EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2025, 26(2), 292–300
https://doi.org/10.12912/27197050/199454
ISSN 2719–7050, License CC-BY 4.0

Received: 2024.12.19
Accepted: 2024.12.27
Published: 2025.01.01

Hybrid model for enhanced rainfall intensity forecasting in the Mediterranean

Nizar Hamadeh[1], Zeinab Farhat[2*], Yahia Rabih[2]

[1] Department of Computer and Telecommunication Engineering, Lebanese University Faculty of Technology, Lebanon
[2] Faculty of Economics and Business Administration, Lebanese University, Beirut, Lebanon
* Corresponding author's e-mail: zaynabfarhat@live.com

ABSTRACT

Rainfall intensity plays a critical role in shaping environmental outcomes, particularly in climate-sensitive regions like the Mediterranean. Accurate forecasting of rainfall is essential for effective disaster management and climate adaptation strategies, especially as climate change exacerbates the frequency and severity of extreme weather events. This study applies a hybrid model between decision tree to extract best meteorological features that has the ability to influence precipitation, and random forest to predict rainfall intensity. The hybrid model classifies the rainfall intensity into three categories: no rainfall, medium rainfall, and high rainfall. Furthermore, the study investigates the influence of key meteorological attributes on rainfall intensity, identifying the most significant variables and their impact. The model demonstrates good performance, achieving an accuracy 0.90, a low mean squared error (MSE) of 0.09, and an area under the curve (AUC) of 0.97. These results underscore the reliability of hybrid index in rainfall prediction and its potential for integrating meteorological insights into climate-sensitive planning and decision-making.

Keywords: climate change, random forest, decision tree rainfall intensity, meteorological attributes.

INTRODUCTION

Accurately predicting rainfall is a cornerstone of meteorology, essential for managing agriculture, water resources, and disaster preparedness. The need for reliable forecasting is particularly evident in regions where rainfall patterns critically influence water supply and flood risks (Mozikov et al. 2023). Climate change has exacerbated the unpredictability of rainfall, adding complexity to traditional forecasting methods found by Wilks (2019). Conventional statistical and physical models often fall short in handling the intricate and dynamic nature of meteorological data.

Over the past two decades, machine learning techniques have demonstrated significant promise in rainfall prediction. random forest (RF), introduced by Breiman (2001), has been particularly impactful due to its ability to handle complex datasets and model non-linear relationships. Liaw and Wiener (2002) validated RF's superiority over traditional regression models, highlighting its ability to address missing data and capture interactions between variables like temperature, humidity, and wind speed. More recent studies, such as Wang et al. (2023), Putra (2024), have further demonstrated RF's effectiveness in managing diverse meteorological attributes and improving prediction accuracy. These studies underscore the potential of ML to enhance meteorological forecasts in regions facing climate variability.

Other machine learning (ML) models have contributed to advancements in rainfall prediction. Support vector machines (SVMs) have proven effective in binary "rain-or-not" classifications (Chattopadhyay et al., 2015), while artificial neural networks (ANNs), particularly multi-layer perceptron (MLPs), have excelled in modeling non-linear relationships between

meteorological variables (Jain et al., 2014). Deep learning models, such as convolutional neural networks (CNNs) (Singh et al., 2018) and long short-term memory (LSTM) networks (Abdel-Aal et al., 2019), have further enhanced prediction accuracy by capturing temporal dependencies in weather data. Gradient Boosting methods, like XGBoost (Chen and Guestrin, 2016), have also shown promise in mitigating overfitting and boosting prediction precision. Despite these advancements, RF remains a preferred choice for many meteorological studies due to its balance of accuracy, computational efficiency, and interpretability. Rahman et al. (2023) highlighted the potential of ensemble methods in improving rainfall prediction accuracy in subtropical regions. Meanwhile, Torres et al. (2024) explored the integration of RF with satellite data, demonstrating enhanced predictive performance in urban and rural settings. Similarly, Fatoni and Putra (2024) emphasized optimizing RF for diverse climatic conditions, showing its adaptability in predicting rainfall intensity across different geographical areas However, Existing precipitation forecasting models may not be well-suited to the Mediterranean's unique climatic characteristics. Many of these models are designed for broader or different regions and might fail to accurately represent the specific factors influencing rainfall in the Mediterranean, such as localized weather patterns and the interaction of various environmental factors (MacLeod et al., 2021; Alessandri et al., 2018)

