




A Hyperparameter Optimization Framework Using Optuna for XGBoost-based Drought Zone Classification in Flores Island

Alfredo Ananta Turkaemli* and Irfan Dwiguna Sumitra 

Universitas Komputer Indonesia, Bandung, Indonesia
*alfredo.75124008@mahasiswa.unikom.ac.id

Abstract. The purpose of this study is to classify drought zones on Flores Island, Indonesia, using the Extreme Gradient Boosting (XGBoost) algorithm. XGBoost is well known for its ability to handle tabular data with high complexity. However, previous studies have shown that the performance of XGBoost largely depends on the choice of hyperparameters. Therefore, in this study, XGBoost was combined with Optuna to identify the optimal hyperparameters during the modeling stage. The dataset was obtained from NASA POWER, covering the period 2015–2024, with a total of 27,214 daily climate records, which were preprocessed and categorized into wet, normal, and dry zones. The model was trained and tested using an 80/20 split, while Optuna was applied with 100 trials for hyperparameter tuning. The final model achieved 99.3% accuracy with balanced precision, recall, and F1-scores across all classes. Feature importance analysis highlighted humidity and maximum temperature, along with precipitation, as the most influential factors in classification. Overall, the study demonstrates that combining XGBoost with Optuna provides a robust framework for drought zone classification and offers valuable insights to support drought mitigation efforts on Flores Island.

Keywords: Classification, XGBoost, Optuna, Hyperparameter, Drought Zone.

1 Introduction

Decreased agricultural productivity and reduced water availability are often tangible consequences of drought, particularly in areas experiencing long-term low rainfall, such as Flores Island in Indonesia. According to Mokhtar et al. (2021), one of the triggers of climate change is human activity affecting the environment [1]. This view aligns with the findings of Ekmekcioğlu (2023) and Ali et al. (2023) [2, 3], who add that drought not only affects the agricultural sector but can also drive population migration. Considering these various impacts, classifying areas potentially vulnerable to drought appears to be a crucial step, allowing data-driven mitigation strategies and emergency response plans to be designed.

Machine learning-based drought prediction appears to be increasingly important in modeling research [1, 4] of the many available algorithms, Extreme Gradient Boosting

(XGBoost) is often the preferred choice, likely due to its ability to process tabular data and handle complex and layered datasets [5, 6]. In several studies predicting drought indices, XGBoost has produced highly precise results [2, 7]. However, this performance is not solely due to the superiority of the algorithm itself but is also heavily influenced by the accuracy of hyperparameter settings; without a thorough optimization process, these advantages can be diminished and the resulting predictions may be less than satisfactory.

That's why using automated hyperparameter optimization through an Automated Machine Learning (AutoML) framework seems like a reasonable path forward. In this case, the process relied on Optuna, a library built on Bayesian Optimization, intended to zero in on the most effective combination of hyperparameters during the modeling phase [8, 9, 10]. A number of studies suggest that Extreme Gradient Boosting (XGBoost) can reach remarkably high prediction accuracy when tuned with Optuna [8, 9]. What's interesting is that this XGBoost–Optuna pairing shows up well beyond the scope of this research it has been applied to forecasting wildfire risks and monitoring soil salinity, with results that seem both efficient and surprisingly accurate, particularly when dealing with tabular datasets [8, 10]

In this study, modeling variables were selected taking into account the unique context of Flores Island, which has a diverse climate that makes it difficult to predict. These variables include rainfall, air temperature, and humidity. These three indicators, when examined theoretically, can improve prediction accuracy compared to conventional approaches. Although numerous studies on drought prediction have been conducted, only a few have specifically addressed drought zone classification on Flores Island with a focus on hyperparameter optimization. This study contributes by integrating XGBoost and Optuna for drought zone classification using ten years of climate data, and demonstrates that hyperparameter optimization can significantly improve model accuracy. It is hoped that the application of the XGBoost algorithm combined with Optuna will not only provide reliable prediction results but also generate information sufficiently robust to serve as a basis for data-driven decision-making, particularly in drought risk mitigation efforts in the Flores Island region.

