IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Rainfall Prediction using Machine Learning with Explainable AI

Monika Ghodekar[1], Vishwajeet Miragane[2], Shubham Musale[3], Sanket Vedpathak[4], Vivek Chirmade[5]
Assistant Professor, Computer Engineering[1]
Students, Computer Engineering[2-5]
All India Shri Shivaji Memorial Society's College of Engineering, Pune
Savitribai Phule Pune University, Pune, India

Abstract:. *Accurate rainfall prediction is essential for agriculture, disaster management, and water resource planning. However, traditional forecasting methods often fail to capture complex and nonlinear relationships among atmospheric variables. This paper presents a machine learning–based rainfall prediction system that combines Extreme Gradient Boosting (XGBoost) with Explainable Artificial Intelligence (XAI) to deliver accurate and interpretable predictions.*
The proposed system focuses on regression-based rainfall prediction and incorporates quantile-based probabilistic outputs to represent uncertainty in weather conditions. A baseline model using Stochastic Gradient Descent (SGD) is also implemented for performance comparison. To enhance transparency, SHAP-based explanations are used to interpret model predictions.
The system demonstrates low prediction error with MAE in the range of 0.003–0.039 mm and RMSE between 0.05–0.36 mm, indicating high accuracy.
The integration of XAI and probabilistic forecasting transforms the model into a reliable decision-support tool suitable for real-world applications.

Keywords: Rainfall Prediction, Machine Learning, XGBoost, Explainable AI, SHAP, Quantile Regression, Weather Forecasting

I. INTRODUCTION

Rainfall prediction plays a critical role in agriculture, disaster management, and climate monitoring. In countries like India, where agricultural productivity is highly dependent on monsoon patterns, accurate rainfall forecasts can significantly impact economic stability and food security.

Traditional weather forecasting methods rely on statistical models and numerical weather prediction systems. While these approaches provide useful insights, they often struggle to accurately model nonlinear and dynamic atmospheric interactions. As a result, prediction accuracy is limited, especially under rapidly changing climatic conditions.

Machine learning (ML) has emerged as a powerful alternative for rainfall prediction. ML models can learn complex relationships from historical data and adapt to changing patterns. Among these models, XGBoost has gained popularity due to its efficiency, scalability, and ability to handle structured data effectively.

Despite their advantages, ML models often operate as black boxes, making it difficult to interpret predictions. This lack of transparency reduces trust, particularly in critical applications such as disaster warning systems. To address this issue, Explainable Artificial Intelligence (XAI) techniques are used to provide insights into model behavior.

This paper proposes a rainfall prediction system that integrates XGBoost with SHAP-based explanations and probabilistic forecasting. The system not only predicts rainfall but also explains the underlying factors influencing the prediction.

DOI: 10.48175/568

II. LITERATURE REVIEW

Rainfall prediction has been widely studied using various approaches ranging from statistical models to advanced machine learning techniques.

Early studies relied on regression-based models, which assumed linear relationships among meteorological variables. However, these models were unable to capture the complex interactions between parameters such as humidity, temperature, and atmospheric pressure.

Machine learning approaches such as Decision Trees, Support Vector Machines (SVM), and Random Forest improved prediction accuracy by modeling nonlinear relationships. Random Forest, in particular, is known for its robustness and resistance to overfitting.

More advanced ensemble methods, such as XGBoost, have shown superior performance due to their gradient boosting mechanism. XGBoost builds models sequentially, reducing prediction errors and improving generalization.

Deep learning models such as Long Short-Term Memory (LSTM) networks have also been explored for rainfall prediction. These models are capable of capturing temporal dependencies but require large datasets and high computational resources.

A major limitation of many existing approaches is the lack of interpretability. To overcome this, XAI techniques such as SHAP and LIME have been introduced. These methods provide insights into feature importance and model behavior.

Comparative Analysis of Existing Methods

Model	Advantages	Limitations
SVM	Effective for classification	Poor scalability
Random Forest	Robust, handles nonlinear data	No uncertainty estimation
XGBoost	High accuracy, efficient	Less interpretable
LSTM	Captures temporal patterns	High computational cost

Rainfall prediction is inherently a complex problem due to the chaotic nature of atmospheric systems. The interaction between temperature, humidity, pressure, and wind does not follow a simple linear pattern but rather a nonlinear and dynamic process influenced by spatial and temporal variations.