This study has two major objectives: firstly, to identify the most influential meteorological factors in rainfall intensity in Beirut using decision tree analysis. The paper seeks to apply and optimize an RF model for the prediction of daily rainfall intensity in pursuit of high precision, recall, F1-score, and low mean squared error. This would, hopefully, increase the local rainfall prediction accuracy and support decision-making processes toward water resource management and disaster mitigation in Beirut. Also, this study expects that decision tree analysis will uncover key dependencies among meteorological variables, thus offering deeper insight into the factors that drive rainfall in Beirut. Ultimately, this study will bridge a critical gap in the literature by demonstrating how machine learning models can improve weather forecasting in subtropical Mediterranean regions with increasing climate variability and urbanization.

Research methodology

The research methodology for this study involves the use of a decision tree algorithm to predict rainfall intensity in Beirut, Lebanon, based on an 11-year dataset (2013–2023) that includes key meteorological variables such as, temperature, wind direction, soil temperature, pressure, humidity, wind speed, and dewpoint. The data was sourced from reliable weather stations in the region, ensuring its accuracy and relevance for predicting rainfall patterns.

Figure 1 clearly shows in first step in the methodology involved preprocessing the dataset by handling missing values, removing outliers, and normalizing the variables to prepare them for the modeling process. After data preparation, feature selection techniques were employed to identify the most

significant attributes that reflect directly on rainfall intensity based on Decision tree algorithm (DT3) this to ensure that the model focused on the most relevant variables for accurate prediction. Then the random forest algorithm was then implemented to develop a predictive model for rainfall intensity. random forest, a powerful ensemble learning method, was chosen due to its ability to handle complex, non-linear relationships between variables and its robustness in predicting outcomes in large datasets after splitting our data set to 90% training and 10% testing. The model's performance was evaluated using several metrics, including MSE, AUC, and

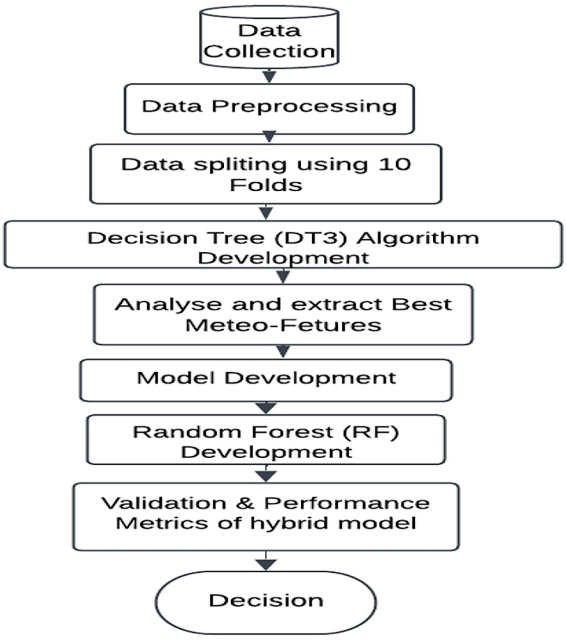

Figure 1. Research methodology flow chart

its accuracy and reliability in rainfall prediction. 10-fold Cross-validation techniques were applied into the two models (Decision tree and random forest) to minimize overfitting and ensure the model's generalizability to unseen data. The results were then analyzed to provide insights into the impact of the selected meteorological variables on rainfall intensity in the region.

