Rainfall Prediction Using Machine Learning and Neural Networks: A Performance Evaluation

Debojit Basak[1*] and Tonni Rani Dey[2††]

[1*]Department of Mathematics, Shahjalal University of Science and Technology, Sylhet, 3100, Sylhet, Bangladesh.
[2]Department of Mathematics, Shahjalal University of Science and Technology, Sylhet, 3100, Sylhet, Bangladesh.

*Corresponding author(s). E-mail(s): debojitbasak102@gmai.com;
Contributing authors: aparnatonni@gmail.com;
†These authors contributed equally to this work.

Abstract

Precise rainfall forecasting is essential for agriculture, managing water resources, and preventing disasters. Conventional approaches to predicting rainfall frequently fall short of the accuracy and flexibility provided by contemporary machine learning methods. In this research, we utilize a dataset comprising meteorological variables like temperature, humidity, pressure, and wind speed to forecast rainfall events through a range of machine learning models and a deep learning neural network. This method seeks to establish a strong framework for tackling issues in forecasting rainfall, which is naturally intricate and affected by numerous interconnected elements.

The dataset experienced thorough preprocessing, such as standardization to maintain consistency in feature scales, addressing class imbalance through SMOTE (Synthetic Minority Oversampling Technique), and applying PCA (Principal Component Analysis) for dimensionality reduction. Exploratory Data Analysis (EDA) uncovered important patterns and correlations among features, including the link between pressure and rainfall, which guided model creation and emphasized essential predictive variables. These preprocessing measures guaranteed that the models developed on the dataset could accurately identify fundamental trends and reduce overfitting.

Six different machine learning models—Logistic Regression, Random Forest, Decision Tree, Support Vector Classifier (SVC), CatBoost, and K-Nearest Neighbors (KNN)—were developed and assessed using performance metrics like accuracy, precision, recall, and F1-score. Furthermore, a neural network featuring an improved architecture that integrates Batch Normalization and Early

Stopping was created to enhance predictive performance, maintain stability in training, and avoid overfitting. The architecture of the neural network comprised several hidden layers equipped with dropout techniques to manage intricate connections among features.
The findings indicated that Random Forest [1] and CatBoost surpassed the other models, reaching an accuracy of 87%, while the Neural Network followed with an accuracy of 85%. Logistic Regression, SVC, and KNN also demonstrated competitive performance, emphasizing their effectiveness for predicting rainfall in resource-limited settings. The Decision Tree showed the least accuracy (74%), suggesting its limitations in generalizability for this purpose. Additionally, comprehensive examinations of confusion matrices and classification reports highlighted the advantages and disadvantages of each model, offering understanding into their real-world uses.
This study highlights the success of machine learning and deep learning methods in forecasting rainfall and offers views on their real-world uses. The findings highlight the significance of preprocessing methods like SMOTE and PCA in creating models that perform well. Future research might aim at broadening the dataset to encompass various geographic and climatic contexts, investigating ensemble learning techniques to improve predictive accuracy, and integrating sophisticated deep learning frameworks such as recurrent neural networks (RNNs) or transformers for forecasting time-series data. By focusing on these aspects, the suggested approach can provide a basis for creating more precise and scalable rainfall forecasting systems.

Keywords: Machine Lerning, Deep Learning, Neural Netwrok, PCA, SMOTE

1 Introduction

1.1 Problem Statement

Forecasting rainfall is an essential element of meteorology, with important consequences for agriculture, disaster response, and water resource management. Precise forecasts can assist farmers in making educated choices regarding irrigation, planting, and harvesting. Nevertheless, conventional statistical models frequently find it challenging to accurately represent the nonlinear and intricate relationships among different meteorological elements, such as humidity, pressure, and temperature, essential for forecasting rainfall. The absence of accurate forecasts may result in negative consequences, such as agricultural failures, water deficits, and financial setbacks.

The emergence of machine learning and deep learning technologies has created new opportunities for tackling these issues. These techniques can detect intricate patterns within extensive datasets, providing a chance to enhance the precision and dependability of rainfall predictions. Even with their promise, successfully implementing these methods demands thoughtful attention to preprocessing, feature selection, and model assessment.

1.2 Importance of Rainfall Prediction in Agriculture and Water Resource Management

Precipitation is crucial in influencing farming output and the accessibility of water resources. In areas that depend largely on rain-fed farming, precise rainfall predictions are essential for reducing risks linked to erratic weather conditions. For example, prompt forecasts can help improve irrigation timing, decrease water loss, and lessen the effects of droughts or floods.

