Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3872~3883
e-ISSN: 2548-6861

Implementation of a Hybrid TabNet–XGBoost Model Based on Radiosonde Data for Predicting Daily Rainfall Intensity in Surabaya

Annabel Gracia Puryani[1]*, Aviolla Terza Damaliana[2], Alfan Rizaldy Pratama[3]***
* Universitas Pembangunan Nasional "Veteran" Jawa Timur
22083010048@student.upnjatim.ac.id [1], aviolla.terza.sada@upnjatim.ac.id [2], alfan.fasilkom@upnjatim.ac.id [3]

Article Info

Article history:

Received 2026-06-29
Revised 2026-08-03
Accepted 2026-08-12

Keyword:

Hybrid TabNet–XGBoost,
Machine Learning,
Rainfall Prediction,
Radiosonde

ABSTRACT

Rainfall prediction plays an important role in supporting hydrometeorological disaster mitigation and weather-related decision-making. However, accurate rainfall prediction remains challenging because atmospheric processes are highly nonlinear and governed by complex interactions among multiple meteorological variables. This study proposes a Hybrid TabNet–XGBoost model for daily rainfall prediction using integrated radiosonde and surface meteorological observations collected at the BMKG Juanda Class I Meteorological Station. The dataset covers the period from 2019 to 2025 and consists of 2,551 daily observations. TabNet was employed to select the fifteen most informative atmospheric variables based on feature importance, while temporal dependencies were incorporated through lag features generated using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analyses. Hyperparameter optimization was performed using Optuna with TimeSeriesSplit cross-validation prior to model training. Experimental results on the testing dataset achieved an RMSE of 19.1347 mm, an MAE of 11.9742 mm, a MAPE of 17.62%, and an R^2 of 0.0449. The proposed model was able to capture the general temporal pattern of daily rainfall and produced satisfactory predictions under the dominant rainfall conditions represented in the dataset. However, the model exhibited reduced sensitivity to high-intensity rainfall events, resulting in the underestimation of extreme rainfall and a relatively low R^2 value, primarily due to the imbalanced rainfall distribution and the complexity of rainfall processes. The optimized model was subsequently applied to generate daily rainfall projections for 2026 based on historical atmospheric observations. Since the corresponding observational data were unavailable at the time of this study, these projections should be interpreted as model-based forecasts rather than validated prediction results. Overall, the proposed Hybrid TabNet–XGBoost framework demonstrates the potential of integrating radiosonde and surface meteorological observations for daily rainfall prediction while highlighting the need for additional atmospheric and spatial information to improve the prediction of extreme rainfall events.

I. INTRODUCTION

Global climate change has increased the frequency and intensity of extreme weather events in various regions around the world [1]. The Intergovernmental Panel on Climate Change (IPCC) AR6 report states that the rise in global average temperature of approximately 1.1°C since the pre-industrial era has caused the atmosphere to hold more water vapor, thereby increasing the likelihood of heavy rainfall [1]. The increase in extreme rainfall events has far-reaching impacts on various sectors, ranging from the environment and infrastructure to the social and economic activities of communities [2]. These conditions underscore the need to develop more accurate rainfall prediction systems as part of efforts to mitigate hydrometeorological disasters [3].

Indonesia is a tropical country with a complex atmospheric profile resulting from the interaction of monsoon winds, the Madden–Julian Oscillation (MJO), air mass convergence, and topographic influences [4]. The interaction of these various atmospheric phenomena causes the distribution and intensity of rainfall in Indonesia to exhibit high variability throughout the year [4]. In addition to being influenced by regional and global factors, rainfall formation in Indonesia is also dominated by convective processes involving temperature, humidity, and atmospheric dynamics across various layers of the atmosphere [5]. These dynamic atmospheric characteristics mean that predicting rainfall intensity in tropical regions remains a challenge in the field of meteorology [6].

This phenomenon has also been observed in the city of Surabaya, which in recent years has experienced an increase in heavy rainfall events with daily rainfall exceeding 150 mm [7]. These extreme rainfall events have caused flooding in various areas, resulting in infrastructure damage and disruptions to community activities [7]. The Surabaya City Government has also noted an increase in the budget required for flood management as a result of the rise in hydrometeorological events [7]. These conditions indicate that improving the ability to predict rainfall intensity is essential for supporting decision-making in disaster mitigation [3].