2 Literature Review

2.1 Classification

In research conducted by Noviandi and Sumitra (2018), it is explained that classification is a process of grouping data into certain classes based on predetermined features, and classification can also be used to predict future trends through the formation of models [11].

2.2 XGBoost

In various studies, XGBoost has proven superior to traditional methods. This algorithm has successfully produced accurate results for drought risk mapping, rainfall prediction,

and groundwater potential modeling [12]. However, several studies emphasize that excessively high accuracy requires caution, as it may not merely reflect the model's strength but rather indicate a strong dependence on specific data patterns.

2.3 XGBoost

Optuna, an automated optimization framework, has been shown to improve accuracy and efficiency compared to using default parameters [13]. Several recent studies in the climate and environment field have shown that Bayesian-based optimization methods can improve predictive performance while reducing computational costs [14].

3 Method

3.1 Data Source

This study used NASA POWER version 2.5.7 for data collection. Climate data was downloaded by location from each Regency on Flores Island. Each Regency had its own dataset downloaded in CSV format, and then combined into a single dataset.

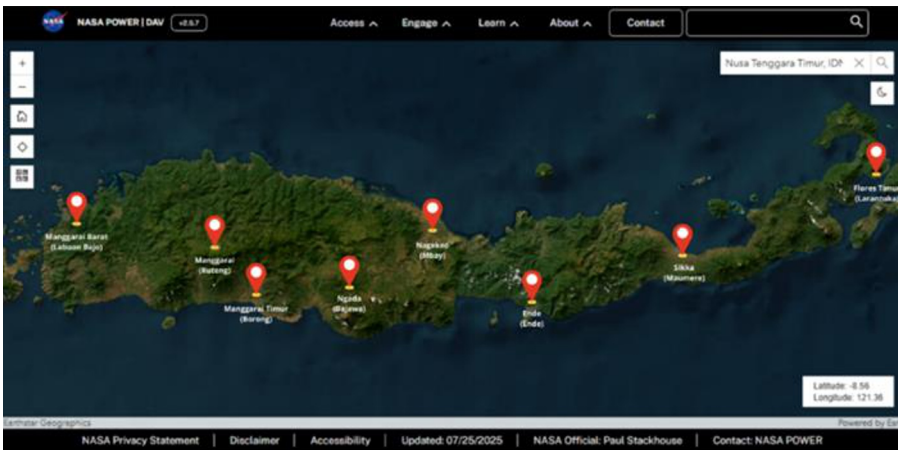


Fig. 1. Data collection location. source: Nasa power (<https://power.larc.nasa.gov/data-access-viewer>).

Fig. 1 shows the location of each regency on Flores Island. These include Manggarai Barat, Manggarai, Manggarai Timur, Ngada, Nagekeo, Ende, Sikka, and East Flores. These locations are determined by the city center of each regency.

3.2 Data Preprocessing

In research conducted by Supatmi et al. (2019), it was explained that the data pre-processing stage includes collecting primary variables. This was also done in this study,

such as determining rainfall, temperature, and humidity variables, which will then become the primary variables in the modeling stage using XGBoost [15].

3.2.1 Data Variables

Data was collected based on the geographic location of each Regency on Flores Island, Indonesia. The variables used were adjusted to meet the needs of the analysis phase. The following are the variables used: year of data recording (YEAR), daily data recording (DOY), precipitation (PRECTOTCORR), average temperature (T2M), minimum temperature (T2M_MIN), maximum temperature (T2M_MAX), and humidity (QV2M).

3.2.2 Adding Columns

Perform a spatial join by adding the Regencys as a variable. This ensures that each data item has Regencys attributes, which facilitate spatial analysis, aggregation of results by administrative region, and integration with external data.