Apart from agriculture, forecasting rainfall is crucial for managing water resources. It assists in the management and functioning of reservoirs, guaranteeing sufficient water availability for household and industrial needs. Precise predictions aid in disaster readiness by offering advance alerts about severe weather occurrences, such as intense rainfall that can cause flooding. By incorporating sophisticated predictive methods, stakeholders can make better-informed choices, strengthening resilience to climate fluctuations and promoting sustainable resource management. [2]

1.3 Objectives of the Study

The main aim of this research is to create a reliable rainfall forecasting system utilizing machine learning and deep learning methods. The particular objectives consist of:

- To prepare and examine a meteorological dataset in order to uncover significant patterns and insights.
- To assess the effectiveness of different machine learning models, such as Logistic Regression, Random Forest, Decision Tree, SVC, CatBoost, and KNN, in forecasting the occurrence of rainfall.
- To develop and execute a neural network featuring an upgraded architecture for better prediction precision.
- To assess the performance of machine learning models and the neural network by utilizing metrics like accuracy, precision, recall, and F1-score.
- To offer insights into the advantages and drawbacks of each model, emphasizing their real-world uses and opportunities for future enhancement.

This research seeks to contribute to developing more dependable and scalable rainfall prediction systems by meeting these goals, thereby enhancing agricultural productivity, water resource management, and resilience to disasters. [3]

2 Literature Review

2.1 Traditional Methods for Rainfall Prediction

Historically, rainfall forecasts have depended on statistical methods and numerical weather prediction (NWP) models. Methods like regression analysis and autoregressive integrated moving average (ARIMA) have been employed to predict rainfall using historical weather records. Although these techniques are simple, they frequently do not account for the nonlinear relationships among variables such as temperature, pressure,

and humidity. For example, Kannan and Ghosh (2011) underscore the shortcomings of regression models in managing multivariate dependencies, resulting in reduced prediction accuracy in intricate meteorological systems. [4]

Another frequently utilized method is employing physical models grounded in atmospheric dynamics. Models like the Weather Research and Forecasting (WRF) model employ equations related to fluid dynamics and thermodynamics to replicate weather patterns. Nonetheless, they demand considerable computational resources and are influenced by initial conditions, as mentioned by Kalnay (2003). [5]

2.2 Machine Learning Approaches in Climate Modeling

Machine learning methods have gained traction in climate modeling due to their ability to handle large datasets and uncover hidden patterns. Techniques such as Support Vector Machines (SVM), Random Forests, and Gradient Boosting Machines have been applied successfully to predict rainfall. According to a study by Kumar et al. (2012), Random Forests outperformed traditional statistical models in predicting monsoonal rainfall in India by effectively handling high-dimensional data. [6]

Deep learning methods, such as Artificial Neural Networks (ANNs), have also been explored for climate prediction. ANNs can approximate complex nonlinear functions, making them suitable for meteorological data.[7] For example, Mishra and Desai (2006) demonstrated the efficacy of ANN-based models in predicting rainfall with higher accuracy compared to regression techniques. However, these methods require extensive training data and are prone to overfitting without proper regularization.

2.3 Advances in Deep Learning for Meteorological Predictions

Recent advancements in deep learning have further enhanced rainfall prediction capabilities. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown promise in extracting spatial and temporal features from meteorological data. CNNs are particularly effective for processing gridded weather data, while LSTMs excel in capturing sequential dependencies. [8]

Studies like Shi et al. (2015) have leveraged LSTM networks for precipitation forecasting, demonstrating their ability to outperform traditional models in time-series predictions. Additionally, hybrid models combining CNNs and LSTMs have been proposed to capture both spatial and temporal dynamics, as highlighted by Rasp et al. (2020). [3]

The application of ensemble deep learning models has also gained attention. Techniques that combine multiple neural networks to improve generalization have been explored in studies like Wang et al. (2019). These advancements highlight the growing potential of deep learning in addressing the complexities of meteorological predictions. [2]

3 Methodology

This section details the dataset description, exploratory data analysis, preprocessing steps, and the model selection and training process used to predict rainfall occurrence. [9]

3.1 Dataset Description

The dataset contains meteorological features relevant to rainfall prediction. While the exact source is unknown, the dataset's structure and content represent key atmospheric and environmental factors influencing weather conditions. The features include:

