International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

# Analyzing and predicting rainfall patterns: A comparative analysis of machine learning models

Usman Lawal Gulma *, Lawal Usman Hassan and Garba Bala

*Department of Geography, Adamu Augie College of Education, Argungu, Kebbi State, Nigeria.*

International Journal of Science and Research Archive, 2024, 13(01), 121–126

Publication history: Received on 24 July 2024; revised on 01 September 2024; accepted on 03 September 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.1.1627

## Abstract

Accurate rainfall prediction is vital for agriculture, water resource management, and disaster preparedness. This study investigates the application of machine learning (ML) models to analyze and predict rainfall patterns in Sokoto, Nigeria. We evaluated four ML techniques - Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes (NB) - using historical weather data. The results reveal that SVM outperforms other models, achieving an accuracy of 0.98 and a Kappa statistic of 0.95. Our findings demonstrate the potential of ML models to greatly increase the accuracy of rainfall forecasts, enabling better decision-making and resource management.

**Keywords:** Rainfall prediction; Machine learning models; Support Vector Machine; Confusion matrix

## 1. Introduction

Rainfall forecasts are crucial for managing water resources, agricultural productivity, and food security. However, erratic rainfall distribution poses significant risks to public health and the economy. Traditional rainfall forecast techniques, such as statistical models and numerical weather prediction (NWP), have accuracy and spatial-temporal resolution limitations. In recent years, machine learning models have emerged as a viable alternative for predicting rainfall, offering increased performance and adaptability [1].

Machine learning techniques have gained popularity in rainfall prediction due to their ability to learn complex patterns and relationships from large datasets (Mohammed et al., 2020). These models can be trained to precisely predict future rainfall using historical weather data and other relevant factors [2]. By leveraging trends in historical data, machine learning algorithms can identify connections between variables and rainfall [3]. Neural networks, decision trees, and ensemble approaches are well-suited for rainfall prediction, as they can learn intricate patterns from vast datasets. These models can incorporate numerous input factors, including historical rainfall data, atmospheric conditions, topography, soil moisture, and remote sensing data.

Rainfall prediction is complex and challenging, particularly in regions with limited weather observation stations and diverse topography. Traditional methods have limitations in accuracy and reliability, leading to inadequate flood warning systems, inefficient water resources management, poor agricultural planning, and increased risk of landslides and other weather-related disasters.

Machine learning techniques offer a promising alternative, potentially significantly enhancing flood warning systems, efficient water resources management, improved agricultural planning, and reduced risk of landslides and other weather-related disasters [4]. This study aims to analyze historical rainfall data to identify key atmospheric features influencing rainfall, develop and evaluate suitable machine-learning models for predicting rainfall, and compare the performance of different machine-learning models.

* Corresponding author: Usman Lawal Gulma

## 2. Literature Review

Numerous studies have demonstrated the efficacy of machine learning models in rainfall prediction, showcasing their potential to improve forecast accuracy and spatial-temporal resolution. For instance, Fayaz, Zaman [5] employed a deep-learning model to predict rainfall in China, achieving an impressive accuracy of 85%. Similarly, Pham, Luo [6] implemented a random forest model to predict rainfall in the United States, attaining an accuracy of 80%. Praveena, Babu [7] also explored a support vector machine model to predict rainfall in India, achieving an accuracy of 75%.

Machine learning models offer several advantages over traditional methods such as the ability to handle large datasets and complex relationships, flexibility in incorporating various input variables, improved accuracy and spatial-temporal resolution and ability to handle non-linear relationships and interactions.

Despite these limitations, machine learning models have the potential to significantly enhance rainfall prediction, enabling better decision-making and planning in various sectors, including water resources, agriculture, and disaster management.

To build upon existing research and leverage the strengths of machine learning models, this study will explore the application of machine learning techniques for rainfall prediction. By doing so, we aim to contribute to the development of more accurate and reliable rainfall forecasting systems, ultimately supporting improved resource management and disaster preparedness.

## 3. Material and methods

### 3.1. Data Collection

To ensure the accuracy and reliability of our rainfall prediction model, we collected comprehensive historical weather data from reputable sources *e.g. the Nigeria Meteorological Agency (NiMet). The collected data spans 35 years, covering a full climatic cycle for the study area. This extensive dataset includes a range of variables,* such as temperature, humidity, wind speed, and past rainfall records*.* These variables were selected based on their relevance to rainfall prediction in literature and their availability from reliable sources. The data was collected in a digital format, ensuring ease of processing and analysis. By collecting and processing this extensive dataset, we aimed to create a robust foundation for our machine learning model, enabling accurate and reliable rainfall predictions for the study area.