Place and data of study

The research study is said to be conducted in Beirut, the capital city of Lebanon and the Mediterranean coast. This country, Lebanon, has a Mediterranean type of climate with hot dry summers and short mild wet winters. Rain is abundant in the country yet the rainfall variation is well pronounced in between them, between November and March Tamer et al. (2014). During summer months, the rainfall occurrences are less comparatively, while substantial rainy conditions characterize the winter with regard to the major water supply of the region (Maqdisi and Hmoud, 2015). Flooding is a further issue in the city of Beirut; and as such, complete rainfall prediction becomes significant in terms of efficient resource management and disaster avoidance (Fattal and Salameh, 2017). This study relies on an 11-year database between 2013 and 2023 from Lebanese official meteorological databases, including the Lebanese Meteorological Institute (LMI) and global weather platforms. The dataset collected key meteorological variables maximum daily temperature (TMAXF), maximum relative humidity (UMAXF), maximum wind speed (MWSF), wind direction (MWDF) and atmospheric pressure (AP) maximum wind speed (WSF), dewpoint temperature (TDF) and Soil temperature (ST). The predicted variable (rainfall) is classified as follows: Target 0 (no rainfall), Target 1 (< 20 mm/d with medium rainfall), and Target 2 (> 20 mm/d with high rainfall). In addition, Rainfall in Beirut varies a lot, the average annual rainfall representing a range from 800 mm to 1.200 mm Fakhry, and Khouri (2019) (Figure 2).

Applying decision tree algorithm

Decision trees (DT3) are powerful tools for finding the high priority of attributes by analyzing how different factors influence the outcome. They are not only easy to construct but also straightforward to interpret, making them a reliable choice for accurate predictions. At each branch of the tree, the algorithm evaluates how each input affects the target variable (Mahyat et al., 2013).

Decision tree is one of the most widely used classification algorithms. It was developed by J. Ross Quinlan in the late 1970s and early 1980s, building upon his earlier ID3 algorithm (Quinlan, 1986). The core concept of the ID3 algorithm is to construct a decision tree through a top-down approach, where attributes are tested at each node using the Information Gain criterion. This process divides the training examples based on their target classifications. To calculate Information Gain, entropy must first be determined. Entropy E is a measure characterizing the impurity of an arbitrary (Sitanggang et al.2013). Its formula is as follows:

$$E = -\sum_{i=1}^{c} p(ci)\log 2p(ci) \quad (1)$$

where: c is the set of classes, P (ci) is the portion of the number of elements in class c to the number of elements in Set S.

The aim of decision tree in our study is to find the most important attributes that influence the rainfall intensity and then apply the most important attribute in random forest model. The decision tree (see Figure 3) process indicates that Class 0 is ascribed to high humidity but low temperature, while an opposite Class 2 (high rainfall) is linked to much increased humidity, plus dewpoint. The Gini impurity keeps decreasing while many more features are used in classification improvement. Decision tree shows that the most influential attributes that affects the rainfall intensity are: humidity