Predicting rainfall intensity remains a challenge because the process by which it forms is influenced by the complex and nonlinear interactions of various atmospheric parameters [8]. The use of global reanalysis data such as ERA5 still has limitations in representing tropical atmospheric conditions, particularly with regard to humidity and atmospheric instability [9]. Therefore, radiosonde data is a relevant source of observations because it provides direct, high-resolution information on the vertical structure of the atmosphere [10]. In addition to measuring temperature, pressure, humidity, and wind, radiosonde data can also be used to calculate various atmospheric indices related to the potential for extreme rainfall [11].

Various studies have utilized atmospheric indices derived from radiosonde observations to support rainfall forecasts. The K-Index has been reported to be capable of identifying most extreme rainfall events in Southeast Asia and is therefore widely used as an indicator of atmospheric instability [12]. In addition, the Lifted Index (LI) and Convective Available Potential Energy (CAPE) are also known to have a strong correlation with the development of convective clouds and heavy rainfall events [13]. These findings indicate that atmospheric parameters derived from radiosonde observations have the potential to be used as inputs in the development of rainfall intensity prediction models [14].

As technology has advanced, various machine learning methods have been applied to model nonlinear relationships in multivariate atmospheric data [15]. TabNet is one such method designed for tabular data that uses a sequential attention mechanism, enabling it to perform adaptive feature selection [16]. On the other hand, XGBoost is widely used as a gradient boosting algorithm because it is capable of building predictive models for data with complex characteristics [17]. However, the use of TabNet as a standalone model still faces challenges in handling imbalanced data distributions and relatively rare extreme weather events [18].

Various machine learning and deep learning approaches, including Random Forest, Support Vector Regression (SVR), Light Gradient Boosting Machine (LightGBM), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), have been widely applied for rainfall prediction because of their ability to model complex nonlinear relationships among meteorological variables [19]–[21]. Although these methods have demonstrated promising predictive performance, most previous studies primarily rely on surface meteorological observations or gridded reanalysis datasets, which may not adequately represent the vertical atmospheric structure associated with convective rainfall formation [22]. Furthermore, many existing studies directly utilize all available predictor variables without adaptive feature selection, potentially introducing redundant information and increasing model complexity [23]. Consequently, the integration of radiosonde-derived atmospheric variables with adaptive feature selection remains relatively limited in rainfall prediction research. Therefore, combining TabNet-based feature selection with XGBoost provides a promising framework for reducing feature redundancy while improving rainfall prediction using vertically observed atmospheric information [24].

Given these conditions, the implementation of the Hybrid TabNet–XGBoost approach for rainfall intensity prediction based on radiosonde data remains an interesting topic for further study [24]. In this study, TabNet was used as an adaptive feature selection method to identify the most relevant atmospheric parameters, while the selected features were used as input for the XGBoost model to predict rainfall intensity [24]. This approach combines TabNet's adaptive feature selection process with XGBoost's ability to build predictive models based on nonlinear relationships in atmospheric data [24]. This study aims to examine the implementation of the Hybrid TabNet–XGBoost model in rainfall intensity prediction based on radiosonde data and to evaluate its performance using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) [25].

II. METHOD

This study proposes a Hybrid TabNet–XGBoost model to predict rainfall intensity using radiosonde data. The research stages include data acquisition, preprocessing, TabNet feature selection, Lag Feature Construction Using ACF and PACF, data splitting, lag feature engineering, hyperparameter optimization using Optuna, training the Hybrid TabNet–XGBoost model, evaluating the model's performance using

the RMSE, MAE, and MAPE metrics, and predicton. The research flowchart is shown in Figure 1.