- **Day**: A unique identifier for each observation, used for indexing.
- **Pressure**: Atmospheric pressure, a critical factor in weather dynamics.
- **Max Temperature (maxtemp)**: The highest temperature recorded during the day, which affects evaporation and humidity levels.
- **Temperature**: The average temperature of the day, providing a general thermal measure.
- **Min Temperature (mintemp)**: The lowest temperature recorded, contributing to diurnal variations.
- **Dew Point**: The temperature at which air becomes saturated, linked to moisture content.
- **Humidity**: The percentage of moisture in the air, a significant determinant of rainfall.
- **Cloud Cover**: The fraction of the sky covered by clouds, directly correlated with rainfall likelihood.
- **Rainfall**: The target variable indicating the occurrence of rainfall (binary: 1 = Rain, 0 = No Rain).
- **Sunshine**: The number of sunshine hours during the day, inversely related to cloud cover.
- **Wind Direction**: The predominant direction of wind, influencing moisture transport.
- **Wind Speed**: The speed of the wind, impacting the movement of weather systems.

Handling Missing Values.

Missing values were handled using statistical imputation techniques. For numerical features, missing values were replaced with the mean of the respective column, preserving the overall data distribution. For categorical features, such as the binary target variable (*Rainfall*), data integrity was maintained during imputation. [10]

3.2 Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns, correlations, and distributions within the dataset. [11]

Feature Correlations.

A correlation heatmap revealed significant relationships among features:

- **Pressure and Rainfall**: Moderate negative correlation, consistent with the role of low-pressure systems in precipitation.
- **Humidity and Rainfall**: Strong positive correlation, indicating higher humidity increases rainfall likelihood.
- **Dew Point and Humidity**: High correlation, reflecting their shared dependence on atmospheric moisture.
- **Cloud Cover and Sunshine**: Strong negative correlation, as increased cloud cover reduces sunshine.

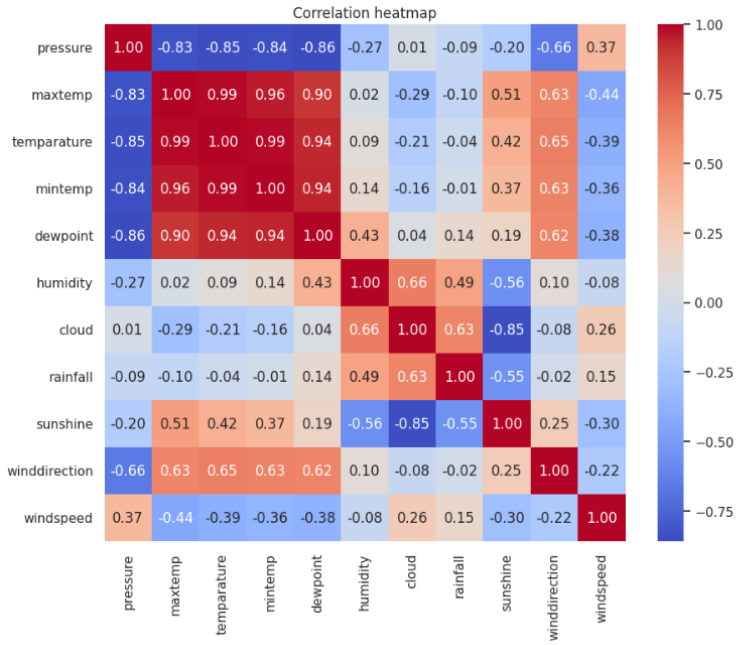

Fig. 1: Correlation matrix

Feature Distributions and Visualizations.

Distributions were analyzed using histograms.

- **Pressure and Temperature**: Exhibited normal distributions with minimal outliers.
- **Humidity**: Left-skewed distribution, indicating generally high humidity levels.
- **Wind Speed**: Showed significant variability, with some outliers representing extreme conditions.

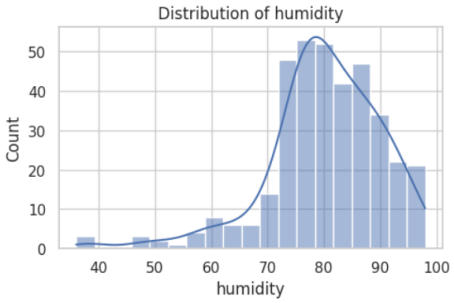

Fig. 2: Humidity.