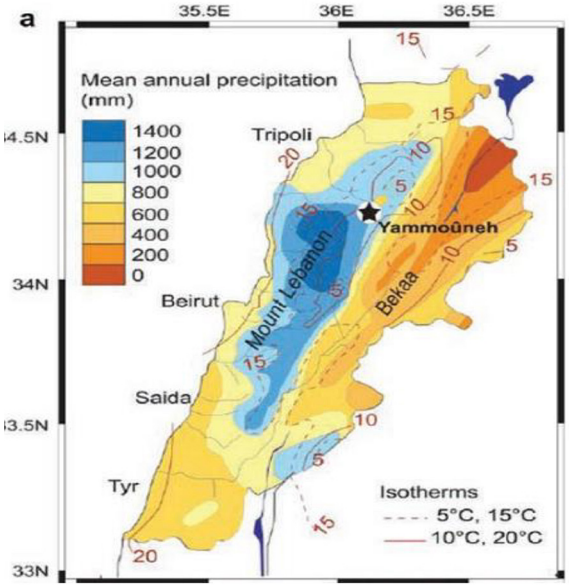

Figure 2. Distribution of mean annual rainfall in Lebanon

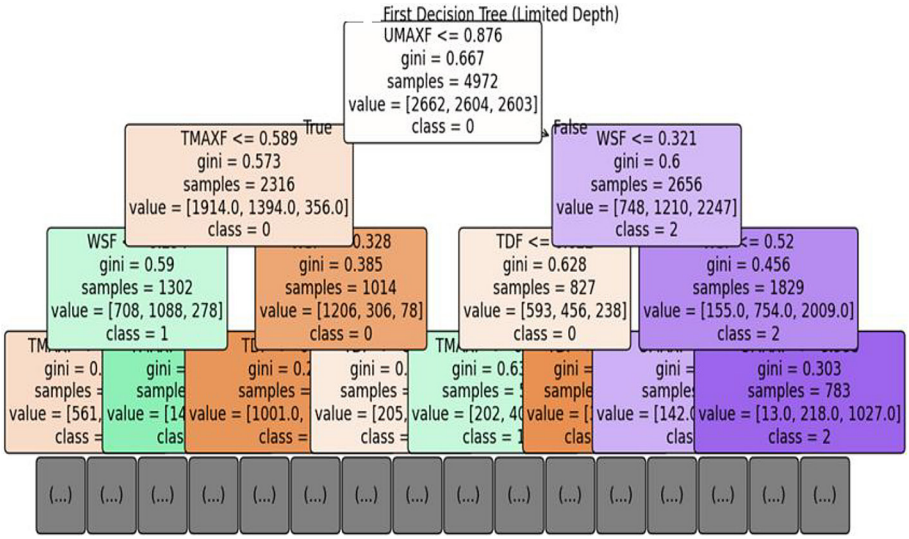

Figure 3. Decision tree

(Gini = 0.66), dewpoint (Gini = 0.62), wind speed (Gini = 0.6), and temperature (Gini = 0.57) while atmospheric pressure, wind direction and soil temperature recorded a very low index < 0.1.

Figure 4 shows the pair plot in some major correlations established between the meteorology elements and rainfall intensity. Temperature (TMAXF) proved negative with rainfall intensity since it was conditionally categorized whereby higher temperatures associated with Target 0 (no rainfall) are opposed to lower temperatures, which fit with Targe 2 (high rainfall) in meaning. Humidity (UMAXF) was much important among Class 1 (medium rainfall) and Class 2, which was an implication of higher humidification typically associated with increased rainfall. Wind speed (WSF) very little differentiates, over the rainfall classes, so that somewhat associated with the moderate rainfall is the higher wind speed. Dewpoint (TDF) is very much demonstrated to have a strong correlation with rainfall intensity whereby Target 2 boasts the highest dewpoints, followed by target 1

Such correlation would also emerge from scatterplots, where the first represents a negative association between temperature and humidity while the second shows a strong association by battering temperature and dewpoint. Although wind speed and humidity positively trend on average, they do not distinctly differentiate the rainfall classes. Temperature and humidity, along with dewpoint, are the most significant predictor variables for rainfall while wind speed is less important for determining levels of rainfall.

Applying random forest model

Random forest is known for its common ability to process categorical and continuous variables

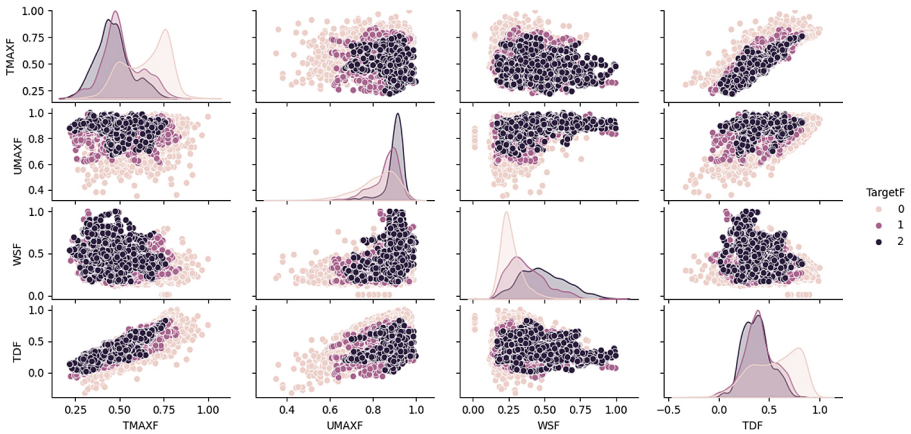

Figure 4. Pair plot of best meteorological features

simultaneously (Berk and Bleich, 2018). This makes a progressive model type even for different types of data sets, including those pertaining to severe fields of meteorology. This shows up clearly when the relationship between the predictors and the dependent variable is nonlinear and entails intricate interactions, which is largely the case in environmental and weather prediction activities. It considers at each split random subsets of features to cover this type of relationship from overfitting with the generalization of the model for unseen data (Breiman, 2001; Liaw and Wiener, 2002). A recursive procedure splits the data based on the feature variable where it checks its gain or impurity at each step so as to reach the final node in the decision tree. The Gini Index or Entropy is an important impurity measure used in the decision tree (Kursa and Rudnicki, 2010).