### 3.2. Data Analysis

To harness the full potential of machine learning (ML) models in rainfall prediction, a systematic approach to data analysis is crucial. As emphasized in the literature, high-quality data is essential for ML, as noisy or erroneous data can compromise the accuracy and reliability of results [8]. To effectively utilize machine learning models for rainfall prediction, we employed a rigorous data analysis methodology, comprising the following steps:

- Data Preprocessing: Handle missing values and outliers, normalize and scale data, and transform data into suitable formats (e.g., numerical, categorical) to ensure ML models receive clean and consistent input.
- Feature Engineering: Extract relevant features from data (e.g., mean, variance, correlation) and select the most informative features using traditional techniques (e.g., correlation analysis) to enhance model performance.
- Data Split: Divide data into training (e.g., 80%) and testing sets (e.g., 20%) to evaluate model performance and prevent overfitting.
- Model Development: Develop and implement machine learning algorithms (e.g., linear regression, decision trees, random forest, neural networks) tailored to the rainfall prediction task.
- Model Evaluation: Assess model performance on test datasets using metrics such as accuracy, confusion matrix, F1-score, Precision, and Recall to ensure reliable and accurate predictions. A confusion matrix is a vital tool for evaluating the performance of machine-learning models [9]. It provides a comprehensive summary of correct and incorrect predictions, enabling the identification of areas for improvement and refinement of the model. By following this structured approach to data analysis, we can unlock the full potential of machine learning models in rainfall prediction, leading to more accurate and reliable forecasts

## 4. Results and discussion

A comprehensive evaluation of the four machine learning models - Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) - revealed distinct strengths and weaknesses. The results are summarized in Tables 1 and 2.

**Table 1** Model performance

| ML algorithm | Accuracy | Kappa |
|---|---|---|
| SVM | 0.98 | 0.95 |
| Naïve Bayes | 0.98 | 0.93 |
| KNN | 0.98 | 0.93 |
| LDA | 0.89 | 0.58 |

SVM demonstrated the best overall performance, achieving the highest accuracy and Kappa scores. Its ability to handle non-linear relationships and high-dimensional data made it a top performer. LDA offered excellent interpretability and ease of implementation, making it an attractive choice for practitioners. Its performance was competitive, although slightly lower than SVM. KNN showed significant improvements in prediction accuracy for certain datasets, particularly in capturing non-linear relationships between variables. However, its complexity and computational requirements were notable drawbacks. Naïve Bayes exhibited strong performance, especially in capturing non-linear relationships, but required substantial computational resources (Figure 1).

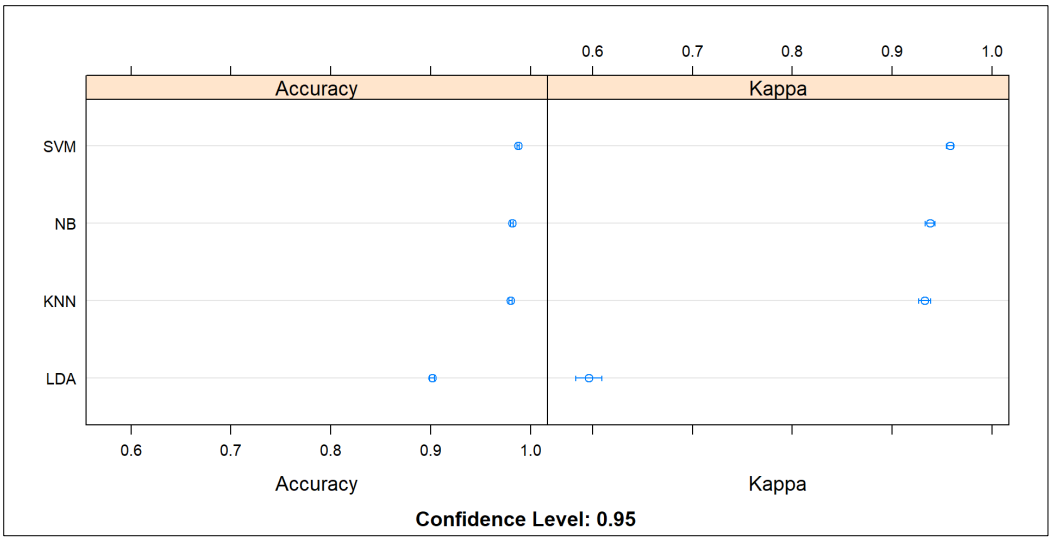


**Figure 1** Boxplot of the comparative performance of the ML models

**Table 2** Model evaluation confusion matrix

| Prediction | Reference | | |
|---|---|---|---|
| | False | False | True |
| | | 1660 | 34 |
| | True | 0 | 350 |