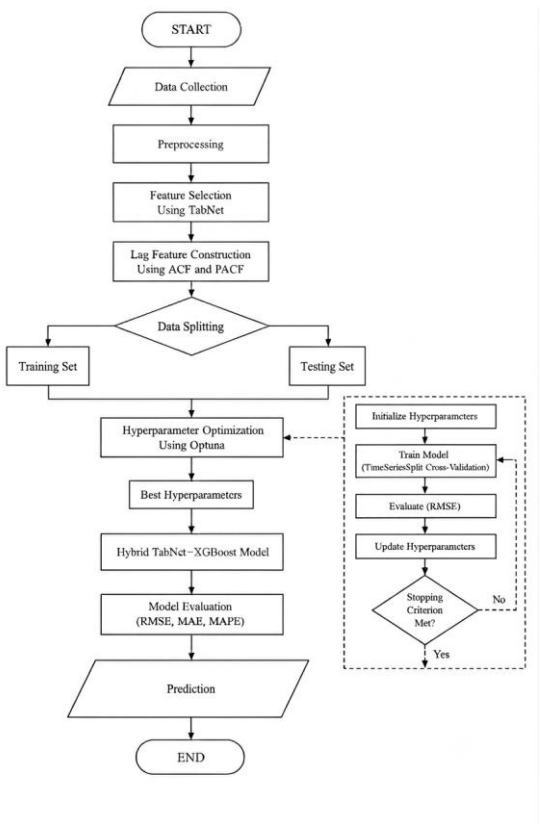

Figure 1 Flowchart Rainfall Prediction

A. Data Collection

The dataset used in this study consists of atmospheric observation data obtained from the BMKG Juanda Class I Meteorological Station. The data comprise upper-air observations from radiosondes, surface meteorological observations, and daily precipitation records covering the period 2019–2025. Radiosonde observations were conducted routinely twice daily at 00 UTC and 12 UTC following the operational procedures of the World Meteorological Organization (WMO). The upper-air observations were subsequently integrated with surface meteorological observations and daily rainfall records to produce a unified dataset for model development. The surface observation variables include station height (Height), surface air pressure (Press0), surface air temperature (Temp0), surface relative humidity (Humi0), wind speed (WS), and wind direction (WD). Meanwhile, the upper-air variables consist of various atmospheric stability indices and thermodynamic parameters, including the Lifted Index (LI), K Index (KI), Showalter Index (SI), Total Totals (TT), Total Precipitable Water (TPW), Convective Available Potential Energy (CAPE), Convective Inhibition (CIN), Boyden Index, and other atmospheric parameters derived from radiosonde observations. The target variable in this study is daily precipitation (CH), while all atmospheric variables are used as predictor variables to develop a Hybrid TabNet–XGBoost model for rainfall prediction.

TABLE 1
DATA STRUCTURES

Tanggal	LI	KI	.	.	Humi0	WS	WD
01/01/19	$x_{1,1}$	$x_{2,1}$	.	.	$x_{21,1}$	$x_{22,1}$	$x_{23,1}$
02/01/19	$x_{1,2}$	$x_{2,2}$	.	.	$x_{10,2}$	$x_{11,2}$	$x_{12,2}$
.	.	.			.	.	.
.	.	.	.	.	.	.	.
30/12/25	$x_{1,2558}$	$x_{2,2558}$	.	.	$x_{21,2558}$	$x_{22,2558}$	$x_{23,2558}$
31/12/25	$x_{1,2558}$	$x_{2,2558}$	.	.	$x_{21,2558}$	$x_{22,2558}$	$x_{23,2558}$

B. Pre-Processing

The preprocessing stage is performed to ensure the quality and consistency of the dataset before it is used in the model development process [26]. This stage includes data cleaning, data type transformation, and missing value handling to produce reliable input data and improve the stability of the prediction model [26]. During the data cleaning process, inconsistencies in data formatting, such as numerical notation and variable naming, are standardized, while all predictor variables are converted into numerical data types compatible with machine learning algorithms [26]. In addition, the date column is transformed into the datetime format to preserve the chronological order of observations, which is essential for time-series modelling [27].

Missing values are handled using the linear interpolation method, which estimates missing observations based on adjacent available values [28]. This method is widely adopted for time-series data because it preserves temporal continuity while minimizing distortion of the original data pattern [28]. Prior to interpolation, the dataset was inspected to identify incomplete observations and ensure the consistency of meteorological measurements. Data cleaning was also performed to standardize variable names, numerical formats, and data types before model development. Furthermore, linear interpolation maintains the number of observations available for model training without introducing abrupt changes in the temporal sequence [28]. As a result, the preprocessing stage produces a complete and consistent dataset that is suitable for subsequent feature engineering and model development [26].

C. Feature Selection Using TabNet

During the training process, TabNet generates an attention mask at each decision step to determine which features should receive greater attention in the subsequent

computation [29]. The attention mask is calculated using the Sparsemax activation function, as expressed in Equation 1.

$$M[i] = \text{sparsemax}\,(h_i(a_i - 1) \cdot P[i]) \quad \text{Equation 1}$$

Where M[i] denotes the attention mask, $h_i(a_i - 1)$ is the transformed activation obtained from the previous decision step, P[i] represents the prior scales that assign the initial weights to each feature, and Sparsemax is a normalization function that produces a sparse weight distribution by assigning zero weights to less relevant features [29].

