Optimizing Rainfall Prediction in Settat, Morocco, Through Machine Learning

Oussama Zemnazi[1], Sanaa El Filali[2], Sara Ouahabi[3], Abderrahim Mouhtadi[4]
Faculty of Sciences Ben M'Sick-Laboratory of Artificial Intelligence and Systems,
University Hassan II - Casablanca, Casablanca, Morocco[1, 2, 3]
General Directorate of Meteorology, Morocco[4]

***Abstract*—Rainfall prediction is still a difficult challenge because rainfall is nonlinear, intermittent, and highly variable, especially in semi-arid climates. Accurate rainfall prediction is crucial for water resource management, agricultural planning, climate-driven decision-making, and more. This study proposes a comparative framework based on machine learning and ensemble learning techniques to predict daily rainfall in Settat, Morocco, as a representative semi-arid region. Five predictive models were trained and evaluated based on meteorological station observations: Random Forest, XGBoost, LightGBM, CatBoost, and a Multilayer Perceptron (MLP). The models' performance was evaluated based on mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), and the coefficient of determination (R-squared). The results demonstrate that the performance and stability of gradient boosting algorithms are superior to all other evaluated models. Specifically, LightGBM produced the fewest erroneous values and explained rainfall variability best. These results underscore the success of boosting-based ensemble techniques in modeling inconsistent precipitation patterns and provide a comparative framework for machine-learning-based rainfall forecasting in semi-arid environments.**

***Keywords*—Rainfall forecasting; machine learning; gradient boosting; LightGBM; semi-arid climate; ensemble learning**

I. Introduction

The forecast of rainfall is vital in water resource management, agricultural productivity, flood risk intervention, and climate resilience strategies, especially in semi-arid environments where precipitation is rare, erratic, and variable. In these regions, rainfall tends to have long dry spells and short, intense storms, leading to uncertain water resource management and a high susceptibility to droughts and flash floods [1], [2].

Traditionally, rainfall prediction has relied on Numerical Weather Prediction (NWP), which is based on the numerical integration of atmospheric physical equations. Although these models provide accurate predictions at synoptic and regional scales, they perform poorly at the local and station levels, especially in data-limited semi-arid areas [3]. Furthermore, NWP models have high computational costs and are driven by initial conditions; they generally do not capture orographic rainfall processes that trigger convection and dominate precipitation variability in dry climates [4].

Recently, machine learning (ML) methods have been established as competitive alternatives for rainfall prediction due to their ability to learn complex nonlinear relationships directly from historical meteorological data. Ensemble learning methods, such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), LightGBM, and CatBoost, are widely used for tabular climate data because of their robustness, ability to handle heterogeneous predictors, and reduced sensitivity to overfitting [5]–[8]. Neural network models, especially the multilayer perceptron (MLP), have been extensively used for rainfall forecasting due to their ability to model complex nonlinear relationships. These models perform effectively when the feature space is well engineered and when large amounts of training data are available, which can undermine their performance in the context of station-based semi-arid areas [9].

Even with the growing use of ML techniques for rainfall prediction, several limitations remain. First, most previous work focuses on temperate or data-rich regions. However, semi-arid settings are under-investigated despite rainfall intermittency, skewed distributions, and heavy-tailed distributions (extreme events) [2], [10]. Second, existing works commonly employ heterogeneous preprocessing techniques, feature representations, and evaluation metrics, which make comparisons among models overly challenging. Consequently, there is still no clear consensus on whether ensemble boosting algorithms consistently outperform neural network models for daily rainfall prediction when evaluated under identical experimental conditions. In addition, relatively few studies provide climate-aware interpretations that explicitly link model performance to the physical characteristics of rainfall processes in semi-arid regions [11].

To bridge the gap, a comprehensive and systematic comparison of five machine learning models for daily rainfall prediction is conducted in Settat, Morocco, a representative semi-arid area. The compared models are four ensemble models: LightGBM, XGBoost, CatBoost, and the baseline neural Multilayer Perceptron (MLP). All models are trained and evaluated with the same 11 years of daily meteorological observation data using the same processing, feature engineering, and evaluation metrics. This unified framework enables a fair and transparent assessment of the relative strengths and limitations of each modeling approach.

The main contributions of this study are summarized as follows:

- A universal benchmarking system to evaluate the ensemble and neural network methods in a daily rainfall forecasting task with identical experimental settings.

- The analysis is based on long-term station observations of a semi-arid climate with high rainfall intermittency.
- A climate-informed assessment of model performance, focusing on the applicability of gradient boosting techniques to tabular meteorological data in data-sparse settings.

