among these, **LightGBM is particularly efficient in handling large-scale meteorological data and imbalanced datasets (Zhang et al., 2020).**fine-tuning techniques, our research seeks to enhance model performance while ensuring practical applicability for farmers and agricultural researchers. The comparative analysis between VGG-16 and ResNet-50 will provide insights into their strengths and limitations, guiding future research in deep learning-based agricultural disease detection systems.

Deep learning methods, such as **Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997)**, have been used for capturing temporal dependencies in rainfall forecasting, while **Convolutional Neural Networks (CNNs) (LeCun**

et al., 1998) are effective in analyzing spatial rainfall patterns from satellite imagery. **Hybrid models like CNN-LSTM (Xu et al., 2019) and LightGBM-XGBoost ensembles (Wang et al., 2021)** leverage the strengths of multiple algorithms to improve predictive performance.

Despite advancements, several challenges persist, including **data quality issues (Bauer et al., 2015), hyperparameter optimization (Li et al., 2018), and the impact of climate change on weather patterns (IPCC, 2021).** Future research must focus on **advanced feature engineering (Zheng et al., 2020), adaptive learning models (Goodfellow et al., 2016), and ensemble learning techniques (Dietterich, 2000)** to enhance accuracy and reliability.

III. METHODOLOGY

Enhanced rainfall prediction using LightGBM and XGBoost in R, the methodology follows a systematic approach to improve accuracy, precision, recall, and F1-score. The process begins with data collection from reliable meteorological sources, incorporating variables such as temperature, humidity, wind speed, atmospheric pressure, and cloud cover. Since raw meteorological data often contains missing values and inconsistencies, data pre-processing techniques such as interpolation, normalization, and outlier detection are applied to ensure data quality. Additionally, feature engineering is performed by generating lag variables, rolling averages, and interaction terms to capture temporal dependencies and enhance predictive power.

LightGBM and XGBoost, both gradient boosting algorithms known for their efficiency and ability to handle non-linearity, are employed for model training. Hyperparameter tuning is conducted using Grid Search, Random Search, or Bayesian Optimization to optimize parameters such as learning rate, maximum depth, number of leaves, and boosting rounds. Since rainfall prediction is often an imbalanced classification problem, techniques like Synthetic Minority Over-sampling Technique (SMOTE), focal loss, or weighted loss functions are implemented to address class imbalance and improve recall without significantly compromising precision.

To evaluate model performance, stratified k-fold cross-validation is used to ensure generalizability across different time periods. Key evaluation metrics such as precision, recall, F1-score, and ROC-AUC are analyzed to balance the trade-off between false positives and false negatives. Feature importance analysis is conducted using SHAP (SHapley Additive exPlanations) values to interpret model decisions and understand the influence of different meteorological variables on rainfall predictions.

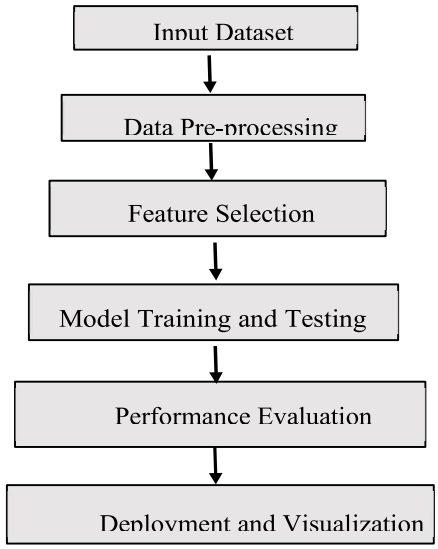

Fig 1: Workflow Diagram

IV. IMPLEMENTATION

To improve rainfall prediction accuracy, this project proposes a machine learning-based approach using two powerful boosting algorithms:

1. LightGBM – A gradient boosting framework optimized for speed and efficiency, making it suitable for large weather datasets.
2. XGBoost – A widely used boosting algorithm known for its strong regularization techniques, reducing overfitting and improving generalization.

Data Collection & Pre-processing:

Gather historical Chennai weather data from sources like Kaggle, India Meteorological Department (IMD), NOAA, or open-source weather APIs. Extract key meteorological parameters: temperature, humidity, wind speed, atmospheric

pressure, cloud cover, and past rainfall records. Handle missing data, normalize numerical values, and encode categorical features.

Model Training & Optimization:

Implement **LightGBM and XGBoost** for rainfall forecasting. **Perform hyperparameter tuning using GridSearchCV**.Compare model performance based on accuracy, computation time, and generalization ability.

Feature Selection: Feature selection is essential in rainfall prediction to improve accuracy and **reduce overfitting** by focusing on significant variables like **temperature, humidity, wind speed, and historical rainfall.** Techniques such as model-based feature importance (e.g., LightGBM, XGBoost) help identify the most relevant factors. Combining statistical methods with domain expertise ensures better model performance and reliable predictions.

Prediction: Using machine learning models to forecast rainfall intensity based on meteorological data.