3.2.3 Data Cleaning

To handle missing values and anomalies, data cleaning is performed. Missing values are addressed using linear interpolation methods or replaced with the mean value, while anomalies are checked and removed if proven to be measurement [16].

3.2.4 Data Aggregation

For drought zone analysis, daily data were combined into monthly data. Total rainfall (PRECTOTCORR) was calculated monthly, while average temperature, humidity, and wind speed were also calculated monthly. This is because drought index calculations are typically performed on a monthly time scale [16].

3.2.5 Normalization

The Min–Max Scale method converts data values into a range of 0 to 1, while the Z-Score method standardizes the data so that it has a mean of 0 and a standard deviation of 1. The application of these techniques ensures that the XGBoost model proportionally prioritizes variables with inherently larger numerical scales, thus maintaining fairness in the model learning process [17].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Equation (1) is applied to normalize data using the Min–Max method. In this equation, X represents the original value of a variable, X_{min} denotes the minimum value, and X_{max} denotes the maximum value in the dataset. The normalized result, indicated as X_1 , is rescaled to lie within the range of 0 to 1.

3.3 Determining Drought Zone Labels

Referring to research by Trabelsi et al. (2022), the zoning was divided into three main classes using the quantile approach [12]. This method is actually simple, but quite relevant because each class has its own specific role. Several previous studies have shown this method to be able to capture the complexity of data variations without over-classifying. The following is the defined zoning classification:

1. Dry Zone, Areas with rainfall values or drought indexes in the lower quantile (< 0.33).
2. Normal Zone, Areas with rainfall values in the middle quantile ($0.33-0.65$).
3. Wet Zone, Areas with rainfall values in the upper quantile (> 0.66).

The reason for choosing this method in this research is simple: it simplifies spatial mapping and maintains balance between classes.

3.4 XGBoost Model Implementation

After determining the drought zone label based on the DryIndex, the XGBoost model was used in this study. The following is a flowchart for the modeling stage using XGBoost (see Fig. 2):

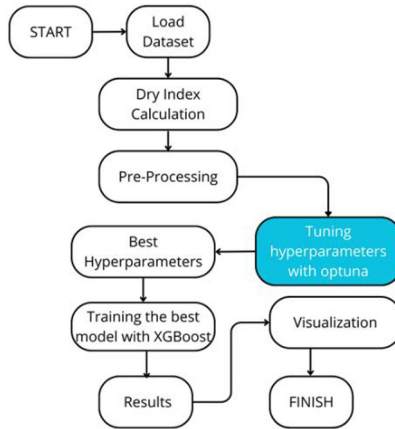


Fig. 2. XGBoost Algorithm Modeling Flowchart.

3.4.1 Specifying Dryindex

The Min-Max scaling method will be used to normalize all parameters selected during the modeling phase. This is crucial because it can impact the modeling results. After normalization, each parameter will be used to calculate the drought index. The following formula will be used:

$$DryIndex = T2M_MAX - PRECOTCORR - QV2M \quad (2)$$

Equation (2) shows how to calculate the DryIndex by utilizing three main climate variables that have been previously normalized. T2M_MAX represents the daily maximum temperature, PRECTOTCORR is the daily rainfall, and QV2M is the specific humidity at a height of two meters. From the calculation using the formula above, we will get the results of the drought index, from these results we can conclude that the higher the value obtained in each Regencys, the higher the possibility of the area experiencing drought, because this calculation is seen from rainfall, air humidity, and temperature of each regency.

3.4.2 Hyperparameter Tuning with Optuna

Before entering the prediction phase, XGBoost needs to be tuned to ensure optimal performance. The tuning phase uses Optuna to find the best parameter combination. The goal is simple: more accurate predictions. The results of this optimization will then be displayed in a diagram.

3.4.3 XGBoost Model Training

The labeled data will then be modeled using XGBoost. The variables used in the modeling phase are temperature, rainfall, and humidity. These three variables will be designated as target variables. In the modeling phase, the data will be divided into two parts: 20% for testing data and 80% for training data. This data will then be tested and trained using parameters from the tuning results using optuna.