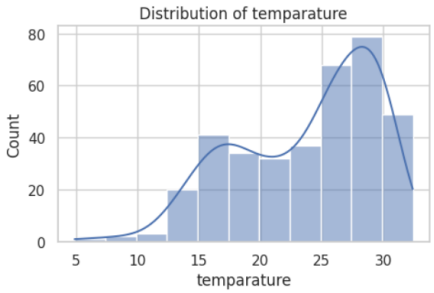

Fig. 3: Temperature.

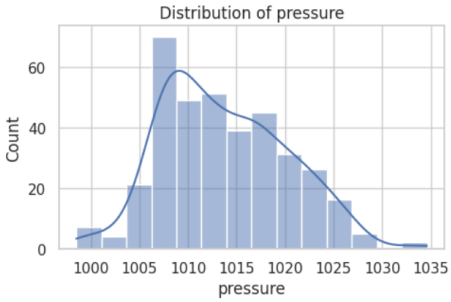

Fig. 4: Pressure.

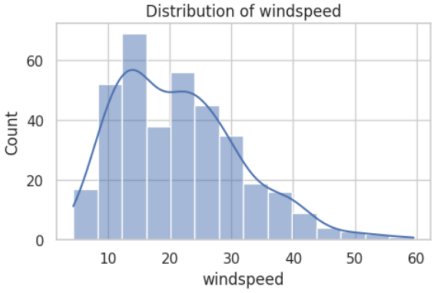

Fig. 5: Wind Speed.

Class Imbalance.

A bar chart of the target variable revealed an imbalance, with non-rainy days (class 0) outnumbering rainy days (class 1). This imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE) during preprocessing. [12]

3.3 Preprocessing Steps

Three key preprocessing steps were applied to prepare the dataset for model training.

Standardization of Features.

Features were standardized to ensure a mean of zero and unit variance:

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where x is the feature value, μ is the mean, and σ is the standard deviation. Standardization was critical for models sensitive to feature scaling, such as Logistic Regression, SVC, and KNN.

Application of SMOTE.

SMOTE balanced the dataset by generating synthetic samples for the minority class. This method interpolates between existing minority samples to create new ones, ensuring class parity without duplication.

Dimensionality Reduction Using PCA.

Principal Component Analysis (PCA) was applied to reduce dimensionality. By retaining components that captured at least 95% of the variance, PCA reduced computational complexity while preserving essential information. [13]

3.4 Model Selection and Training

This study implemented six machine learning models and a neural network to predict rainfall occurrence.

3.4.1 Overview of Machine Learning Models

The models included:

- **Logistic Regression**: Estimates the probability of rainfall using the logistic function:
$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}, \tag{2}$$
where X represents features, and β are coefficients.
- **Random Forest**: Constructs multiple decision trees and aggregates their predictions: [14]
$$\hat{y} = \text{mode}(\{T_1(X), T_2(X), \ldots, T_k(X)\}). \tag{3}$$
- **SVC**: Finds the hyperplane that maximizes the margin between classes:
$$w^T X + b = 0. \tag{4}$$
- **Decision Tree**: Splits data based on feature thresholds to minimize Gini impurity:
$$G = 1 - \sum_{i=1}^{C} p_i^2. \tag{5}$$
- **CatBoost**: Uses gradient boosting with ordered boosting to minimize log loss:
$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]. \tag{6}$$
- **KNN**: Predicts the majority class among k-nearest neighbors using Euclidean distance:
$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}. \tag{7}$$

3.4.2 Hyperparameter Tuning and GridSearchCV

GridSearchCV was used to optimize hyperparameters for each model. The F1-score was used as the evaluation metric: [15]

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

3.4.3 Neural Network Architecture

The neural network architecture included:

- **Hidden Layers**: Each layer computes:

$$a^{(l)} = \sigma(W^{(l)} a^{(l-1)} + b^{(l)}), \tag{9}$$

  where $W^{(l)}$ and $b^{(l)}$ are weights and biases, and σ is the activation function.
- **Batch Normalization**: Normalizes activations:

$$\hat{z}_i = \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}}. \tag{10}$$

- **Dropout**: Regularizes the model:

$$y_i = z_i \cdot r_i, \quad r_i \sim \text{Bernoulli}(1 - p). \tag{11}$$

- **Loss Function**: Binary cross-entropy:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{12}$$

Training optimizations included Early Stopping, batch size of 32, and up to 100 epochs.