The rest of the study is organized as follows: Section II presents related work on machine learning-based rainfall prediction. Section III provides the area of study, the dataset, and the preprocessing. The modeling techniques and evaluation metrics are described in Section IV. Section V presents the experimental results, and we discuss our findings and their limitations in Section VI. Section VII concludes the study and outlines future research directions.

II. Related Work

A. Rainfall Forecasting Using Machine Learning

The prediction of rainfall has long been acknowledged as a difficult problem, primarily because precipitation processes are known to be both nonlinear and nonstationary with high intermittency. Traditional statistical methods, including linear regression and autoregressive models, have been widely used for rainfall prediction. However, their predictive performance is typically weak under strong climate fluctuations and extreme rainfall events. These constraints have led to the growing use of machine learning (ML) approaches as data-driven solutions for rainfall prediction.

Machine learning techniques have demonstrated strong capabilities for capturing complex nonlinear associations between rainfall and atmospheric predictors. Various ML techniques, such as decision trees, support vector machines, artificial neural networks, and ensemble learning methods, have been used to predict rainfall using station-based gauges, satellite-derived products, or reanalysis data. Several researchers have found that machine learning (ML) methods can outperform traditional statistical approaches for daily rainfall forecasting, especially when multiple meteorological variables are used as inputs [2], [12]. Reanalysis data, including ERA5, have also been increasingly adopted in ML-based rainfall prediction to provide spatially and temporally consistent atmospheric information [10]. Nevertheless, the performance of ML techniques on rainfall prediction is still sensitive to regional climate conditions, data availability, and experimental settings. Most previous research has been carried out in wet or temperate zones, since rainfall patterns there are relatively regular. In contrast, semi-arid areas (with low, unevenly distributed rainfall) have received less attention [2]. This disparity underscores the need for region-specific studies that explicitly address the challenges of rainfall prediction in semi-arid environments.

B. Boosting and Neural Network Models for Rainfall Prediction

Ensemble-based methods and neural networks are the two main modeling paradigms for rainfall forecasting among machine learning techniques. Tree-based ensemble learners, including Random Forest, have been extensively used in rainfall prediction because of their strength in modeling when interactions exist among input predictors and their efficacy in handling the heterogeneity of meteorological covariates [5]. Random Forest models have been shown to perform well in daily rainfall prediction and can yield better results than classical regression methods, particularly in systems with high rainfall variability [12], [13].

Gradient boosting methods such as XGBoost and LightGBM have received growing attention in recent rainfall prediction research. These models refine prediction accuracy by correcting residual errors. They are found to perform significantly better than single decision trees and classical neural networks in several rainfall estimation studies [6], [14], [15]. Their excellent ability to handle small sample sizes, missing values, and skewed rainfall distributions makes them particularly desirable for station-based rainfall forecasting in data-limited areas. CatBoost has also been investigated for precipitation prediction, with reported improvements in robustness and reduced bias in simulating complex rainfall patterns [8], [16].

Neural network techniques have also been widely studied in the context of rainfall prediction. Multilayer Perceptron (MLP) networks have been employed for daily rainfall forecasting using data from meteorological stations. They can produce satisfactory results when large volumes of training data are used and careful input feature selection is performed [9], [17]. More complex neural network structures have also been investigated to capture temporal relationships in precipitation series. However, these methods usually require substantial data and extensive hyperparameter tuning, making them challenging to use in station-based and semi-arid contexts.

Comparison analysis shows that boosting-based ensemble models generally yield more stable and robust predictions than neural network methods for tabular meteorological data and daily rainfall forecasting, especially in the presence of rainfall intermittency and scarce data [14], [15]. However, model performance is heavily dependent on climate, predictor selection, and evaluation methods, underscoring the need for consolidated benchmarking.

C. Research Gaps and Challenges in Semi-Arid Regions

Although machine learning models have attracted attention in rainfall prediction, there is still little research on this topic, especially in semi-arid areas. First, many studies were conducted under heterogeneous conditions for preprocessing, feature engineering, and evaluation metrics. This lack of uniformity prevents comparisons of reported results and makes it difficult to draw strong conclusions about model performance. Second, in semi-arid climates, rainfall occurs as long dry spells punctuated by brief, heavy rainfall events. This type of sparsity, intermittency, and skewness can pose major challenges in the development and performance of data-driven models when identifying hidden patterns from scarce rainfall observations. Therefore, performance observed in humid or data-rich regions cannot be extrapolated to semi-arid conditions without rigorous validation. Finally, few studies offer climate-sensitive interpretations of the performance of machine learning models directly related to the physical characteristics of rainfall processes in arid areas. Existing work tends to be accuracy-oriented and pays little attention to what makes the model perform better or worse in a particular climate. Addressing these

challenges requires unified experimental frameworks, region-specific analysis, and interpretative discussions that link model behavior to rainfall variability and data characteristics in semi-arid regions.