3.4.4 Model Evaluation

To ensure classification performance in each class (Wet, Normal, and Dry), the trained models were tested using metrics such as accuracy, precision, recall, F1 score, and confusion matrix. A significant feature analysis was also performed to identify the climate parameters most influential in predicting drought zones.

3.5 Visualization of Results

To demonstrate the distribution of drought zone classifications based on XGBoost model predictions, the results are presented in a bar chart. The three drought zone categories are Dry Zone, NormalZone, and Wet Zone. This visualization facilitates a comparison of drought zones across regencies.

4 Results and Discussion

4.1 Dataset Preparations

This study uses NASA POWER daily climate data from 2015 to 2024. The collected data consists of 27,214 data points. The data is divided into a training set of 21,771 samples (80%) and a testing set of 5,443 samples (20%), where all model evaluation

results are reported based on the testing data. Several key climate parameters are analyzed, such as total rainfall (PRECTOT), average temperature (T2M), maximum temperature (T2M_MAX), minimum temperature (T2M_MIN), and specific humidity (QV2M). All data are collected based on the study area, which covers all regencies on Flores Island.

4.2 Pre-Processing

The datasets containing climate parameters (T2M_MAX, PRECTOTCORR, QV2M) were merged, cleaned of missing values, normalized using Min-Max Scaling, and then used to calculate the DryIndex (T2M_MAX_norm - PRECTOTCORR_norm - QV2M_norm). The results were classified into three zones (Wet, Normal, and Dry), and the *Drought_Zone* was saved as the label for XGBoost model training.

4.3 Implementing Optuna to determine hyperparameters

This optimization is performed using Optuna to obtain the best hyperparameters for the next stage. The optimization results are shown in Table 1.

Table 1 Hyperparameter Optimization Results using Optuna.

Hyperparameter	Score	Function
n_estimators	351	From the results obtained, it can be concluded that the model to be processed will be increasingly complex but still efficient.
max_depth	9	At the decision tree formation stage, based on the resulting values, it can be concluded that the complexity is very high. This is evident from the large number of nodes.
learning_rate	0.1962	The obtained value represents this hyperparameter, which will work accurately and stably. The smaller the value, the more precise the process.
subsample	0.7217	The results obtained show that the sample used was 72% and was selected randomly.
colsample_bytree	0.8899	The feature used is 88.99%, which also helps to prevent overfitting.
gamma	0.2055	The resulting value explains that there is control, but it is not strict. The aim is to prevent overfitting.
min_child_weight	4	This is a control that helps prevent the model from creating nodes that are too small. The results obtained can be said to be quite balanced.
reg_alpha (L1)	0.9180	This aims to ensure that the model does not use excessive features, where some features that are considered unimportant will be removed so that the structure becomes simpler but remains effective.
reg_lambda (L2)	0.6456	It aims to make the model work with more stable predictions.

The implementation results show that the parameters applied in the analysis phase are `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `gamma`, `min_child_weight`, `reg_alpha`, `reg_lambda`. These parameters achieved the best scores,

classifying drought zones with an accuracy rate of up to 99.3%. From these results, Optuna was able to find the best parameters to optimize the performance of the XGBoost algorithm.

4.4 XGBoost Model Evaluation

The evaluation results indicate that the XGBoost model optimized with Optuna performs very well in classifying the three drought zone categories: wet (0), normal (1), and dry (2), as shown in the Table 2.

Table 2. Matrix report.

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	1796
1	0.99	0.99	0.99	1796
2	1.00	1.00	1.00	1851
Accuracy	-	-	0.99	5433
Macro Avg	0.99	0.99	0.99	5433
Weighted Avg	0.99	0.99	0.99	5433

Judging from the results in Table 2, the overall accuracy is 99%. For the wet class, a precision of 0.99 and a recall of 0.99 indicate near-perfect prediction, with an F1-score of 0.99. For the normal class, the results are similar to those for the wet class, indicating balanced model performance across all classes, resulting in a very low probability of error. Finally, for the dry class, a precision of 100% indicates no false positives and a recall of 100% indicates that all data labeled as dry zones were successfully detected. Overall, it can be concluded that this model is capable of detecting categories with high accuracy.