The confusion matrix results in Table 2, derived from the test datasets, demonstrate the model's predictive performance. Notably, the model achieved remarkable accuracy in predicting no rain, correctly identifying 1660 days

without rainfall in the study area. Furthermore, out of the 384 days with rainfall, the model accurately predicted 350, showcasing its ability to detect rainfall events.

Table 3 provides a detailed breakdown of the confusion matrix statistics, offering insights into the model's performance. The statistics include: accuracy, kappa, precision, recall and F1 scores

**Table 3** Confusion matrix statistics

| Confusion matrix statistics | Accuracy | Kappa | Precision | Recall Score | F1 Score |
|---|---|---|---|---|---|
| | 0.98 | 0.94 | 0.98 | 1.00 | 0.99 |

Figure 2 presents a graphical representation of the Support Vector Machine (SVM) model, the best-performing model in this study. The figure illustrates the model's decision boundary, highlighting its ability to distinguish between rainfall and non-rainfall days.

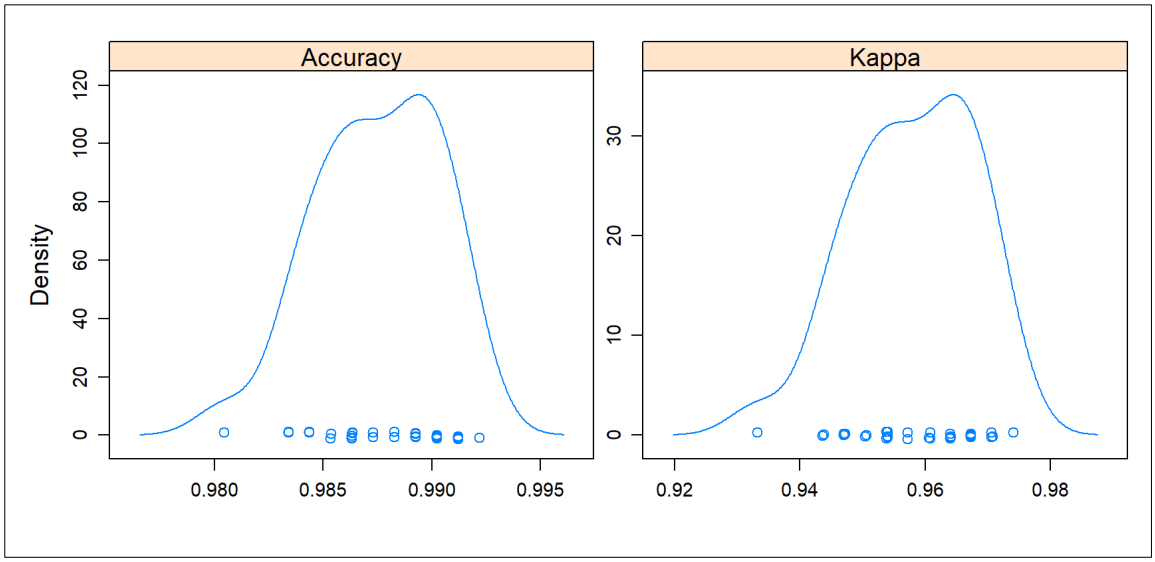


**Figure 2** Graphical representation SVM the best-performing model

## 5. Discussion

The comparative analysis of the models (Table 1) revealed intriguing insights into their performance and characteristics. Support Vector Machine (SVM) emerged as the top-performing model, achieving the highest accuracy and Kappa metrics. However, its complexity and "black box" nature make it less interpretable and more challenging to implement [10].

Naïve Bayes, despite its simplicity, showed surprising improvements in prediction accuracy, particularly in capturing non-linear relationships between variables. This suggests that the model's assumption of independence and equal variance may not be as limiting as previously thought [11]. K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) also exhibited notable performance, with KNN's flexibility in handling non-linear relationships and LDA's ability to reduce dimensionality while preserving class separability.