III. Study Area and Data Description

A. Study Area

The research was carried out in Settat, a city situated in west-central Morocco, and covers the semi-arid climatic zone. Settat is located at altitudes ranging from 200 to 416 m above sea level and is characterized by high interannual and intra-annual rainfall variability. The area is known for a regional weather regime characterized by long, dry seasons terminated by short, sporadic rainfall events, making precipitation prediction particularly difficult.

Seasonal temperature variations are pronounced. Maximum air temperatures in the summer period commonly exceed 35°C and can reach up to 46°C, while those in winter are generally milder (5 to 15°C). Precipitation is low throughout the year and varies from one month to another. These conditions render Settat a suitable study area for assessing the reliability of machine learning techniques for precipitation estimation in semi-arid climates, where precipitation processes are highly intermittent and nonlinear.

B. Data Sources and Description

The meteorological data used in this work are obtained from the Settat weather station of the General Directorate of Meteorology (GDM), Morocco. The dataset comprises daily records for 2014-2024. Owing to institutional policy on data sharing, raw data will not be made publicly accessible unless an individual requests it with the GDM.

The dataset consists of eight daily meteorological variables selected for their potential influence on rainfall generation processes. These variables include maximum air temperature (TMPMAX, °C), minimum air temperature (TMPMIN, °C), mean station pressure (Mean_Pressure, hPa), maximum relative humidity (HUMAXQ, %), minimum relative humidity (HUMINQ, %), maximum wind speed (FFXINS, m/s), maximum wind direction (DDXINS, °), and daily rainfall (RRQUOT, mm). Daily rainfall (RRQUOT) was used as the target variable, with all other weather variables as predictors. The selection of these variables is guided by their physical importance to the process and by their accessibility at the station scale.

C. Data Preprocessing

A uniform and well-designed preprocessing workflow was adopted for the dataset, maintaining its temporal structure and improving the model's learning. In a first step, meteorological data were ordered chronologically to match the time scales. In cases of missing values, we linearly interpolated them to produce a continuous time series while avoiding the introduction of artificial discontinuities. Cyclical encoding of the calendar variables was used to model the seasonal and temporal patterns. Precisely, the day and month information was processed by sine and cosine functions, and binary dummy variables were included to denote the four meteorological seasons. Furthermore, lagged features at several past time points were derived from the original predictors to capture temporal dynamics. This was in addition to rolling statistical features (daily and cumulative means, maxima, minima, standard deviations, and daily differences), which were also calculated for persistence and short-term variability in meteorological conditions.

Since tree-based ensemble models (Random Forest, XGBoost, LightGBM, and CatBoost) are not sensitive to feature scaling, we used the generated features directly. On the other hand, we get that the MLP model requires feature scaling; therefore, Min-Max normalization was applied only after the chronological data split to avoid data leakage. The data was divided into 70% for training, 15% for validation, and 15% for testing in a strict temporal order. Finally, non-informative attributes such as station IDs and raw date columns were dropped from the dataset, leaving only those with significance for weather conditions or the period. Linear interpolation was selected due to the relatively low proportion of missing values and its ability to preserve the temporal continuity of daily meteorological time series without introducing artificial variability.

It is important to note that this study is based on data from a single meteorological station located in Settat. This choice is primarily due to the limited availability of high-quality and continuous multi-station meteorological data in the region.

However, the Settat station is representative of the semi-arid climatic conditions of central Morocco, characterized by low, irregular rainfall. Therefore, the findings of this study provide meaningful insights into rainfall prediction in similar semi-arid environments.

Future work will focus on extending this approach to multi-station datasets and incorporating spatial variability to improve model generalization further.

IV. Methodology

A. Experimental Framework

In this research, a consolidated, organized experimental framework is proposed to enable an equitable comparison of machine learning models for daily rainfall prediction. The purpose of this study was to compare the predictive performance of ensemble tree models and neural network approaches under the same data requirements and evaluation protocols. All models were trained using the feature sets generated by the preprocessing pipeline described in Section III. The datasets were divided into 70% for training, 15% for validation, and 15% for testing while preserving the time sequences (no data leakage). Hyperparameters were optimized for the training and validation subsets only, and the test set was held off until the final performance assessment. This controlled setup ensures both methodological comparability and statistical reliability when comparing models. The overall workflow of the proposed methodology is illustrated in Fig. 1.