4.5 Matrix Confusion Results

The confusion matrix was able to accurately predict samples in the dry class, but there were some errors in the analysis. This occurred in the normal class where in some data samples, the normal class was classified into the wet class several times. However, it can be concluded that the confusion matrix analysis results show a very low level of classification error because the results of the analysis show a balanced classification distribution. The following is a visualization of the confusion matrix from this study (see Fig. 3).

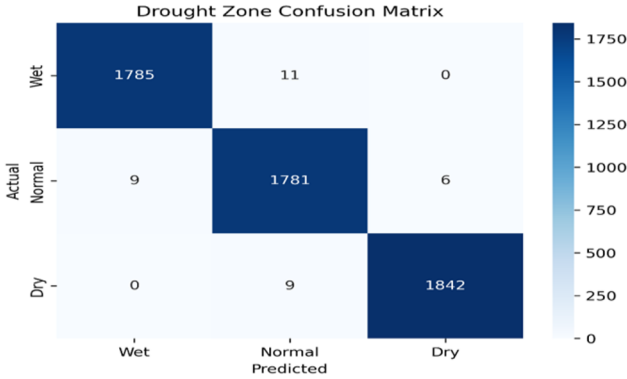


Fig. 3. Confusion Matrix Results.

4.6 Feature Importance Analysis

The analysis results show that the QV2M or specific humidity variable has the highest level of importance compared to other variables. This indicates that humidity levels have a significant role in determining the drought zone on Flores Island. Next, daily data recording (DOY) and maximum temperature (T2M_MAX) have a direct correlation because it can be concluded that when the temperature is high, the air humidity will be lower. This variable is very important for classification, because humidity and temperature can also affect water availability in the soil, while rainfall (PRECTOTCORR) and average temperature (T2M) also contribute significantly. T2M_MIN (minimum temperature) has a smaller influence, but is still relevant in helping the model detect extreme conditions (see Fig. 4).

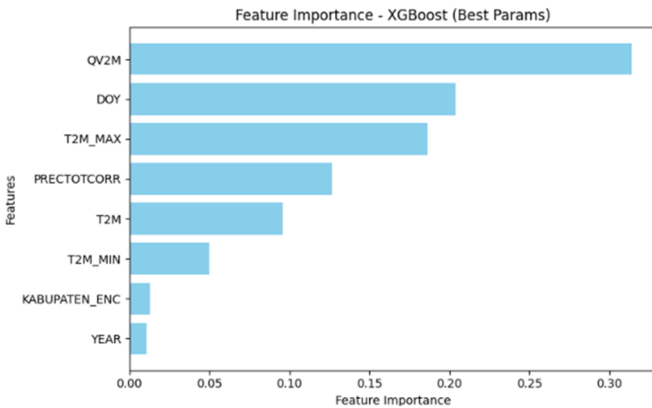


Fig. 4. Feature Importance Results.

4.7 Comparison of Default XGBoost and XGBoost Using Optuna

XGBoost with default parameters produced 98.0% accuracy, 0.96 precision, 0.98 recall, and 0.97 F1-score (see Table 3). After hyperparameter optimization using Optuna,

performance increased to 99.3% accuracy with precision, recall, and F1-score consistently at 0.99. This confirms that despite the robustness of the XGBoost baseline, hyperparameter tuning still provides significant performance improvements. These results align with Trabelsi et al. (2022) who reported the superiority of ensemble algorithms in environmental modeling, and Ekmekcioğlu (2023) who achieved >95% accuracy in XGBoost-based drought prediction [1, 12]. However, these findings are higher than those of Akiba, who reported 80–88% accuracy for environmental modeling. This difference can be explained by the selection of more relevant input variables and the application of hyperparameter optimization in this study.