Traditional statistical models assume fixed relationships between variables, which limits their ability to adapt to changing climatic conditions. In contrast, machine learning models learn these relationships directly from data, enabling them to capture hidden dependencies and interactions.

Ensemble learning methods such as XGBoost are particularly effective because they iteratively refine predictions by correcting previous errors. This allows the model to focus on complex patterns that simpler models fail to capture.

The integration of Explainable Artificial Intelligence (XAI) adds another important dimension by transforming the model from a black box into an interpretable system. This allows users to understand the reasoning behind predictions, improving trust and usability.

III. SYSTEM ARCHITECTURE

The proposed rainfall prediction system has been implemented as a multi-layer pipeline that processes real-time meteorological data and generates accurate and interpretable rainfall predictions. The architecture integrates machine learning, probabilistic modeling, and Explainable Artificial Intelligence (XAI) into a unified framework.

The system is designed and tested using real-world weather data and consists of the following major components:

• Data Acquisition Layer
• Data Preprocessing and Feature Engineering Layer
• Machine Learning Model Layer
• Prediction Module

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

• Explainable AI (XAI) Layer
• Visualization and Decision Support Layer

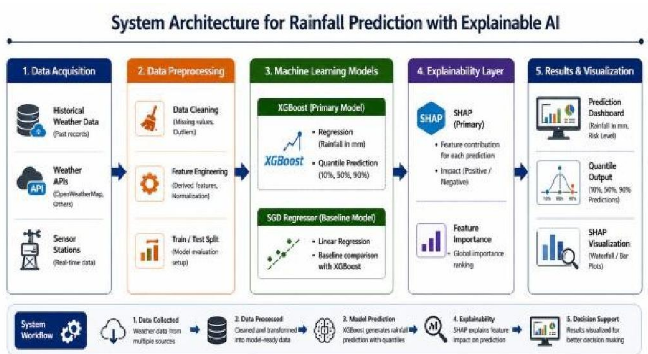

Fig 1. a. System Architecture

a. Data Acquisition Layer

The system collects meteorological data from external sources, including weather APIs and historical datasets. Unlike traditional models that use a fixed dataset, the proposed system introduces an adaptive data selection mechanism based on user input.

The user can dynamically control the amount of historical data used for model training through a configurable parameter called "Historical Memory (Years)". This allows the system to train the model using data from the past 1 to 5 years.

b. Data Preprocessing and Feature Engineering

The collected data is processed to ensure quality and consistency before being used for model training and prediction. This stage includes:

• Data Cleaning: Handling missing values and removing outliers
• Feature Engineering: Creating derived features such as normalized values and domain-specific indicators
• Train-Test Split: Dividing the dataset for model training and evaluation

These steps improve the reliability of the input data and enhance the learning capability of the model.

c. Machine Learning Models Layer

The core prediction engine of the system is implemented using machine learning models.

• XGBoost (Primary Model): Used for regression-based rainfall prediction. The model predicts rainfall in millimeters and also generates quantile outputs (10%, 50%, 90%) to represent uncertainty.

• SGD Regressor (Baseline Model): Used for comparison with the primary model. It provides a linear approximation and helps evaluate the performance improvement achieved by XGBoost.

The results from implementation showed that XGBoost performs significantly better due to its ability to model nonlinear relationships among atmospheric variables.

d. Prediction Module

The prediction module processes the output from the trained model and generates:

• Rainfall prediction in millimeters
• Quantile-based outputs representing optimistic, median, and pessimistic scenarios

This allows the system to provide both precise predictions and uncertainty estimates.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, June 2026 Impact Factor: 8.2

e. Explainable AI (XAI) Layer

To enhance transparency, the system integrates Explainable Artificial Intelligence using SHAP. This layer provides:

• Feature contribution for each prediction
• Positive and negative impact of input variables
• Global feature importance ranking

The SHAP analysis helps in understanding why a particular rainfall prediction was generated, making the system more interpretable and trustworthy.

This layer enhances transparency and helps users understand the influence of different features on the prediction outcomes.

f. Results and Visualization Layer

The final stage of the system presents outputs through an interactive dashboard.