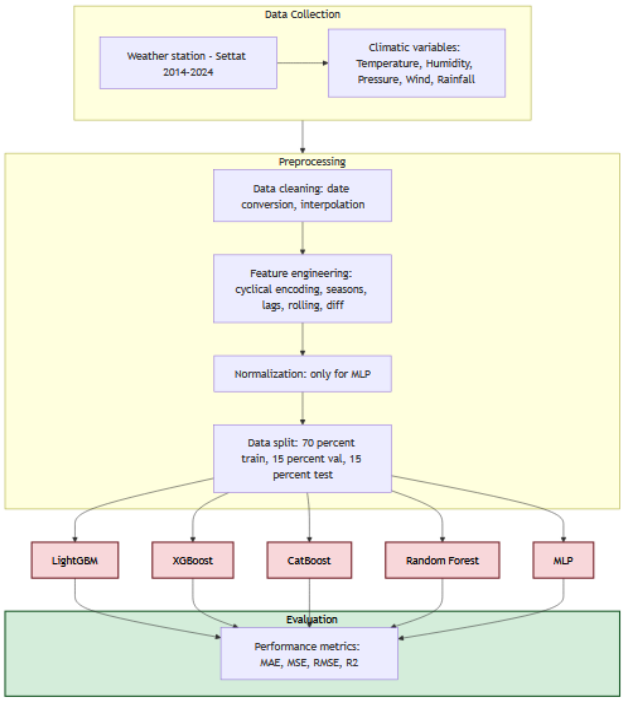

Fig. 1. Proposed methodology flowchart.

B. Machine Learning Models

The five supervised regression models used in this study were Random Forest, XGBoost, LightGBM, CatBoost, and Multilayer Perceptron (MLP). These models were selected to represent two dominant paradigms in tabular data modeling: ensemble-based tree methods and neural network architectures.

1) Random Forest (RF): Random Forest is an ensemble learning method that creates multiple decision trees via bootstrap sampling and random feature selection. The final prediction is obtained by averaging the three results. This strongly reduces variance and enhances generalization. Random forest is well-suited for a precipitation forecast model in a semi-arid region with an irregular and highly variable rainfall distribution because of its noise insensitivity and ability to handle multicollinearity and non-linear interactions.

2) Extreme Gradient Boosting (XGBoost): XGBoost is a gradient-boosting machine learning method that trains decision trees sequentially to minimize a regularized objective. The model adjusts the residuals from the previous trees at each iteration. The model also includes regularization to prevent overfitting and ensure the stability of the training model. XGBoost is well-known for its excellent predictive performance on structured data with nonlinear patterns.

3) Light Gradient Boosting Machine (LightGBM): LightGBM is a gradient boosting framework that uses a histogram-based learning algorithm, and it grows the tree leaf-wise. These models improve computational efficiency while preserving high predictive performance. LightGBM is particularly suited to structured meteorology datasets with engineered statistical and temporal features.

4) CatBoost: CatBoost is a gradient boosting algorithm that aims to reduce prediction bias by using ordered boosting methods. Although the data source used in this study contains mostly numerical features, we included CatBoost because it has been reported to be robust to noise and performs well for general tabular regression problems.

5) Multilayer Perceptron (MLP): The Multilayer Perceptron is a feedforward artificial neural network consisting of fully connected layers with nonlinear activation functions. For this analysis, an MLP was used as a regression model to capture complex nonlinear relationships between meteorological variables and daily rainfall. Neural Networks are sensitive to variations in feature scale. Therefore, the input variables were normalised using min–max normalisation before model training.

C. Hyperparameter Optimization

Hyperparameter optimization was performed on the training and validation subsets. Ensemble model hyperparameters, including the number of trees, maximum tree depth, and learning rate, were optimized to maximize predictive performance without overfitting. For the MLP network, the hidden neurons, learning rate, and number of training iterations were tuned. The final hyperparameter settings for each model were chosen based on validation performance. The test set was held out entirely during model tuning to obtain an unbiased estimate. The final selected hyperparameter values for each model are summarized in Table I.

TABLE I. HYPERPARAMETER SETTINGS OF THE EVALUATED MODELS

Model	Hyperparameters
LightGBM	learning_rate = 0.03; num_leaves = 31; feature_fraction = 0.8; bagging_fraction = 0.8; min_data_in_leaf = 20
XGBoost	learning_rate = 0.03; max_depth = 3; subsample = 0.8; colsample_bytree = 0.8; min_child_weight = 2
CatBoost	iterations = 1000; learning_rate = 0.03; depth = 6; l2_leaf_reg = 3
Random Forest	n_estimators = 200; max_depth = None
MLP	hidden_layers = (64, 32); activation = ReLU; learning_rate = 0.001; batch_size = 32; epochs = 120

Early stopping was applied for gradient boosting models and the MLP model to determine the optimal number of training iterations and prevent overfitting.