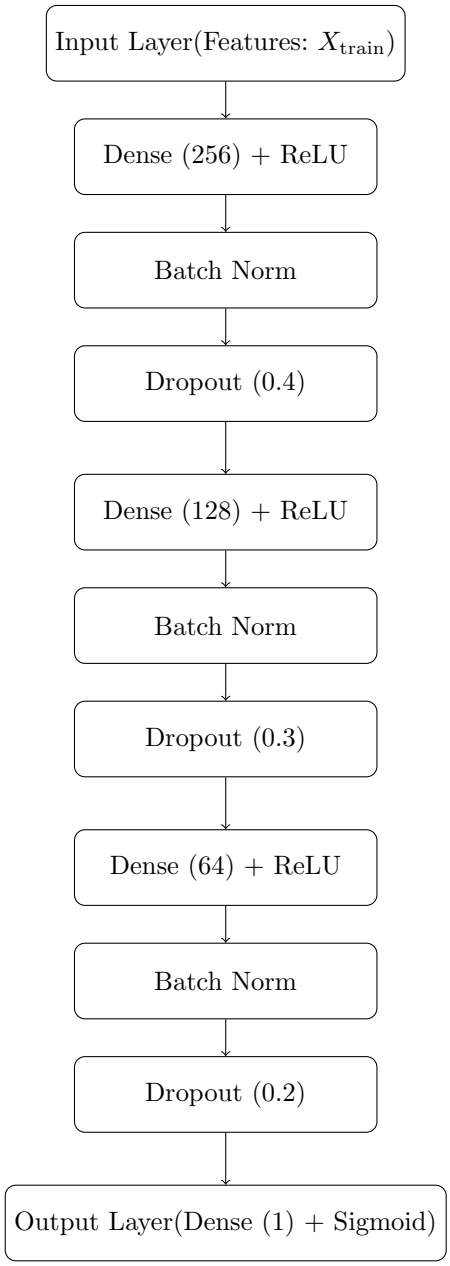

Fig. 6: Neural network architecture.

4 Results and Analysis

This section provides a comprehensive evaluation of model performance using key metrics, a comparative analysis of their effectiveness, and a detailed discussion of their strengths and weaknesses.

4.1 Performance Metrics for Each Model

The performance metrics for each model, including accuracy, precision, recall, and F1-score, are summarized in Table 1. These metrics provide an overview of the classification capabilities of each model.[16] [17]

Table 1: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.79	0.80	0.81	0.79
Random Forest	0.87	0.87	0.88	0.87
Decision Tree	0.74	0.75	0.76	0.74
SVC	0.76	0.78	0.78	0.76
CatBoost	0.87	0.87	0.87	0.87
KNN	0.86	0.85	0.86	0.86
Neural Network	0.85	0.85	0.86	0.85

4.2 Comparative Analysis

To visually compare the model performance, Figure **??** illustrates a bar chart of accuracy and F1-score across all models. Additionally, confusion matrices for the two best-performing models, CatBoost and Random Forest, are displayed in Figure **??**. [14, 18] [19]

The results from Table 1 and Figures **??** and **??** highlight the differences in model performance. Random Forest and CatBoost demonstrated the highest accuracy (0.87), with strong precision and recall, making them suitable for real-world applications requiring high reliability.

4.3 Discussion of Model Performance

The performance metrics reveal the following insights into the models:

- **Random Forest and CatBoost**: Both models achieved the highest accuracy (0.87). Random Forest leverages ensemble learning to reduce overfitting and capture nonlinear relationships. CatBoost's ordered boosting technique mitigates overfitting and enhances performance on categorical data.
- **Logistic Regression, KNN, and Neural Network**: Logistic Regression performed well, achieving a balance between precision (0.80) and recall (0.81). KNN showed comparable performance (accuracy: 0.86), benefiting from its non-parametric nature. The Neural Network effectively captured complex patterns, achieving a high F1-score (0.85) but at a higher computational cost.