Table 3. Comparison of default param and using optuna.

Model	Accuracy (%)	Precision	Recall	F1-Score
XGBoost (Default)	98	0.96	0.98	0.97
XGBoost+ Optuna (Tuned)	99.3	0.99	0.99	0.99

4.8 Visualization of Modeling Results

The results of this study indicate that each Regency differs in the amount of data included in the three categories: dry, normal, and wet. REGENCIES with the highest amount of data in the dry category can be considered more vulnerable to drought compared to other Regencys. Conversely, regencies with more data in the wet category indicate better water availability. The following are the prediction results obtained from the XGBoost model with optimized hyper parameters (see Table 4 and Fig. 5).

Table 4. Drought Ranking Table.

Rank	Regency	Number of Dry Zones
1	Sikka	1.477
2	Ende	1.304
3	Ngada	1.173
4	Nagekeo	1.172
5	Flores Timur	1.101
6	Manggarai Barat	1.027
7	Manggarai	1.027
8	Manggarai Timur	972

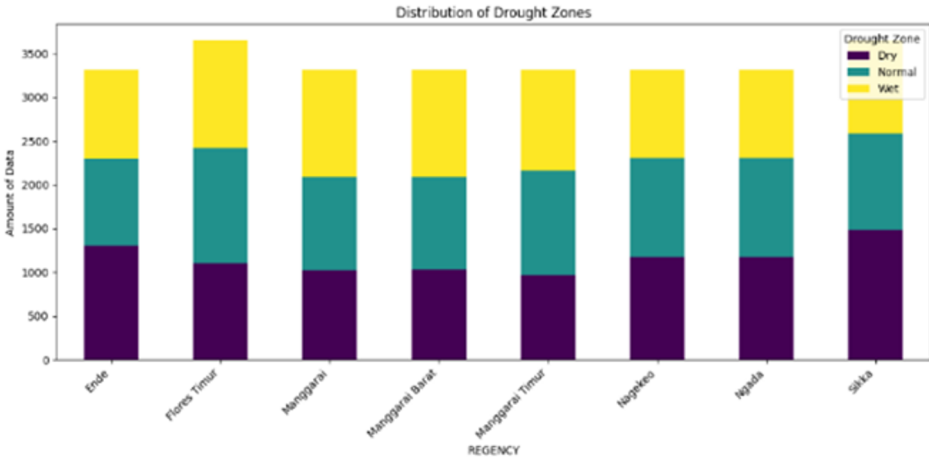


Fig. 5. Classification Results.

Furthermore, this visualization was used to identify the five Regencys with the highest levels of drought. Sikka has the largest number of dry zones, followed by Ende, Nagekeo, Ngada, and East Flores. The results indicate that drought risk reduction should be a top priority in these five regions.

5 Conclusion

This study demonstrates that applying Optuna for hyperparameter optimization in the XGBoost algorithm significantly improves model performance, achieving an accuracy of 99.3% with consistent precision, recall, and F1-scores across all classes. The findings highlight that relevant climate variables, particularly humidity, play a key role in classifying drought zones into three categories (wet, normal, dry). The main contribution of this study is the integration of XGBoost and Optuna, which proved effective for drought mapping in Flores Island. However, the study is limited to a single region and relies only on basic climate variables without incorporating additional factors such as vegetation indices or land use. As a next step, the model should be tested in other regions with the inclusion of more environmental variables to evaluate its generalizability and further enhance predictive accuracy.

References

1. Mokhtar, A., Jalali, M., He, H., Al-Ansari, N., Elbeltagi, A., Alsafadi, K., Rodrigo-Comino, J.: Estimation of SPEI meteorological drought using machine learning algorithms. *IEEE Access* **9**, 65503–65523 (2021).
2. Ekmekcioglu, Ö.: Drought Forecasting Using Integrated Variational Mode Decomposition and Extreme Gradient Boosting. *Water* **15**(19), 3413 (2023).