D. Evaluation Metrics

Model performance was evaluated using four complementary regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The MAE is the mean of the differences between observed and predicted rainfall values. MSE gives the mean of the squared errors and is sensitive to large deviations. RMSE: The square root of the MSE, RMSE, quantifies rainfall prediction error (millimetres), enabling interpretation. R^2 is the fraction of the variability in observed rainfall that the model explains. Their joint consideration enables accurate measurement of the quality and performance of models in semi-arid regions with low precipitation.

V. RESULTS

Before evaluating model performance, it is important to analyze the statistical characteristics of the rainfall data. The dataset is characterized by a high proportion of zero rainfall values, corresponding to long dry periods typical of semi-arid climates. In contrast, rainfall events occur sporadically and are often characterized by high intensity, resulting in a highly skewed distribution.

This intermittency and variability make rainfall prediction particularly challenging, as models must simultaneously capture both the absence of rainfall and sudden extreme events. These characteristics highlight the importance of using robust machine learning models that can handle nonlinear and imbalanced data distributions.

A. Overall Predictive Performance

The performance of the tested machine learning models was evaluated on the test set using four evaluation metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2). A comparative result for Random Forest (RF), XGBoost, LightGBM, CatBoost, and Multilayer Perceptron (MLP) is shown in Table II.

TABLE II. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS FOR DAILY RAINFALL PREDICTION.

Metrics	Method				
	LightGBM	***XGBoost***	***CatBoost***	***RF***	***MLP***
MAE	0.0507	0.0423	0.0534	0.0683	0.1628
MSE	0.0456	0.0530	0.0536	0.1199	0.3096
RMSE	0.2135	0.2301	0.2314	0.3463	0.5564
R^2	0.9868	0.9846	0.9844	0.9652	0.9107

The performance of the evaluated machine learning models is summarized in Table II using four evaluation metrics: MAE, MSE, RMSE, and R^2. The results clearly indicate that gradient boosting models outperform the other approaches across all metrics.

LightGBM achieved the best overall performance, with the lowest error values and the highest coefficient of determination ($R^2 = 0.9868$), indicating its strong ability to accurately model rainfall variability. XGBoost and CatBoost also demonstrated competitive performance, with slightly higher error values but still maintaining high predictive accuracy.

In contrast, the Random Forest model showed moderate performance, with higher error rates than boosting methods. The MLP model exhibited the weakest performance, with significantly higher errors and lower R^2 values, indicating difficulties in learning from intermittent, highly variable rainfall patterns.

The superior performance of gradient boosting models can be attributed to their sequential learning strategy, which iteratively reduces prediction errors and enhances model generalization. These models are particularly effective in handling nonlinear relationships and skewed data distributions, which are common characteristics of rainfall in semi-arid regions.

Furthermore, boosting algorithms are more robust to sparse data and are better suited to capture rare but significant rainfall events. In contrast, neural network models such as MLPs require larger, more balanced datasets to achieve stable performance, which limits their effectiveness in this context.

B. Feature Importance Analysis

To further interpret the predictive behavior of the best-performing model, a feature importance analysis was conducted using the LightGBM algorithm. This analysis provides insight into the relative contribution of each input variable to the rainfall prediction task.

The results indicate that meteorological variables such as relative humidity, temperature, and wind speed play a significant role in rainfall prediction. In particular, relative humidity was identified as one of the most influential features, as it directly reflects atmospheric moisture conditions necessary for precipitation formation.

Temperature variables also contribute significantly, as they influence evaporation and condensation processes. Wind-related features further enhance prediction performance by capturing atmospheric dynamics and moisture transport.

Overall, the feature importance analysis confirms that the selected input variables are physically relevant and consistent with the underlying climatic processes governing rainfall in semi-arid regions. This strengthens the reliability and interpretability of the proposed machine learning approach.

C. Observed vs. Predicted Daily Rainfall

To better understand the predictive behavior of the evaluated models, a visual comparison between observed and predicted daily rainfall values is presented for selected models. The selected models include LightGBM as the best-performing model, XGBoost as a competitive gradient boosting approach, and the Multilayer Perceptron (MLP) as a neural network baseline.