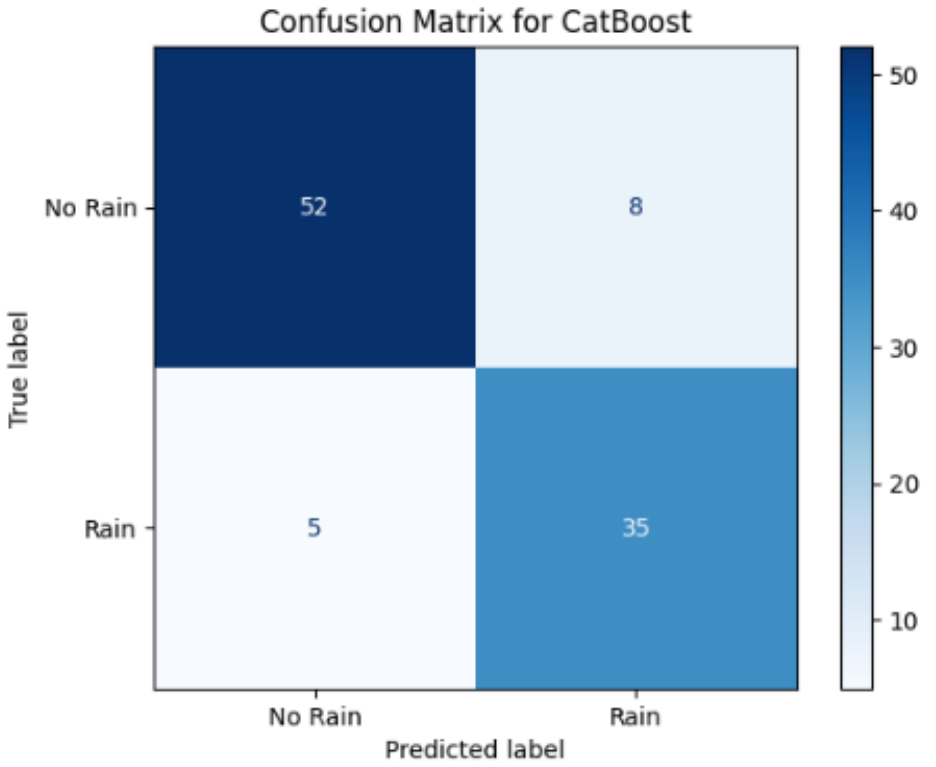

Fig. 7: Confusion Matrix of CatBoost

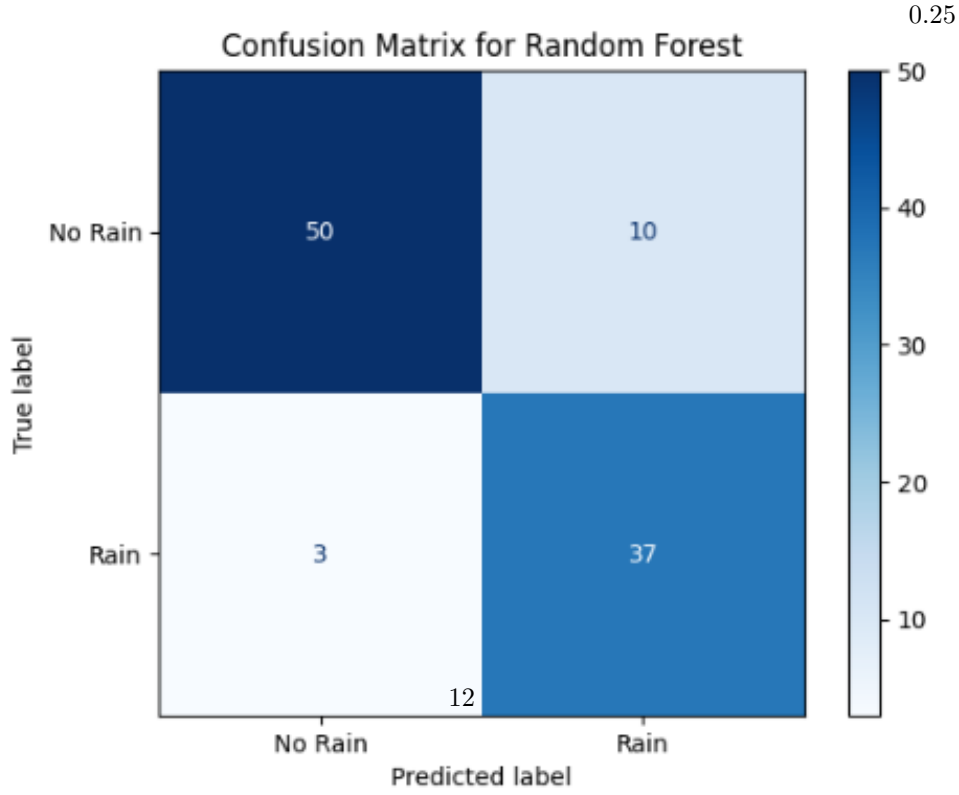

Fig. 8: Confusion Matrix of Random Forest

- **SVC and Decision Tree**: SVC achieved moderate accuracy (0.76), constrained by the dataset's complexity. Decision Tree exhibited the lowest accuracy (0.74), highlighting its sensitivity to noise and tendency to overfit.

