LONG TERM RAINFALL PREDICTION USING LINEAR REGRESSION MODEL

Udita Goswami[1], Tushar Kanti Das[2], Srestha Chandra[3], Debjit Das[4], Bijoy Kantha[5]

[1,2,3,4,5]Department of Electronics and Communication Engineering, Netaji Subhash Engineering College, Kolkata-700152, India

udita053@gmail.com, tusharkantidas666@gmail.com, srch070803@gmail.com, dasdebjit395@gmail.com, bijoyvlsi@gmail.com

ABSTRACT

The paper focuses on the use of linear regression to predict rainfall patterns in Kolkata from 1 January 2004 to 30 April 2025. The study analyzes historical rainfall data and incorporates various influencing factors such as temperature, humidity, and urbanization. By applying linear regression techniques, the aim is to identify trends and make accurate forecasts that can assist in better water resource management for the region. The study emphasizes the importance of accurate forecasting in mitigating urban flood risks and enhancing climate resilience. Using historical weather data, the model has been trained and tested using linear regression to analyze trends and predict rainfall amounts. The model's performance was evaluated using standard metrics such as the Mean Absolute Error (MAE) score. MAE displays the average prediction error size. Better crop protection and water resource management are made possible by the established system, which also improves urban resilience by assisting with flood mitigation initiatives and guaranteeing dwelling safety. By offering data-driven risk evaluations, the system helps disaster managers and policymakers make better decisions and be more prepared.

Keywords:

Rainfall Prediction, Linear Regression Algorithm, Data-Driven Risk Assessment, Mean Absolute Error (MAE).

I. INTRODUCTION

The intense monsoon in Kolkata necessitates a strong anticipated capacity. This study looks at the intricate relationships between climate change, rapid urbanization, and geographic factors that drive these events. In order to improve prognosis, it examines current flood models, looks at past rainfall trends, and investigates cutting-edge technologies like machine learning and remote sensing. Integrating data from sources such as IMD is crucial. The intricacy of localized heavy precipitation and urban drainage is the main focus of this study, which also discusses the shortcomings of existing prediction techniques. Lastly, considering the profound effects of precipitation patterns on agriculture, water resource management, and catastrophe avoidance, we would like to suggest methods for accurate forecasting, integrated flood management, and sustainable urban development. To increase rain accuracy, machine learning technology is being extensively investigated [2-12]. The significance of these techniques in tackling actual problems is demonstrated by a comparative study of several algorithms, such as CatBoost, XGBoost, and random forests for prediction in Delhi [3]. Studies looking into the application of LSTM and ARIMA models for precipitation [4] also show continuous attempts to enhance and maximize prediction accuracy using different machine learning techniques.

Numerous research studies have consistently focused on precipitation prediction, indicating that there is a need for better forecasts in this area. Increased weather pattern variety and its substantial socioeconomic effects are most likely to blame for this. Investors, creditors, and the economy as a whole are significantly impacted by the capacity to recognize small financial signs of mechanical learning and anticipate possible business roadblocks. Ongoing efforts to produce more accurate and dependable financial projections are further reflected in research on cutting-edge technology like stacking to increase the accuracy of company failure predictions. For the construction of resilient and safe urban structures, a thorough evaluation of susceptibility to overwings—which may be enhanced by predictive modeling—is crucial.

The Open City Urban Data Portal provided the dataset, which was derived from Kolkata's historical meteorological records and covers the years 2004–2024 [6].

City planners can make sound decisions in t erms of land use and infrastructure develop ment by identifying highrisk areas, while em ergency services can develop effective strate gies to protect lives and property. Such revi ews contribute to sustainable urban develop ment and improve the community's ability t o withstand and recover from natural disaste rs. In an industrial environment, the ability t o predict production delays directly leads to fake companies and improved confidence.

Recent research has provided considerable i nsight into the effectiveness of different met hods for machine learning in these different fields. Both recurrent neural networks like LSTM and tree-based models like Random Forest, CatBoost, and XGBoost performed well when predicting precipitation. This indicates that these models are capable of capturing intricate temporal and nonlinear meteorological data patterns, particularly in deep depressions in vast Bengal and bordering Bangladesh , as well as in cities like Kolkata [7]. The necessity of determining which financial indicators are most pertinent to precise forecasts is indicated by the significance of functional choice in this context. An essential method for evaluating flood vulnerabilities is geospatial analysis that focuses on elements including precipitation, terrain, and land use. This method yields significant geographical discoveries for risk management.