Fig. 2 illustrates the comparison between observed and predicted daily rainfall using the LightGBM model on the testing dataset. The predicted rainfall closely follows the observed temporal patterns, effectively capturing both prolonged dry periods and rainfall events. This demonstrates the model's strong capability in representing the nonlinear and intermittent nature of rainfall in semi-arid regions. However, slight underestimation is observed during high-intensity rainfall events, which is consistent with the difficulty of modeling extreme values in skewed precipitation distributions.

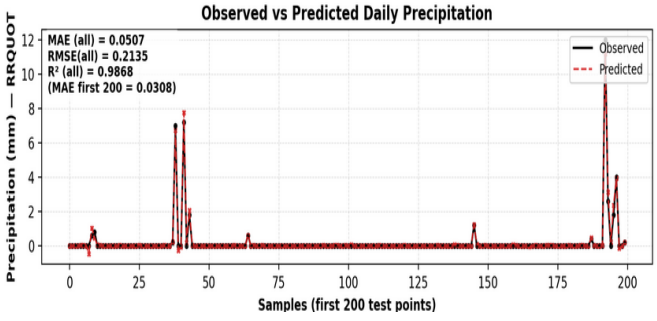

Fig. 2. Observed vs. predicted daily rainfall using LightGBM.

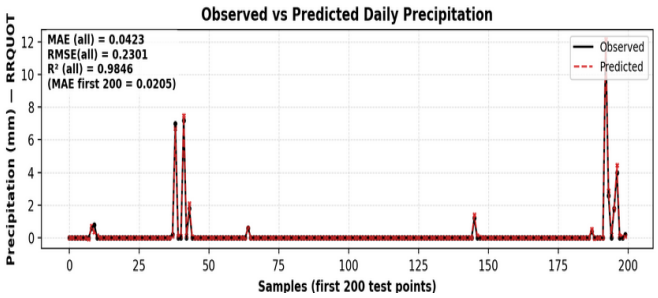

Fig. 3. Observed vs. predicted daily rainfall using XGBoost.

Fig. 3 presents the observed and predicted rainfall values using the XGBoost model. Similar to LightGBM, XGBoost can capture general rainfall dynamics and temporal variability with high accuracy. Nevertheless, minor deviations are evident compared to LightGBM, particularly during peak rainfall periods, suggesting slightly lower precision in modeling extreme events.

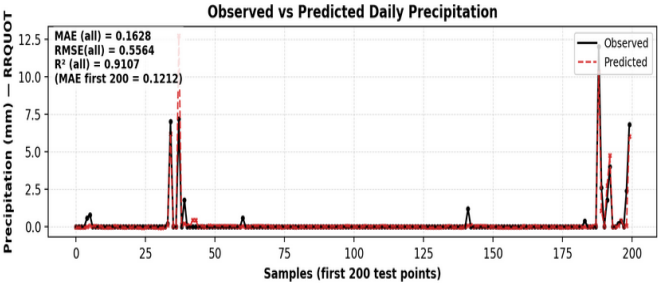

Fig. 4. Observed vs. predicted daily rainfall using MLP.

Fig. 4 shows the MLP model's performance in predicting daily rainfall. Unlike the boosting-based methods, the MLP model exhibits larger discrepancies between observed and predicted values. The model struggles to accurately reproduce rainfall variability, especially during abrupt transitions and extreme rainfall events. This behavior reflects the limitations of neural networks when applied to relatively small and highly intermittent meteorological datasets.

Overall, the visual comparison confirms the quantitative results presented earlier, demonstrating that gradient boosting models, particularly LightGBM and XGBoost, provide more accurate and stable predictions than the neural network model. These models are better suited to capture the intermittent, nonlinear, and skewed characteristics of rainfall in semi-arid environments.

D. Comparative Analysis of Model Behavior

Model comparison reveals the robustness and stability of gradient boosting models (LightGBM, XGBoost, and CatBoost) across rainfall intensity levels.

LightGBM has the lowest MAE, MSE, and RMSE values and the highest R^2, indicating a stronger ability to minimize prediction error while explaining more rainfall variability.

In contrast, the MLP model displays greater variability in prediction errors at all-time steps, especially during rapid precipitation changes. Random forests show reasonable robustness but are still behind boosting-based methods on the respective metrics evaluated.

These results indicate that boosting algorithms are particularly ideal for simulating the intermittent and skewed rainfall distributions commonly found in semi-arid environments.

The graphical comparison between observed and predicted rainfall values further supports these findings. As illustrated in Fig. 2 to Fig. 4, the boosting-based models provide a closer fit to the observed data, whereas the MLP model exhibits larger deviations, particularly during abrupt rainfall changes.