Overall, the results demonstrate the effectiveness of advanced machine learning and deep learning methods in predicting rainfall. Random Forest and CatBoost are well-suited for applications requiring high accuracy and interpretability, while the Neural Network is advantageous for capturing complex nonlinear relationships. [12, 20]

5 Discussion

This section discusses the key findings from the study, potential practical applications, and the limitations of the proposed models and methodology.

5.1 Key Findings

The evaluation of the models revealed several important findings:

- **Random Forest and CatBoost**: These models emerged as the best-performing classifiers, each achieving an accuracy of 87%. Their ability to handle complex feature interactions and class imbalance contributed to their superior performance. Random Forest benefited from ensemble learning, while CatBoost's ordered boosting algorithm minimized overfitting. [15]
- **Neural Network**: The neural network achieved an accuracy of 85%, highlighting its ability to capture nonlinear patterns in the dataset. Optimizations such as Batch Normalization and Early Stopping improved stability during training and prevented overfitting, making it a robust solution for complex meteorological data. [19]
- **Logistic Regression, SVC, and Decision Tree**: These models exhibited limitations in handling imbalanced datasets, even after applying SMOTE. Logistic Regression performed well for a simpler linear relationship, while Decision Tree and SVC were less effective in capturing the complexities of the dataset.

These findings demonstrate the potential of advanced machine learning and deep learning models to improve rainfall prediction accuracy and reliability.

5.2 Practical Applications

The results of this study highlight several practical applications of the proposed rainfall prediction system:

- **Agricultural Planning and Flood Prevention**: Accurate rainfall forecasts can assist farmers in optimizing irrigation schedules, selecting appropriate planting times, and preparing for droughts or floods. In flood-prone regions, timely predictions can aid in disaster preparedness and mitigation efforts, reducing loss of life and property.
- **Integration into Climate Monitoring Systems**: The models can be integrated into real-time climate monitoring systems to provide actionable insights for government agencies, environmental organizations, and policymakers. This integration can enhance water resource management and ensure sustainable agricultural practices.

- **Improving Disaster Resilience**: Advanced rainfall prediction systems can support communities in building resilience against extreme weather events by providing reliable and timely information, enabling proactive decision-making.

These applications underline the practical value of deploying machine learning and deep learning models for addressing climate-related challenges.

5.3 Limitations

While the proposed methodology demonstrates promising results, it is essential to acknowledge the limitations of this study:

- **Dataset Size and Regional Specificity**: The dataset used in this study was relatively small and region-specific, which may limit the generalizability of the models to other geographic regions or climates. Expanding the dataset to include diverse meteorological conditions would enhance the robustness of the models.
- **Computational Cost**: Training advanced models such as CatBoost and neural networks requires significant computational resources. While these models achieve high accuracy, their deployment in resource-constrained environments may be challenging.
- **Class Imbalance**: Although SMOTE was used to address class imbalance, further exploration of advanced resampling techniques or cost-sensitive learning methods could improve model performance, particularly for minority classes.

Addressing these limitations in future studies can further improve the accuracy, scalability, and applicability of the proposed rainfall prediction system.

6 Conclusion and Future Work

This section provides a summary of the key findings from the study and outlines recommendations for future research to further enhance rainfall prediction models.

6.1 Summary of Findings

This study conducted a comprehensive comparative analysis of various machine learning and deep learning models for rainfall prediction. The findings can be summarized as follows:

- **Performance of Machine Learning Models**: Random Forest and CatBoost were the best-performing models, each achieving an accuracy of 87%. These models effectively captured the nonlinear relationships in the dataset and handled class imbalance with high precision and recall. The neural network, with an accuracy of 85%, demonstrated its capability to model complex patterns using advanced optimizations such as Batch Normalization and Early Stopping.
- **Importance of Preprocessing**: The preprocessing steps played a crucial role in the success of the models. The use of SMOTE addressed the class imbalance problem by generating synthetic samples for the minority class, ensuring more balanced training data. PCA (Principal Component Analysis) helped reduce the dimensionality of

the dataset, preserving 95% of the variance while minimizing computational complexity. These preprocessing techniques improved model performance and reduced the risk of overfitting.
- **Comparative Analysis**: Logistic Regression, SVC, and KNN demonstrated moderate performance, while Decision Tree exhibited the lowest accuracy. The results underscored the importance of model selection, especially for datasets with nonlinear dependencies and imbalanced target variables.