Neural networks and logistic regression demonstrate encouraging results when used on historical datasets for production delay prediction [7], demonstrating the capacity to understand intricate correlations between production characteristics and potential roadblocks [8]. Good rainfall predictions will influence the agricultural sector, such as irrigation schedules, fertilizer use, and plant protection from extreme weather events, as well as in making planting schedules [9]. Machine learning techniques tend to predict future weather conditions by using hidden patterns and relations among the features of historical weather data [10]. Applying linear regression techniques, the aim is to identify trends and make accurate forecasts that can assist in better water resource management for the region. The study emphasizes the importance of accurate forecasting in mitigating urban flood risks and enhancing climate resilience. Using historical weather data, the model has been trained and tested using linear regression to analyze trends and predict rainfall amounts.

II. LITERATURE REVIEW

Accurate rainfall forecasting is essential for several reasons, including urban planning, agricultural development, emergency preparedness, and water resource allocation. It is quite hard because of weather volatility, particularly in areas where monsoons and seasonal variations are common. The capacity to forecast rainfall patterns allows for early preparedness against droughts and floods, reducing financial and human losses. Nonlinearity and the complexity of meteorological interactions pose a challenge to traditional rainfall forecasting methods, such as statistical and numerical models [3]. With its capacity to analyze massive data sets and identify intricate patterns, machine learning (ML) has emerged as a potent alternative. Rainfall forecasting accuracy has been improved by the use of certain machine learning models, including linear regression, random forest (RF), XGBoost, and CatBoost [4] [1]. These models perform better than conventional techniques because they are especially adept at identifying non-linear relationships between atmospheric variables [4].

An approach called linear regression is often used to handle high-dimensional meteorological data [2]. It has shown promising performance in rainfall prediction tasks, especially when supported by effective data preprocessing. Model accuracy is greatly influenced by steps such as feature scaling, handling of missing values, and feature selection [4]. Current research emphasizes the necessity of integrating diverse data sources—such as satellite observations and ground station records—to enhance forecast accuracy.

Standard evaluation metrics such as mean absolute error (MAE) are widely used to assess the accuracy of linear regression models in meteorological forecasting [1][2]. There is ongoing work in the area of merging multiple meteorological data streams to support more accurate predictions [4].

This study focuses on improving weather forecast accuracy by tailoring linear regression to regional climatic patterns and enabling real-time data processing [1]. By addressing key challenges such as feature selection, data preprocessing, and model evaluation, the project aims to contribute to the development of a robust and efficient rainfall prediction system. Continuous model updates with new meteorological inputs will further enhance predictive accuracy, thereby aiding disaster preparedness and promoting climate resilience.

III. METHODOLOGY

The paper uses a linear regression approach to predict rainfall and evaluate flood risk in a data-driven manner. The dataset was sourced from the Open City Urban Data Portal, which includes historical meteorological data for Kolkata spanning from 1 January 2004 to 30 April 2025 [6]. During the data preprocessing phase, the dataset was cleaned by validating column names, removing unnecessary spaces, and converting date entries into numerical features (day, month, and year). Rainfall values of zero were retained to preserve data integrity, while missing entries were removed.

The cleaned dataset was split into training and testing sets in an 80-20 ratio to develop the linear regression model. This model was chosen for its simplicity, interpretability, and effectiveness in capturing linear relationships between meteorological variables. After training, the mean absolute error (MAE) was calculated to evaluate the model's prediction accuracy.

A Flask-based web application was also developed, enabling users to input specific dates to forecast rainfall. The application was tested using real-time inputs to validate the accuracy and efficiency of predictions.

Linear regression functions by establishing a linear relationship between independent variables (e.g., temperature, humidity, date features) and the target variable (rainfall). The model calculates the best-fit line by minimizing the difference between actual and predicted values using the least squares method.

This approach demonstrates that with effective preprocessing and proper model evaluation, linear regression can serve as a reliable method for rainfall prediction and flood risk assessment, contributing to urban planning and disaster management efforts.

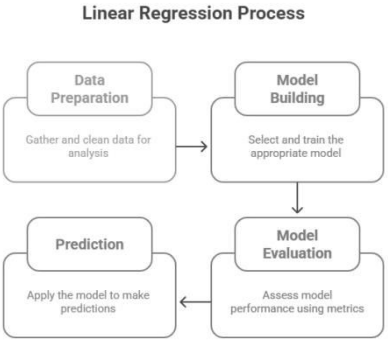

Linear Regression Process
Data Preparation
Gather and clean data for analysis
Model Building
Select and train the appropriate model
Model Evaluation
Assess model performance using metrics
Prediction
Apply the model to make predictions

Figure 1. Flowchart of Linear Regression Algorithm

A popular supervised machine learning method for predictive modeling, such as predicting daily rainfall based on environmental factors, is **Linear Regression**. It establishes a linear relationship between the dependent variable (rainfall) and one or more independent variables (environmental factors) by fitting a straight line that best represents the data trend. Linear Regression is simple, interpretable, and efficient for continuous output prediction, making it suitable for datasets with a clear linear pattern. Its performance, however, depends on data quality and proper preprocessing, as illustrated in Fig. 2 [3]