VI. Discussion

A. Interpretation of Model Performance

The experimental results indicated that all gradient boosting models, especially LightGBM, performed strongly in reducing prediction errors and explaining rainfall variability when predicting daily rainfall in the semi-arid region of Settat. For instance, a lower MSE value implies greater reliability against large prediction errors, and lower MAE and RMSE suggest improved consistency in reproducing rainfall intensity. The higher R^2 score also suggests that this model better explains rainfall variability than those analyzed.

Boosting algorithms are particularly strong because they iteratively refine residual errors to improve generalization via a sequential error-correction mechanism. This framework is especially beneficial for modelling nonlinear relationships between meteorological predictors and for accommodating skewed rainfall distributions. There have been similar findings in semi-arid rainfall prediction studies using ensemble and boosting methods [18], [19]. In contrast, broader reviews stress that AI-based rainfall forecasting frameworks are highly robust to climatic variability [20].

Conversely, the MLP showed greater variation in prediction errors, especially during rapid rainfall changes. Neural network models need more extensive, balanced datasets to achieve stable performance. Similar findings from comparative studies across various climatic regions indicate that deep learning models may struggle when trained on limited or highly intermittent rainfall datasets [21].

B. Comparison with Existing Literature

To evaluate the effectiveness of the proposed approach, the obtained results were compared with findings from recent studies on rainfall prediction in semi-arid regions.

Recent studies have demonstrated the effectiveness of machine learning and ensemble techniques for precipitation modeling. For instance, a recent study conducted in Morocco [18] reported high predictive performance using CatBoost and XGBoost models, with R^2 values reaching approximately 0.98. Similarly, other studies have demonstrated that machine learning and hybrid approaches achieve strong predictive performance in rainfall prediction tasks across different climatic conditions [22], [23].

In comparison, the proposed LightGBM model achieved an R^2 of 0.9868, demonstrating competitive, robust predictive performance. These results confirm that ensemble learning approaches are well-suited for modeling the nonlinear and highly variable nature of rainfall in semi-arid environments.

C. Implications and Study Limitations

Semi-arid areas are characterized by long dry periods interspersed with short-duration, high-intensity events, making rainfall forecasting a fundamental challenge in these regions. This intermittency results in highly skewed data distributions, making models sensitive to extreme values. The reduction in RMSE for the LightGBM across both extremes was observed, confirming its resilience to these extremes and further suggesting a move towards semi-arid rainfall modelling with this algorithm.

However, some limitations need to be considered. First, the analysis is based on a single meteorological station, which limits spatial generalizability. Second, daily one-step-ahead forecasting was only examined when multiple-step-ahead prediction frameworks may yield more operationally relevant insights [24]. Third, the more sophisticated spatiotemporal deep learning models were not applied in this study [25].

Nevertheless, the general, unified evaluation framework used in this study provides a clear standard against which to compare machine learning algorithms in semi-arid climates. It contributes to understanding the suitability of algorithms under intermittent precipitation regimes.

VII. Conclusion

This study presented a comprehensive evaluation of machine learning models for daily rainfall prediction in a semi-arid region of Morocco. Several models, including LightGBM, XGBoost, Random Forest, CatBoost, and MLP, were assessed using standard performance metrics.

The results demonstrated that the LightGBM model achieved the best performance, with a high coefficient of determination (R^2 = 0.9868) and low prediction errors, indicating its strong capability to model nonlinear and highly variable rainfall patterns.

These findings highlight the effectiveness of ensemble learning approaches for rainfall prediction, particularly in data-scarce and semi-arid environments.

However, this study is limited by the use of data from a single meteorological station, which may affect the generalization of the results.

Future work will focus on extending this approach to multi-station datasets, incorporating spatial and temporal variability, and exploring advanced deep learning architectures, such as LSTMs and hybrid models, to further improve prediction accuracy.

Acknowledgment

The authors would like to thank the General Directorate of Meteorology (DGM) in Morocco for providing the meteorological data used in this study.

References

[1] M. V. K. Sivakumar, A. C. Ruane, and J. Camacho, "Climate change in the West Asia and North Africa region," in Climate Change and Food Security in West Asia and North Africa, Dordrecht, The Netherlands: Springer, 2013, pp. 3–26, doi: 10.1007/978-94-007-6751-5_1.

[2] L. V. Alexander et al., "Global observed changes in daily climate extremes of temperature and precipitation," J. Geophys. Res. Atmos., vol. 111, no. D5, Art. No. D05109, 2006, doi: 10.1029/2005JD006290.