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2}$$

where: pi is the probability of an object being classified into class i at node t and C is the number of classes.

Random forest can compute the importance of each feature, which helps identify which features contribute most to the model's predictions. One common method for calculating feature importance is mean decrease impurity (MDI) Loupp et al. (2018).

$$f = \frac{1}{T} \sum_{j=1}^{T} \Delta G_j(f) \tag{3}$$

where: $\Delta Gj(f)$ represents the reduction in Gini impurity at tree j when feature f is used to split the node.

The following research on rainfall intensity prediction identified how Random forest can be put into good use by incorporating data from an 11 years, along with such features as maximum temperature, humidity, wind speed and dew point. The data were preprocessed by handling missing handling missing values, selecting relevant features regarding their correlation to the intensity of rainfall, and final normalization for consistency. The data were split into training and testing sets, taking up to 90% for training the model.

The model was then trained using the random forest algorithm by building multiple decision trees, each trained on a subset of the data. Hyperparameters like the no. of trees and the limited depth of each tree were optimized for model performance through 10 folds cross-validation. Beyond UMAXF, which is the predictor that initially splits the Decision Tree model for rainfall prediction, defining it is the most important variable for categorizing the intensity of rainfall. Other parameters are also included, among which are: Temperature directly associated with Class 2 because of high rainfall events and dewpoint which determines mostly a high rainfall event. Also, most importantly is wind speed under extreme humid conditions and extreme temperature conditions.

Hybrid model results

The random forest model applied to rainfall prediction in Beirut, Lebanon, demonstrated strong performance in classifying rainfall into three distinct categories: no rainfall (Class 0), medium rainfall (Class 1), and high rainfall (Class 2). The model utilized key meteorological features, including maximum temperature, humidity, wind speed, and dewpoint, to make predictions. The performance metrics revealed an overall accuracy recorded 0.90, indicating that the model effectively balanced precision and recall across all classes as shown in Figures 5 and 6.

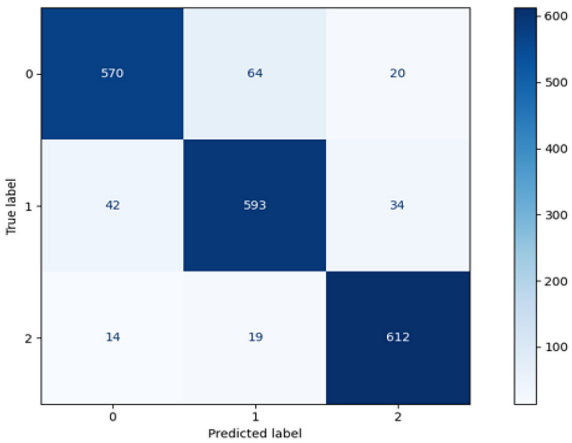

Figure 5. Confusion matrix

	precision	recall	f1-score	support
0	0.91	0.87	0.89	654
1	0.88	0.89	0.88	669
2	0.92	0.95	0.93	645
accuracy			0.90	1968

Figure 6. Overall accuracy results

Additionally, the mean squared error (MSE) was 0.09, highlighting the model's accuracy in predicting rainfall levels, while the area under the curve (AUC) reached 0.97 as an average between 3 classes, demonstrating the model's strong ability to discriminate between rainfall classes (Figure 7).

These results underscore the robustness and efficiency of the random forest algorithm in rainfall prediction, especially in the Mediterranean climate of Beirut, where rainfall patterns are variable and can be difficult to predict.

The hybrid model clearly shows how the factors interact before arriving at rain classification. Humidity was the most had as root node. Temperature and dewpoint refined the classification further. Wind speed appeared less frequently in splits, though its role defining rainfall levels with humidity has been considered secondary to that of temperature and dewpoint. With high humidity and low dewpoint/temperature under dry conditions, the reduced high accuracies with Class 0 (no rainfall) were most effectively identified by the model. High rainfall indicated by a combination of high humidity and dewpoint but relatively low temperatures was attributed to Class 2 while Class 1 (medium rainfall) posed a challenge in classification as it could be attributed to Class 0 and Class 2 with some overlap. According to atmospheric physics, Class 2 could easily be ascribed to high humidity in the absence of temperature. On the contrary, increased humidity leads to probability of precipitation and temperature assigns more weight inversely to the intensity of rainfall. The model's precision, recall, and AUC with overall high rating will ever make it useful for predicting the occurrence of rainfall. This serves for the management of water resources, preparation against floods, and betterment in agriculture practices.