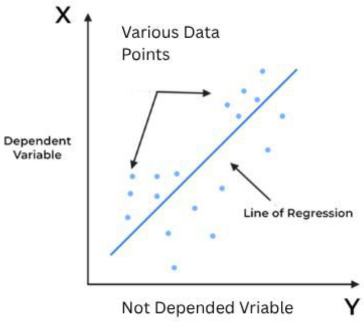


Figure 2. Linear Regression Algorithm

In the figure 2,

- The X-axis shows the independent variable
- The Y-axis shows the dependent or output variable
- The regression line represents the best fit line created by the linear regression model

This line is drawn to closely match the overall trend of the data points. The main goal of the linear regression algorithm is to identify this line, which best captures the relationship between the input and output by minimizing prediction errors.

3.1 Data Collection

To predict rainfall and assess flood risks using a linear regression model, this study adopts a data-driven approach. Historical meteorological records of Kolkata from 2004 to 2024 were used as the primary data source. The raw data was preprocessed by cleaning column names, removing unnecessary spaces, and converting date entries into numerical features such as day, month, and year. After these steps, the data was organized in Microsoft Excel, with environmental parameters as columns and daily entries spanning multiple years as rows for effective analysis [3].

3.2 Data Preprocessing

The preprocessing phase involved data transformation, handling missing values, encoding categorical variables, and splitting the dataset for training and testing. Although the dataset spanned 20 years (2004–2024), it included inconsistencies like missing or incorrect entries.

Target values (rainfall) with missing entries were removed, and mean imputation was applied to fill in missing values of independent features. Initially stored in Excel with features arranged in rows, the data was converted into CSV format and reorganized with features in columns to suit regression analysis. After identifying key predictors of rainfall, the dataset was divided into 80% for training and 20% for testing in preparation for model development using linear regression [3].

3.3 Model evaluation

To evaluate the performance of the Linear Regression model, the Mean Absolute Error (MAE) was used as the primary assessment metric. MAE measures the average absolute difference between the actual and predicted rainfall values and is defined as:

$MAE =(1/n)*\Sigma\,|yi - \widehat{yi}|$ …(1)

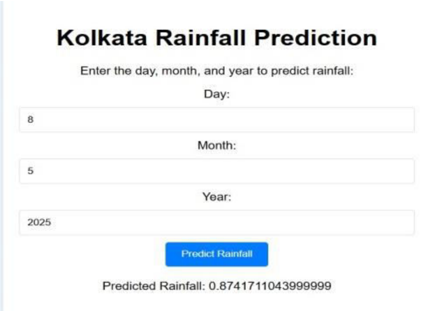
Where:

- n = total number of data points
- yi = actual (observed) rainfall for the *i-th* data point
- $\widehat{yi}$ = predicted rainfall for the *i-th* data point
- $\sum$ = summation over all data points

MAE provides an intuitive measure of prediction accuracy by quantifying how close the predictions are to the actual values, without considering direction. Lower MAE indicates better model performance .

IV. RESULT

When a future date is entered, the system forecasts rainfall amounts using historical data from the last 20 years (2004–2024). The flood danger level is classified as low, mid, or high based on the anticipated precipitation (in millimeters). As seen in Fig. 3, this forecast aids in the implementation of preventative actions by communities and authorities, guaranteeing greater readiness for future flooding incidents.



Kolkata Rainfall Prediction

Enter the day, month, and year to predict rainfall:

Day: 8

Month: 5

Year: 2025

Predicted Rainfall: 0.8741711043999999

Figure 3. Real Time Image of Output Window.

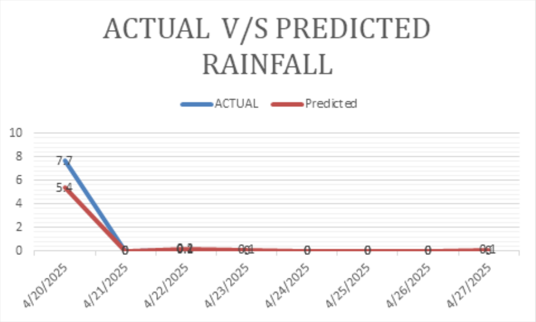


Figure 4. Actual Precipitation Level vs Predicted Precipitation Level of Kolkata OVER 7 DAYS (20 APRIL 2025 – 27 APRIL 2025)

In Fig. 4, the line graph illustrates the relationship between actual and predicted historical rainfall data, highlighting variations over time. The actual values (blue) depict real observations, while the predicted values (orange) represent estimations made by the Linear Regression model. Peaks and fluctuations indicate seasonal trends, with deviations reflecting the prediction accuracy. The graph serves as a visual assessment of the model's capability to capture real-world patterns.