[3] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," Nature, vol. 525, no. 7567, pp. 47–55, 2015, doi: 10.1038/nature14956.

[4] M. G. Schultz et al., "Can deep learning beat numerical weather prediction?" Phil. Trans. R. Soc. A, vol. 379, no. 2194, Art. no. 20200097, 2021, doi: 10.1098/rsta.2020.0097.

[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discover. Data Min. (KDD), San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[7] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 3146–3154.

[8] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Montréal, QC, Canada, 2018, pp. 6638–6648.

[9] A. Maier and A. Dandy, "Neural networks for the prediction and forecasting of water resources variables: A review," Environmental Modelling & Software, vol. 15, no. 1, pp. 101–124, 2000, doi: 10.1016/S1364-8152(99)00007-9.

[10] H. Hersbach et al., "The ERA5 global reanalysis," Q. J. R. Meteorol. Soc., vol. 146, no. 730, pp. 1999–2049, 2020, doi: 10.1002/qj.3803.

[11] M. Reichstein et al., "Deep learning and process understanding for data-driven Earth system science," Nature, vol. 566, no. 7743, pp. 195–204, 2019, doi: 10.1038/s41586-019-0912-1.

[12] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," Journal of Big Data, vol. 8, no. 153, 2021, doi: 10.1186/s40537-021-00545-4.

[13] M. Abu Saleh, H. M. Rasel, and B. Ray, "A comprehensive review towards resilient rainfall forecasting models using artificial intelligence techniques," Green Technologies and Sustainability, vol. 2, No. 3, Art. no. 100104, 2024, doi: 10.1016/j.grets.2024.100104.

[14] A. Sreenivasu, S. Rafi, and C. Rajani, "Rainfall prediction using machine learning," in 2024 IEEE International Conference on Machine Learning and Computing (ICMACC), 2024, pp. 86–90, doi: 10.1109/ICMACC62921.2024.10894486.

[15] 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, vol. 39, no. 1, pp. 271–290, 2025, doi: 10.1007/s11269-024-03969-8.

[16] V. Kumar, N. Kedam, K. V. Sharma, K. M. Khedher, and A. E. Alluqmani, "A comparison of machine learning models for predicting rainfall in urban metropolitan cities," Sustainability, vol. 15, no. 18, Art. no. 13724, 2023, doi: 10.3390/su151813724.

[17] I. V. Necesito, D. Kim, Y. H. Bae, K. Kim, S. Kim, and H. S. Kim, "Deep learning-based univariate prediction of daily rainfall: application to a flood-prone, data-deficient country," Atmosphere, vol. 14, no. 4, Art. no. 632, 2023, doi: 10.3390/atmos14040632.

[18] A. Elmotawakkil, A. Moumane, A. Sadiki, et al., "Forecasting short-term rainfall patterns in arid and semi-arid regions using machine learning and deep learning models: A case study from Morocco," Theoretical and Applied Climatology, 2025, doi: 10.1007/s00704-025-05677-8.

[19] M. El Hafyani, K. El Himdi, and S. E. El Adlouni, "Improving monthly precipitation prediction accuracy using machine learning models: A multi-view stacking learning technique," Frontiers in Water, vol. 6, 2024, doi: 10.3389/frwa.2024.1378598.

[20] F. A. F. Sham, A. El-Shafie, W. Z. Wan Jaafar, M. Sherif, and A. N. Ahmed, "Advances in AI-based rainfall forecasting: A comprehensive review of past, present, and future directions with intelligent data fusion and climate change models," Results in Engineering, vol. 27, 2025, doi: 10.1016/j.rineng.2025.105774.

[21] O. A. Wani et al., "Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the north-western Himalayas," Scientific Reports, vol. 14, 2024, doi: 10.1038/s41598-024-77687-x.

[22] A. K. Sharma, R. Gupta, and S. Verma, "How accurate are machine learning models in improving monthly rainfall prediction in hyper-arid environments?," Journal of Hydrology, vol. 633, 2024, doi: 10.1016/j.jhydrol.2024.131040.

[23] M.-H. Lee and Y. J. Chen, "Precipitation modeling for extreme weather based on sparse hybrid machine learning and Markov chain random field," Water, vol. 13, no. 9, p. 1241, 2021, doi: 10.3390/w13091241.

[24] S. Narejo et al., "Multi-step rainfall forecasting using deep learning approach," PeerJ Computer Science, vol. 7, e514, 2021, doi: 10.7717/peerj-cs.514.

[25] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," arXiv preprint arXiv:1706.03458, 2017.