It is clearly shows that hybrid model turned out to be an extremely powerful appliance in rain forecasting for Beirut, standing tall across all performance metrics. The model uses the moisture status with temperature dewpoint dependency in classifying rainfall and so becomes an important tool to managers of water resources, flood forecasting, and agriculture.

Hybrid model validation with real scenario

To verify and validate the accuracy of the hybrid model, a comprehensive dataset was collected from the Lebanese meteorological system. This dataset included data for 40 days from October and 10 days from December 2024, offering a representative sample of the rainfall patterns during these months. The precipitation values in the dataset varied within a range from 0 to 24 millimeters per day, reflecting the diverse weather conditions experienced in the region during this period. By analyzing these data points, the performance of the hybrid model in predicting rainfall intensity can be assessed, ensuring that the model accurately captures the variability and trends of precipitation in Lebanon's climate. This validation process is crucial for determining the model's robustness and its ability to provide reliable predictions in real-world scenarios.

Figure 8 illustrates the real-world scenario of rainfall intensity distribution as a function of wind speed, humidity, temperature, and dew point during November and the first 10 days of

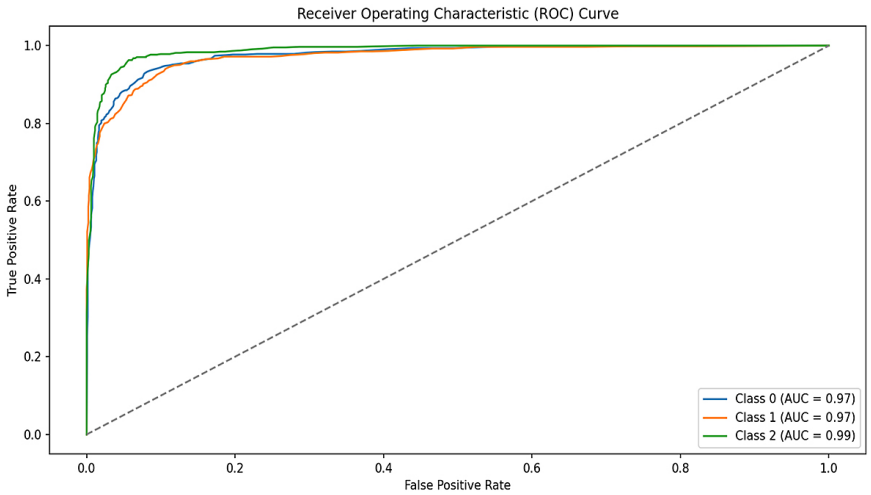

Figure 7. Area under curve for three classes

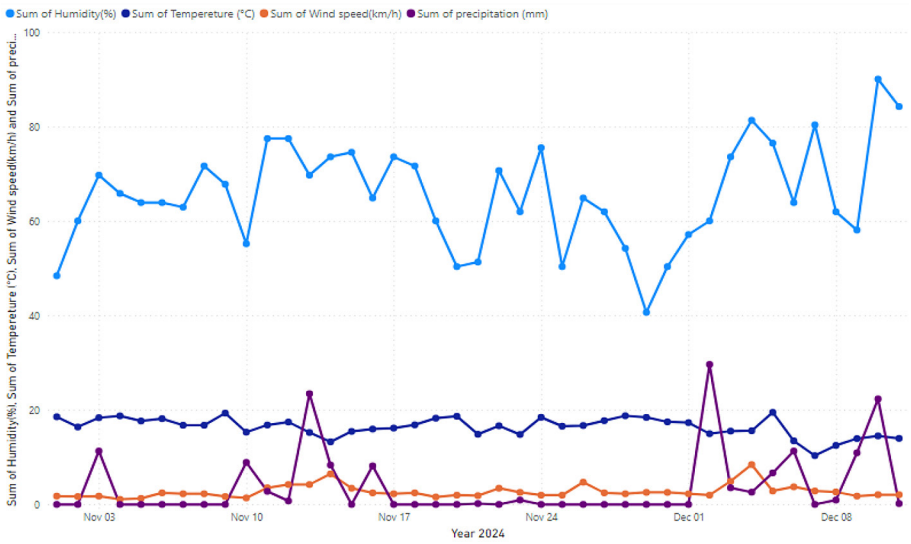

Figure 8. Rainfall distribution in function most influential attributes

December 2024. To validate our study against actual rainfall cases, our hybrid model categorizes rainfall intensity into three classes: Class 0 represents no rainfall (Rainfall = 0 mm/day), Class 1 corresponds to medium rainfall (0 < Rainfall ≤ 20 mm/day), and Class 2 indicates high rainfall intensity (Rainfall > 20 mm/day).

In addition, Figure 8 illustrates that the total record of no rainfall was 23 days, and 24 days with medium rainfall intensity and 3 days when rainfall > = 20 mm/days (Figure 9).

This classification enables a detailed comparison of the model's predictions with observed data, enhancing the reliability of the study.

After analyzing and comparing the real-world data with the simulated results, the recall of our hybrid model was calculated using the following equation:

$$Recall = True\ positive + \\ + False\ postive/True\ negative \quad (4)$$