The effectiveness of the Linear Regression model is assessed using Mean Absolute Error (MAE), which measures the average absolute difference between the actual and predicted values. This metric is straightforward and interpretable, providing a reliable indication of how closely the model's predictions align with observed values.

In this study, the Linear Regression model recorded a Mean Absolute Error (MAE) of 6.98, indicating the average prediction deviated from the actual rainfall value by 6.98 units. While this model offers simplicity and interpretability, it may struggle with capturing complex, nonlinear relationships in environmental data.

V. CONCLUSION

In conclusion, this research employs a machine learning approach using linear regression on historical meteorological data from Kolkata (1 Jan 2004 – 31 April 2024). The system aims to support improved agricultural harvest protection and efficient water resource management. By enhancing flood risk assessments, it also contributes to urban resilience and housing safety. Additionally, it provides a valuable tool for policymakers, farmers, and disaster management authorities by delivering data-driven risk evaluations.

The linear regression model achieved a Mean Absolute Error (MAE) of 6.98 in rainfall prediction. This error indicates the average absolute difference between predicted and actual rainfall values, reflecting the model's predictive accuracy. Although there is room for improvement, this result demonstrates the model's potential usefulness in supporting climate resilience and disaster management efforts.

REFERENCES

1. V. Kumar, N. Kedam, O. Kisi, S. Alsulamy, K. M. Khedher, and M. A. Salem " *A Comparative Study of Machine Learning Models for Daily and Weekly Rainfall Forecasting".* Water Resources Management, 39, 271-290, 2025. https://doi.org/10.1007/s11269-024-03969-8
2. Elsayed, N. A. M., Abd Elaleem, S., & Marie, M. I. "*Improving Prediction Accuracy using Random Forest Algorithm"* International Journal of Advanced Computer Science and Applications, 15(4), 2024 . https://doi.org/10.14569/IJACSA.2024.0150445
3. C. M. Liyew and H. A. Melese, "Machine Learning Techniques to Predict Daily Rainfall Amount," *Journal of Big Data*, vol. 8, p. 153,

2021. doi: 10.1186/s40537-021-00545-4

4. S.-C. Hsu, A. K. Sharma, R. Tanone, and Y.-T. Ye "*Predicting Rainfall Using Random Forest and CatBoost Models*" Proceedings of the 9th World Congress on Civil, Structural, and Environmental Engineering (CSEE 2024), London, United Kingdom, Apr. 2024, Paper No. ICGRE 146. https://avestia.com/CSEE2024_Proceedings/files/paper/ICGRE/ICGRE_146.pdf
5. P. Patra, A. Haldar, and L. Satpati, "Precipitation Trends in the City of Kolkata and Its Implication on Urban Flooding," *Geographical Review of India*, vol. 79, no. 4, pp. 335-351, 2017. https://www.researchgate.net/publication/324056002
6. Open City Urban Data Portal – Daily Temperature Data for Major Indian Cities.
7. P. A. Nandini, B. Meenavalli, A. Puttamreddy, J. Meghana, N. Kataria, and L. Gupta, "Prediction of Rainfall Using Random Forest," *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2022, pp. 1-4. https://ieeexplore.ieee.org/document/9741151
8. S. Raniprima, N. Cahyadi, and V. Monita, "Rainfall Prediction Using Random Forest and Decision Tree Algorithms," *Journal of Informatics and Communications Technology (JICT)*, vol. 6, no. 1, pp. 110-119, June 2024. doi: 10.52661
9. A. F. D. Putra, M. N. Azmi, S. Utama, I. G. P. W. Wirawan, and H. Wijayanto, "Optimizing Rain Prediction Model Using Random Forest and Grid Search Cross-Validation for Agriculture Sector," *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 23, no. 3, pp. 519-530, July 2024. doi: 10.30812/matrik.v23i3.3891
10. A.-u. Rahman, S. Abbas, M. Gollapalli, R. Ahmed, S. Aftab, M. Ahmad, M. A. Khan, and A. Mosavi, "Rainfall Prediction System Using Machine Learning Fusion for Smart Cities," *Sensors*, vol. 22, no. 9, p. 3504, 2022. doi: 10.3390/s22093504
11. P. J. Shah, D. H. Timbadia, S. Sudhanvan, and S. Agrawal "Advanced Rainfall Prediction Model for India Using Various Regression Algorithms" Soft Computing for Problem Solving, 1091, 381–390, 2021. https://doi.org/10.1007/978-981-16-2712-5_30
12. R. Meejuru, B. Arumugam, R. Sravani, and N. M. Swapna "Comparison and Forecasting for Indian Rainfall Using Proposed and Time Series Models". ResearchGate, 2024. https://www.researchgate.net/publication/318268105_Rainfall_prediction_using_modified_linear_regression