The dashboard includes:

• Predicted rainfall value (in mm)
• Risk level classification (e.g., negligible rainfall)
• Quantile-based prediction distribution
• SHAP-based feature visualization

This layer ensures that technical outputs are converted into actionable insights for users.

g. System Workflow

1. User selects location and historical memory (1–5 years)
2. System retrieves corresponding historical dataset from API
3. Data is preprocessed and used for model training
4. Prediction is generated using trained model
5. Results are explained using SHAP
6. Output is displayed on dashboard

IV. SYSTEM METHODOLOGY

The proposed rainfall prediction system follows a structured methodology that integrates data collection, preprocessing, model training, prediction, and explainability into a unified workflow. The methodology is designed to reflect the actual implementation of the system using real-world meteorological data.

a. Data Collection and Adaptive Selection

Meteorological data is collected from external weather APIs and historical datasets. The system includes an adaptive mechanism that allows users to select the amount of historical data used for training through the "Historical Memory (Years)" parameter.

Based on user input, the system dynamically selects data from the past 1 to 5 years. This allows the model to adapt to:

• Short-term trends (recent data)
• Long-term seasonal patterns (extended data)

This adaptive data selection improves model flexibility and ensures that predictions remain relevant under changing climatic conditions.

b. Data Preprocessing

The collected data is processed to ensure consistency and quality before being used for model training. The preprocessing steps include:

• Handling missing values

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

• Removing noise and outliers
• Feature normalization and scaling
• Time-based feature encoding

These steps ensure that the dataset is suitable for machine learning algorithms and reduces the impact of noisy data on predictions.

c. Feature Engineering

Feature engineering is performed to extract meaningful information from raw meteorological data. Additional features are generated to capture atmospheric behavior more effectively.

Key engineered features include:

• Temperature–dew point difference
• Pressure-based indicators
• Temporal features

These features help the model understand complex atmospheric interactions rather than relying solely on raw inputs.

d. Model Training

The system uses two machine learning models:

• XGBoost (Primary Model)
• SGD Regressor (Baseline Model)

The dataset is divided into training and testing sets, and models are trained using supervised learning techniques.

The XGBoost model is optimized to minimize prediction error and capture nonlinear relationships among meteorological variables. The SGD model serves as a baseline for comparison.

e. Quantile-Based Prediction

To incorporate uncertainty, the system uses quantile regression with XGBoost. Instead of predicting a single value, the model generates multiple outputs:

• (Q{0.1}): Lower bound (Optimistic)
• (Q{0.5}): Median prediction
• (Q{0.9}): Upper bound (Pessimistic)

This allows the system to represent uncertainty and provide a range of possible rainfall outcomes.

f. Explainable AI Integration

To improve transparency, SHAP (SHapley Additive Explanations) is used to interpret model predictions.

SHAP provides:

• Feature contribution values
• Direction of impact (increase/decrease rainfall)
• Global and local interpretability

This ensures that predictions are not only accurate but also understandable.

g. Prediction and Output Generation

The final stage of the methodology involves generating rainfall predictions and presenting them through the system interface.

The output includes:

• Rainfall prediction (in mm)
• Probability distribution (quantile values)
• Feature contribution analysis

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

• Risk classification
These outputs are displayed through an interactive dashboard, enabling real-time decision-making.

V. RESULT AND DISCUSSION

The performance of the proposed rainfall prediction system was evaluated using real-world meteorological data obtained through external weather APIs. The system was trained on approximately 8,700+ data samples, and predictions were generated using the XGBoost model with quantile-based outputs. A baseline comparison was also performed using the SGD regressor.

a. Model Performance Evaluation

The performance of the model was evaluated using standard regression metrics:
• Mean Absolute Error (MAE)
• Root Mean Squared Error (RMSE)
The obtained results are:
• MAE: 0.003 – 0.039 mm
• RMSE: 0.05 – 0.36 mm

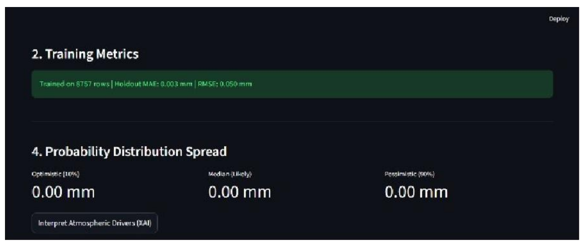

Fig. 2. Training Metrics Dashboard (MAE, RMSE, Rows Processed)