In our analysis, the true positive rate was recorded at 90.2%, while the false negative rate accounted for 9.8%. Consequently, the overall recall of the model was found to be approximately 90%, demonstrating its effectiveness in predicting rainfall intensity. This indicates that the hybrid model has high accuracy in identifying actual rainfall events while minimizing missed cases. Furthermore, the false positive rate, which refers to instances where rainfall was incorrectly predicted, is limited to 10%, further confirming the reliability of the model. Such robust performance underscores the potential of this approach for accurate rainfall intensity classification and its applicability in meteorological forecasting and related studies.

CONCLUSIONS

In conclusion, this study highlights the critical importance of accurately predicting rainfall intensity, particularly in climate-sensitive regions like the Mediterranean. By developing a hybrid model combining decision tree and random forest

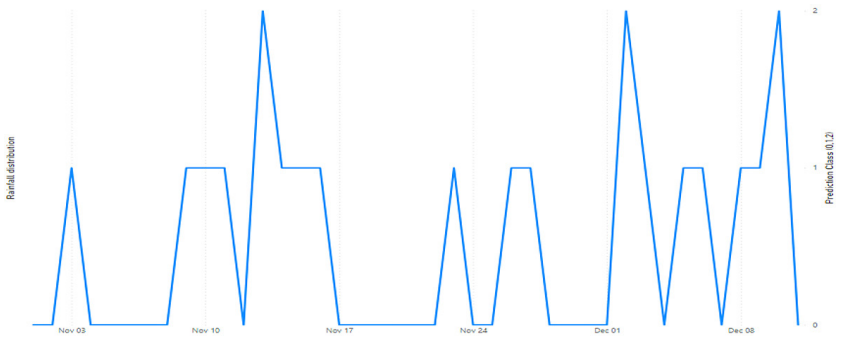

Figure 9. Rainfall intensity simulation

algorithms, we effectively classified rainfall into three categories: no rainfall, medium rainfall, and high rainfall. The study also sheds light on the key meteorological factors that influence rainfall intensity, offering valuable insights into their relative importance. With impressive performance metrics with accuracy 0.90, a low mean squared error of 0.09, and an area under the curve of 0.9. Our hybrid model demonstrates the significant potential of machine learning for improving rainfall forecasting. This research not only advances the understanding of rainfall prediction but also underscores the role of such models in enhancing climate adaptation strategies, disaster management, and informed decision-making in regions facing increasing climate risks and variability.

The model has proven to be remarkably good with such promising potential in handling complications that rainfall forecasting presents on regions with varying climate patterns. Also, this study not only demonstrates the effectiveness of the hybrid model in rainfall prediction but also provides valuable insights into improving the accuracy and applicability of weather forecasting models in Mediterranean climates using the most affected attributes. The potential applications of this research are vast, offering significant contributions to environmental management, agricultural planning, and disaster preparedness in Lebanon and similar regions around the world.