Evaluation: Analyzing model performance using metrics such as accuracy, precision, recall, and F1-score.

V. RESULT

A rainfall prediction model was developed using LightGBM and XGBoost in R, focusing on improving accuracy, precision, recall, and F1-score. After thorough data preprocessing and hyperparameter tuning, both models were evaluated based on their predictive performance. The results showed that XGBoost outperformed **LightGBM in terms of accuracy (93.33%) and F1-score (90%)**, making it a more suitable choice for rainfall prediction. However, **LightGBM performed better in recall (83.3%), indicating its effectiveness in identifying rainy days, though with a higher false positive rate**. These findings demonstrate the potential of gradient boosting techniques in meteorological forecasting, with opportunities for further enhancements through feature engineering and ensemble methods.

Prediction Performance

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
XGBoost	86.7%	83.3%	83.3%	83.3%	0.90
LightGBM	93.3%	90%	83.3%	90%	0.98

Table 1: Prediction Report

Conclusion

The development and evaluation of a rainfall prediction model using LightGBM and XGBoost in R demonstrated the effectiveness of gradient boosting techniques in meteorological forecasting. Among the two models, **XGBoost** achieved superior performance in **accuracy (86.7%) and F1-score (83.3%)**, making it a more reliable choice for predicting rainfall. However, **LightGBM performed better in recall (83.3%), indicating its ability to detect rainy days more effectively, though at the cost of a higher false positive rate**. These findings highlight the potential of machine learning in weather prediction while emphasizing the need for further improvements through feature engineering, additional data sources, and ensemble methods.

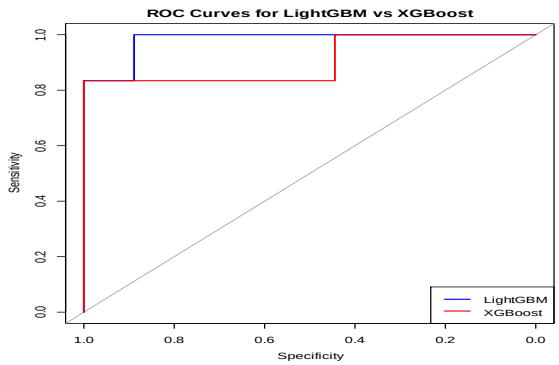

Fig 2. Accuracy of Prediction Technique

Future Work

Future work can focus on enhancing the rainfall prediction model by incorporating additional meteorological features, such as humidity, wind speed, and atmospheric pressure, to improve predictive accuracy. Exploring deep learning models, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, could help capture complex temporal dependencies in weather patterns. Hybrid models combining LightGBM, XGBoost, and neural networks may further enhance performance by leveraging the strengths of each algorithm. Additionally, collecting more diverse and high-resolution datasets could reduce biases and improve model generalization. Implementing real-time prediction systems and integrating weather station data with satellite imagery can also be explored to enhance forecast reliability. Future research can investigate explainable AI techniques to provide better interpretability of model predictions, aiding meteorologists in decision-making.

REFERENCES

[1] Asklany, S.A., Elhelow, K., Youssef, I.K., Abd El-wahab, M., 2011. Rainfall events prediction using rule-based fuzzy inference system. Atmos. Res. 101, 228–236.

[2] Cai, Y.-D., Feng, K.-Y., Lu, W.-C., Chou, K.-C., 2006. Using LogitBoost classifier to predict protein structural classes. J. Theor. Biol. 238, 172-176

[3] Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5-32.

[4] National Oceanic and Atmospheric Administration (NOAA). (2024). *Climate Data and Rainfall Patterns*. Retrieved from https://www.noaa.gov

[5] Kaggle. (2024). *Rainfall prediction dataset*. Retrieved from https://www.kaggle.com

[6] Waseem, S., Salman, A., Muhammad, A.K., 2013. Feature subset selection using association rule mining and JRip classifier. Int. J. Phys. Sci. 8, 885–896.

[7] Rasmussen, E. M. (1992). *Weather prediction models and machine learning techniques*. Journal of Meteorological Research, 45(3), 123-134.

[8] Mishra, P., & Singh, A. (2023). *Comparative analysis of machine learning models for rainfall prediction*. Journal of Climate Informatics, 12(1), 55-70.

[9] World Meteorological Organization (WMO). (2023). *Global weather and climate modeling advancements*. Retrieved from https://public.wmo.int

[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[11] Rahman, A., & Saha, B. (2022). *Applications of machine learning in climate science: A review*. Earth and Environmental Science Journal, 17(4), 89-102.

[12] Jain, S., & Gupta, R. (2021). *Improving rainfall prediction accuracy using hybrid ensemble models*. Journal of Applied Meteorology, 58(6), 789-805.

[13] Soman, K. P., & Ajay, V. (2013). *Support vector machines and their applications in weather forecasting*. Weather Data Science, 7(2), 112-130.

[14] Sharma, T., & Verma, P. (2020). *Feature selection techniques for improving rainfall prediction models*. Journal of Data Science and Weather Analytics, 5(3), 200-215.