The study highlights the potential of machine learning and deep learning models in rainfall prediction, emphasizing the significance of careful preprocessing and model optimization.

6.2 Recommendations for Future Work

While this study provides valuable insights, several avenues for future research can be explored to enhance the accuracy, scalability, and applicability of rainfall prediction models:

- **Expanding the Dataset**: Future studies should focus on incorporating larger and more diverse datasets from multiple regions and climatic conditions. This would improve the generalizability of the models and their ability to handle variations in meteorological patterns.
- **Exploring Ensemble Learning Approaches**: Ensemble techniques such as stacking, blending, or boosting can be investigated to combine the strengths of multiple models. These approaches have the potential to enhance predictive accuracy and robustness.
- **Investigating Transformer-Based Architectures**: Transformers have shown remarkable success in time-series forecasting and sequence modeling. Adapting transformer-based architectures for rainfall prediction could leverage their ability to capture long-term dependencies in meteorological data.
- **Integration with Real-Time Systems**: Future work could focus on integrating these models with real-time data collection and monitoring systems, enabling dynamic and adaptive predictions.
- **Cost-Effective Models for Resource-Constrained Environments**: Developing lightweight and efficient models that maintain high accuracy while reducing computational overhead would make these systems more accessible for broader adoption.

By addressing these areas, future studies can further refine rainfall prediction systems, ensuring their utility in a wider range of practical applications and climatic conditions.

References

[1] Elbeltagi, A., Pande, C.B., Kumar, M., Tolche, A.D., Singh, S.K., Kumar, A., Vishwakarma, D.K.: Prediction of meteorological drought and standardized precipitation index based on the random forest (rf), random tree (rt), and gaussian

process regression (gpr) models. Environmental Science and Pollution Research **30**(15), 43183–43202 (2023)

[2] Ren, X., Li, X., Ren, K., Song, J., Xu, Z., Deng, K., Wang, X.: Deep learning-based weather prediction: a survey. Big Data Research **23**, 100178 (2021)

[3] Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., Woo, W.-c.: Deep learning for precipitation nowcasting: A benchmark and a new model. Advances in neural information processing systems **30** (2017)

[4] Kannan, S., Ghosh, S.: Prediction of daily rainfall state in a river basin using statistical downscaling from gcm output. Stochastic Environmental Research and Risk Assessment **25**, 457–474 (2011)

[5] Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability vol. 341. Cambridge University Press, ??? (2003)

[6] Kumar, A.S., Roshan, S.A., Dutta, A., Ray, S., Masadeh, S.R., Lakshmi, G.P., Michalopoulos, D., Nyayapati, R., Musirin, I.B., Kaur, G.: Rainfall prediction using machine learning. In: Advancements in Climate and Smart Environment Technology, pp. 100–113. IGI Global, ??? (2024)

[7] Singh, P., Borah, B.: Indian summer monsoon rainfall prediction using artificial neural network. Stochastic environmental research and risk assessment **27**, 1585–1599 (2013)

[8] Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N.: Weatherbench: a benchmark data set for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems **12**(11), 2020–002203 (2020)

[9] Ahrens, C.D.: Meteorology Today: an Introduction to Weather, Climate, and the Environment. Cengage Learning Canada Inc, ??? (2015)

[10] Wilks, D.S.: Statistical Methods in the Atmospheric Sciences. Academic press, ??? (2011)

[11] McKinney, W., *et al.*: Data structures for statistical computing in python. In: SciPy, vol. 445, pp. 51–56 (2010)

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[13] Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences **374**(2065), 20150202 (2016)

[14] Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)

[15] Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade: Second Edition, pp. 437–478. Springer, ??? (2012)

[16] Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process **5**(2), 1 (2015)

[17] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information processing & management **45**(4), 427–437 (2009)

[18] Probst, P., Wright, M.N., Boulesteix, A.-L.: Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery **9**(3), 1301 (2019)

[19] Dorogush, A.V., Ershov, V., Gulin, A.: Catboost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 (2018)

[20] Goodfellow, I.: Deep learning. MIT press (2016)