These results indicate that the model achieves high accuracy with minimal deviation between predicted and actual rainfall values. The low MAE reflects consistent performance, while RMSE confirms that larger prediction errors are effectively minimized.

b. Comparison with Baseline Model

To evaluate the effectiveness of the proposed approach, the XGBoost model was compared with the SGD (Stochastic Gradient Descent) model.
The comparison shows that:
• XGBoost produces significantly lower error values
• SGD, being a linear model, fails to capture nonlinear atmospheric relationships
• XGBoost provides more stable and reliable predictions

Fig. 3. XGBoost vs SGD Performance Comparison

This confirms that ensemble-based methods are better suited for rainfall prediction tasks involving complex meteorological data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

c. Impact of Adaptive Historical Memory

To evaluate the effect of training data size on model performance, experiments were conducted using different historical data ranges: 2 years, 3 years, and 5 years. The results demonstrate how the system adapts to varying data windows selected by the user.

Case 1: 2-Year Historical Data

The model trained using 2 years of data showed:

• MAE: 0.0026 mm
• RMSE: 0.0425 mm
• Rows Processed: 17,517

Fig. 4: Performance with 2-Year Data

These results indicate very low prediction error, as the model focuses on recent weather patterns. However, limited historical data may reduce the ability to capture long-term seasonal trends.

Case 2: 3-Year Historical Data

With 3 years of data, the model achieved:

• MAE: 0.0391 mm
• RMSE: 0.3690 mm
• Rows Processed: 26,277

Fig. 5: Performance with 3-Year Data

This configuration balances short-term and medium-term trends, providing improved generalization compared to 2-year data while maintaining reasonable accuracy.

Case 3: 5-Year Historical Data

Using 5 years of data, the model produced:

• MAE: 0.1299 mm
• RMSE: 0.6153 mm
• Rows Processed: 43,797

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, June 2026
Impact Factor: 8.2

Fig. 6: Performance with 5-Year Data

Although this configuration captures long-term seasonal behavior, the error values increase slightly due to the inclusion of older and potentially less relevant data patterns.

d. Probabilistic Rainfall Prediction

Unlike traditional systems that provide a single output, the proposed system generates quantile-based predictions to represent uncertainty.

The system produces:

• 10% Quantile (Optimistic)

• 50% Quantile (Median)

• 90% Quantile (Pessimistic)

In the observed case, all quantile outputs indicated:

• Rainfall: 0.00 mm

Fig. 7: Quantile Prediction Output (10%, 50%, 90%)

This suggests stable atmospheric conditions with a very low probability of rainfall.

e. Explainable AI (SHAP) Analysis

To interpret the model's predictions, SHAP analysis was performed. The SHAP visualization provides a breakdown of how each feature contributes to the final prediction.

Key observations include:

• HourCos, Dew Point, and Pressure contributed negatively, reducing rainfall probability

• Cloud Cover and gradient features contributed slightly toward increasing rainfall

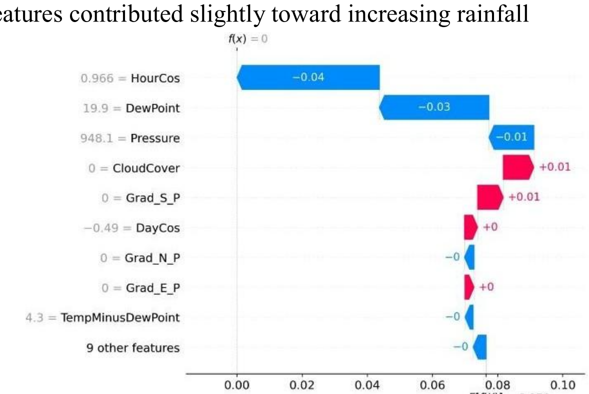

Fig. 8: SHAP Waterfall Plot (Feature Contribution)

These results align with meteorological principles, where stable pressure and low cloud cover indicate minimal chances of precipitation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 10, June 2026 Impact Factor: 8.2

f. Final Prediction Output

The system provides the final rainfall prediction along with a risk assessment. In the evaluated scenario:

• Predicted Rainfall: 0.00 mm
• Risk Level: Negligible

Fig. 9: Final Dashboard Output (Rainfall + Risk Level)

The output confirms that the model correctly identifies non-rain conditions based on the given atmospheric parameters.