3. Ali, S., Khorrami, B., Jehanzaib, M., Tariq, A., Ajmal, M., Arshad, A., Shafeeque, M., Dilawar, A., Basit, I., Zhang, L., Sadri, S., Niaz, M.A., Jamil, A., Khan, S.N.: Spatial Downscaling of GRACE Data Based on XGBoost Model for Improved Understanding of Hydrological Droughts in the Indus Basin Irrigation System (IBIS). *Remote Sensing* **15**(4), 873 (2023).
4. Zhao, Y., Zhang, J., Bai, Y., Zhang, S., Yang, S., HENCHIRI, M., Seka, A.M., Nanzad, L.: Drought Monitoring and Performance Evaluation Based on Machine Learning Fusion of Multi-Source Remote Sensing Drought Factors. *Remote Sensing* **14**(24), 6398 (2022).
5. Zeng, F., Gao, Q., Wu, L., Rao, Z., Wang, Z., Zhang, X., Yao, F., Sun, J.: Modeling Short-Term Drought for SPEI in Mainland China Using the XGBoost Model. *Atmosphere* **16**(4), 419 (2025).
6. Zhang, B., Salem, F.K.A., Hayes, M.J., Tadesse, T.: Quantitative Assessment of Drought Impacts Using XGBoost based on the Drought Impact Reporter. arXiv preprint arXiv:2211.02768 (2022)
7. Duan, S., Zhang, X.: AutoML-based drought forecast with meteorological variables. arXiv preprint arXiv:2207.07012 (2022)
8. Liu, X., Hu, Y., Li, X., Du, R., Xiang, Y., Zhang, F.: An Interpretable Model for Salinity Inversion Assessment of the South Bank of the Yellow River Based on Optuna Hyperparameter Optimization and XGBoost. *Agronomy* **15**(1), 18-30 (2025).
9. Khan, M.S., Peng, T., Akhlaq, H., Khan, M.A.: Comparative analysis of automated machine learning for hyperparameter optimization and explainable artificial intelligence models. *IEEE Access* (2025)
10. Na, R., Gantumur, B., Du, W., Bayarsaikhan, S., Shan, Y., Mu, Q., Bao, Y., Tegshjargal, N., Vandansambuu, B.: Daily-Scale Fire Risk Assessment for Eastern Mongolian Grasslands by Integrating Multi-Source Remote Sensing and Machine Learning. *Fire* **8**(7), 273 (2025).
11. Noviani, I., Sumitra, I.D.: Classification consumer credit for missing value dataset. In: IOP Conference Series: Materials Science and Engineering, **407**(1), p. 012173 (2018).
12. Trabelsi, F., Bel Hadj Ali, S., Lee, S.: Comparison of novel hybrid and benchmark machine learning algorithms to predict groundwater potentiality: case of a drought-prone region of Medjerda Basin, northern Tunisia. *Remote Sensing* **15**(1), 152 (2022)
13. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631 (2019).
14. Janizadeh, S., Vafakhah, M., Kapelan, Z., Mobarghaee Dinan, N.: Hybrid XGBoost model with various Bayesian hyperparameter optimization algorithms for flood hazard susceptibility modeling. *Geocarto International* **37**(25), 8273–8292 (2022).
15. Supatmi, S., Hou, R., Sumitra, I.D.: Study of hybrid neurofuzzy inference system for forecasting flood event vulnerability in Indonesia. *Computational Intelligence and Neuroscience* **2019**(1), 6203510 (2019).
16. Phan, Q.T., Wu, Y.K., Phan, Q.D.: A hybrid wind power forecasting model with XGBoost, data preprocessing considering different NWP. *Applied Sciences* **11**(3), 1100 (2021).
17. Rusdah, D.A., Murfi, H.: XGBoost in handling missing values for life insurance risk prediction. *SN Applied Sciences* **2**(8), 1336 (2020).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