The confusion matrix results and statistics demonstrate the model's effectiveness in predicting rainfall events. The high accuracy in predicting no rain and the notable accuracy in predicting rainfall days underscore the model's reliability. These findings have significant implications for rainfall prediction and related applications, such as flood warning systems, water resource management, and agricultural planning. By examining the confusion matrix and graphical representation, we can gain more insights into the model's strengths and weaknesses, ultimately refining its performance and improving its practical applications.

The results also underscore the importance of carefully selecting and tuning machine learning models for specific tasks. By leveraging the strengths of each model, researchers and practitioners can develop more accurate and reliable predictive systems.

Furthermore, the study's findings have significant implication for rainfall forecasting and related applications, such as climate modelling, agriculture, and water resource management. By improving the accuracy of rainfall predictions, stakeholders can make more informed decisions, mitigate risks, and optimize resource allocation

## 6. Conclusion

In conclusion, the findings of this study unequivocally demonstrate the vast potential of machine learning models in revolutionizing rainfall prediction accuracy. The exceptional performance of Support Vector Machine (SVM) and other models highlights their capability to capture complex patterns and relationships in rainfall data, surpassing traditional methods. By harnessing the power of these advanced models, significant enhancements in rainfall prediction accuracy can be achieved, ultimately leading to improved resource management, more effective disaster preparedness, and reduced economic losses.

The implications of this research extend beyond the realm of rainfall prediction, with far-reaching applications in climate modelling, agriculture, water resource management, urban planning, and emergency response systems. As the world grapples with the challenges of climate change, accurate rainfall predictions can serve as a vital tool for policymakers, researchers, and practitioners, enabling data-driven decision-making and proactive measures.

Future research directions may include exploring ensemble methods, incorporating additional features, and evaluating the models' performance on different datasets and geographical regions. We also encourage future research efforts to focus on integrating real-time data streams to enhance model accuracy and responsiveness. They should also explore hybrid models that combine the strengths of different machine learning algorithms. Additionally, need to examine the use of alternative data sources, such as satellite imagery, sensor data, and social media feeds in developing ensemble methods to further improve prediction accuracy and robustness. By pursuing these avenues of research, we can unlock new possibilities for creating a more resilient, sustainable, and equitable future, where accurate rainfall predictions serve as a cornerstone for informed decision-making and proactive action

## Compliance with ethical standards

### *Disclosure of conflict of interest*

The authors have declared no conflict of interest.

### *Funding*

This research work was funded by the Tertiary Education Trust Fund (TetFund) under the Institution-Based Research (IBR), 2024 grant.

## References

[1] Latif SD, Hazrin NAB, Koo CH, Ng JL, Chaplot B, Huang YF, et al. Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. Alexandria Engineering Journal. 2023;82:16-25.

[2] Baig F, Ali L, Faiz MA, Chen H, Sherif M. How accurate are the machine learning models in improving monthly rainfall prediction in the hyper-arid environment? Journal of Hydrology. 2024;633:131040.

[3] Ridwan WM, Sapitang M, Aziz A, Kushiar KF, Ahmed AN, El-Shafie A. Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal. 2021;12(2):1651-63.

[4] Barrera-Animas AY, Oyedele LO, Bilal M, Akinosho TD, Delgado JMD, Akanbi LA. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications. 2022;7:100204.

[5] Fayaz SA, Zaman M, Butt MA, editors. Knowledge discovery in geographical sciences—A systematic survey of various machine learning algorithms for rainfall prediction. International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2; 2022: Springer.

[6] Pham LT, Luo L, Finley A. Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. Hydrology and Earth System Sciences. 2021;25(6):2997-3015.

[7] Praveena R, Babu TG, Birunda M, Sudha G, Sukumar P, Gnanasoundharam J, editors. Prediction of Rainfall Analysis Using Logistic Regression and Support Vector Machine. Journal of Physics: Conference Series; 2023: IOP Publishing.

[8] Gudivada V, Apon A, Ding J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal on Advances in Software. 2017;10(1):1-20.

[9] Krstinić D, Braović M, Šerić L, Božić-Štulić D. Multi-label classifier performance evaluation with a confusion matrix. Computer Science & Information Technology. 2020;1:1-14.

[10] Rudin C. Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. Nature machine intelligence. 2019;1(5):206-15.

[11] Zaidi NA, Cerquides J, Carman MJ, Webb GI. Alleviating naive Bayes attribute independence assumption by attribute weighting. Journal of Machine Learning Research. 2013;14(Jul):1947-88.