VI. CONCLUSION

This study presented a rainfall prediction system that integrates machine learning, probabilistic forecasting, and Explainable Artificial Intelligence into a unified framework. The system was implemented using the XGBoost model as the primary predictor, with the SGD regressor used as a baseline for performance comparison.

The experimental results demonstrate that XGBoost achieves significantly lower error values compared to the baseline model, confirming its effectiveness in capturing complex nonlinear relationships among meteorological variables. The model achieved low Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating high prediction accuracy and stability.

A key contribution of this work is the introduction of an adaptive historical memory mechanism, which allows users to dynamically select the training data range between 1 to 5 years. The analysis showed that shorter data ranges improve prediction accuracy for recent conditions, while longer data ranges capture broader seasonal trends. This flexibility enhances the practical usability of the system and allows it to adapt to different forecasting requirements.

In addition, the use of quantile-based prediction enables the system to represent uncertainty by providing optimistic, median, and pessimistic rainfall estimates. This improves reliability compared to traditional single-value forecasts.

The integration of Explainable AI using SHAP further strengthens the system by providing clear insights into feature contributions, allowing users to understand the reasoning behind predictions. This transforms the model from a black-box system into a transparent decision-support tool.

Overall, the proposed system successfully combines accuracy, adaptability, and interpretability, making it suitable for real-world applications such as agriculture, disaster management, and environmental monitoring. The results highlight the importance of integrating machine learning with domain knowledge and explainability to build reliable and user-centric weather prediction systems.

REFERENCES

[1] R. Sharma, R. Verma, and S. K. Singh, "Explainable artificial intelligence for climate and weather forecasting: A comprehensive review," IEEE Access, vol. 13, pp. 11245–11267, 2025.

[2] A. Elbeltagi et al., "Explainable artificial intelligence for environmental prediction," Scientific Reports, 2025.

[3] D.Verma and K.Rajput, "XAI-based rainfall prediction using multivariate data," IEEE, 2025.

[4] N. Kumar and P. Mishra, "Machine learning-based rainfall prediction using ensemble and explainable AI models," IEEE Access, vol. 12, pp. 98765–98779, 2024.

[5] C. Ma et al., "Prediction of precipitation using machine learning techniques," Journal of Hydrology, 2024.

[6] Y. Wang and H. Lin, "Explainable LSTM model for weather prediction," Atmospheric Research, 2024.

[7] T. Hossain et al., "Explainable ensemble learning for rainfall prediction," IEEE Access, 2024.

[8] S. Gupta and P. Mehta, "Explainable AI for meteorological forecasting," IEEE Transactions on Artificial Intelligence, 2024.
[9] R. K. Jena, P. Pradhan, and S. Das, "Short-term rainfall prediction using machine learning techniques," IEEE Access, vol. 11, pp. 55421–55435, 2023.
[10] I. Mosavi, P. Ozturk, and K. Chau, "Flood and rainfall prediction using machine learning models: Progress and challenges," Water, vol. 15, no. 2, pp. 1–28, 2023.
[11] A. Jain and M. Dandapani, "A data-driven rainfall forecasting model using XGBoost," Neural Computing and Applications, 2023.
[12] S. Kanani et al., "An AI-enabled ensemble method for rainfall forecasting," Mathematical Biosciences and Engineering, 2023.
[13] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 66, no. 1, pp. 36–43, 2023.
[14] P. Singh and R. K. Mishra, "Rainfall forecasting using hybrid machine learning models," Journal of Hydrology, vol. 620, 2022.
[15] S. Chatterjee, A. Sarkar, and D. Biswas,"Rainfall prediction using ensemble learning techniques," Applied Soft Computing, vol. 109, 2021.
[16] IPCC,"Climate Change 2021: The Physical Science Basis," Cambridge University Press, 2021.
[17] M. Velmurugan et al., "Evaluation of explainable AI methods using LIME and SHAP," IEEE Transactions, 2020.
[18] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016.