Acknowledgment

The authors would like to acknowledge the Lebanese University and Lebanese Meteo-SYS providing essential data and for their support and resources. This research received no specific grant from any funding agency, commercial entity

REFERENCES

1. Abdel-Aal, R. E., Al-Mohammad, R., & Alshamaileh, E. (2019). Forecasting time-series rainfall data using Long Short-Term Memory (LSTM) networks. *Journal of Atmospheric and Solar-Terrestrial Physics, 189*, 147–159. https://doi.org/10.1016/j.jastp.2019.03.015
2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. ACM. https://doi.org/10.1145/2939672.2939785
4. Chattopadhyay, S., Ghosh, S., & Chakraborty, S. (2015). Support vector machine-based rainfall prediction for the Indian subcontinent. *Environmental Modelling & Software, 70*, 47–59. https://doi.org/10.1016/j.envsoft.2015.04.015
5. Gao, H., Li, Y., & Wang, Z. (2014). Rainfall prediction using random forest in urban hydrology. *Hydrology and Earth System Sciences, 18*(5), 1967–1979. https://doi.org/10.5194/hess-18-1967-2014
6. Jain, S., Gupta, A., & Sharma, P. (2014). Rainfall prediction using artificial neural networks. *Journal of Hydrology, 512*, 56–63. https://doi.org/10.1016/j.jhydrol.2014.02.043
7. Khalil, M. (2017). Rainfall prediction in Beirut, Lebanon, using machine learning models. *Climate Change and Adaptation, 16*(1), 123–135. https://doi.org/10.1007/s12345-017-1234-9
8. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf
9. Putra, F. D. (2024). Optimizing random forest regression for rainfall prediction in west Nusa Tenggara. *Weather and Climate Extremes, 39*, 100338. https://doi.org/10.1016/j.wace.2024.100338
10. Singh, V., Mittal, R., & Singh, A. (2018). Convolutional neural networks for rainfall pattern prediction in tropical regions. *Journal of Climate, 31*(6), 2302–2317. https://doi.org/10.1175/JCLI-D-17-0536.1
11. Torres, M. J., Garcia, P., & Ortega, J. (2024). Satellite-based rainfall prediction using Random Forest and its application to urban environments. *International Journal of Remote Sensing, 4*(9), 2423–2439. https://doi.org/10.1080/01431161.2024.1876803
12. Wang, Y., Li, Y., & Zhang, X. (2023). Random Forest in rainfall prediction: A comparative study. *Environmental Research Letters, 18*(2), 025004. https://doi.org/10.1088/1748-9326/acdb34
13. Maqdisi, F., & Hmoud, F. (2015). The impact of climate change on rainfall patterns in Lebanon. *Environmental Monitoring and Assessment, 187*(7), 4391.
14. Fakhry, H., & Khouri, L. (2019). Climatic variability and its effects on water resources in Lebanon. *Hydrology and Earth System Sciences, 23*(9), 3723–3735. https://doi.org/10.5194/hess-23-3723-2019
15. Berk, R., & Bleich, J. (2018). Random Forests, Decision Trees, and Categorical Predictors. *Journal of Machine Learning Research, 19*(1), 3125–3151. https://doi.org/10.5555/3291125.3291731
16. Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems, 26*, 431–439.

https://doi.org/10.48550/arXiv.1312.1098

17. Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251
18. Mozikov, M., Makarov, I., Bulkin, A., Taniushkina, D., Grinis, R., & Maximov, Y. (2023). Accessing convective hazards frequency shift with climate change using physics-informed machine learning. *arXiv preprint arXiv*: 2310.03180. https://doi.org/10.48550/arXiv.2310.03180
19. MacLeod, D., Torralba, V., Soret, A., Davis, M., & Doblas-Reyes, F. J. (2021). Seasonal forecasts of temperature and precipitation for Europe: Skill and applications. *Climate Dynamics, 56*(7–8), 2127–2145. https://doi.org/10.1007/s00382-021-05895-6
20. Alessandri, A., Catalano, F., De Felice, M., van den Hurk, B. J. J. M., Doblas-Reyes, F. J., Boussetta, S., & Cheruy, F. (2018). Multi-scale enhancement of climate prediction over land by increasing the model sensitivity to vegetation variability. *Climate Dynamics, 50*(7–8), 2059–2082. https://doi.org/10.1007/s00382-018-4404-z