The selected features are then processed by the Feature Transformer to generate a new activation a_i which represents a nonlinear combination of the selected features [29]. Before producing the activation, the Rectified Linear Unit (ReLU) activation function is applied to preserve positive values while setting negative values to zero, thereby maintaining nonlinear representations and improving training stability [29]. Subsequently, the prior scales are updated to reduce the probability of repeatedly selecting the same features in subsequent decision steps, as expressed in Equation 2.

$$P[i + 1] = P[i] \odot (1 - M[i]) \quad \text{Equation 2}$$

where P[i+1] denotes the updated prior scales for the next decision step, P[i] is the prior scale at the current decision step, and 1−M[i] reduces the weights of previously selected features, encouraging the model to explore other informative features [29]. Finally, the activations obtained from all decision steps are aggregated to produce the final feature representation, as expressed in Equation 3.

$$a_{final} = \sum_{i=1}^{N} a_i \quad \text{Equation 3}$$

where a_{final} denotes the final feature representation, a_i is the activation generated at the i-th decision step, and N is the total number of decision steps in the TabNet architecture [30]. The aggregated representation is subsequently used to calculate the feature importance score of each input variable, and the variables are ranked according to their importance scores [30]. The variables with the highest feature importance values are then selected as the input features for the subsequent prediction model, reducing data dimensionality while preserving the most informative features [30].

In this study, all predictor variables were initially provided as input to the TabNet model to learn their relative contributions to rainfall prediction through its sequential attention mechanism. Based on the resulting feature importance scores, the predictor variables were ranked according to their relevance to the prediction task. The variables with the highest importance scores were subsequently retained for the next stage of lag feature construction and Hybrid TabNet–XGBoost model development. This approach reduces feature redundancy while preserving the most informative atmospheric variables, thereby improving model efficiency without sacrificing predictive capability [25], [26].

D. Lag Feature Construction Using ACF and PACF

Lag feature generation is performed to incorporate temporal information by representing the relationship between current observations and their historical values [31]. This process utilizes the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify the temporal dependency of each variable and determine the most appropriate lag based on its autocorrelation characteristics [31]. The selected lag values are then used to construct lagged features, enabling the model to capture historical patterns and temporal dependencies within the dataset [31]. The generated lag features are subsequently used as input variables for the prediction model in the next stage [32]. Based on the ACF and PACF analyses, lag features were constructed to incorporate temporal dependencies between previous atmospheric conditions and current rainfall observations. The lag selection process was performed individually for each predictor variable by considering the autocorrelation pattern and statistical significance obtained from the ACF and PACF analyses. The selected lag features were subsequently combined with the predictor variables used as input to the Hybrid TabNet–XGBoost model. The detailed lag values selected for each variable are presented and discussed in the Results and Discussion section.

E. Data Splitting

The dataset is divided into training and testing sets using a chronological split approach, where the observations are separated according to their temporal order without randomization [33]. This approach preserves the sequential nature of time-series data, ensuring that the model is trained only on historical observations before being evaluated on future data [33]. In this study, observations from January 2019 to December 2024 were used as the training set, while observations from January 2025 to December 2025 were reserved as the testing set [33]. Such a data partitioning strategy provides a more realistic evaluation of the model while preventing data leakage that may occur if future observations are included during the training process [33]. Since rainfall data are time-series observations, the dataset was divided chronologically to preserve the temporal order of the observations and prevent data leakage between the training and testing sets. This approach ensures that the prediction model is always evaluated using future observations that were not available during training, thereby providing a more realistic assessment of model performance. The detailed composition of the training and testing datasets is presented in the Results and Discussion section.

F. Hyperparameter Optimization Using Optuna

Hyperparameter optimization was performed using Optuna, a define-by-run optimization framework that automatically searches for the optimal hyperparameter combination based on a predefined objective function [34]. To preserve the temporal structure of the dataset, the

optimization process employed TimeSeriesSplit as the cross-validation strategy, while the Root Mean Square Error (RMSE) was used as the optimization objective [34].

Hyperparameter optimization was performed using the Optuna framework to identify the optimal configuration of the Hybrid TabNet–XGBoost model. Optuna automatically explored combinations of hyperparameter values within predefined search spaces using a trial-based optimization strategy. The optimization objective was to minimize the prediction error while maintaining the model's generalization capability through TimeSeriesSplit cross-validation. This automated optimization approach reduces the need for manual parameter tuning and improves the efficiency of the model development process [34].

TABLE 2
SELECTED SEARCH SPACES

Hyperparameter	Search Space
Booster	{**gbtree**, **dart**}
Number of Trees (n_estimators)	250–400 (step = 25)
Maximum Tree Depth (max_depth)	2–4
Learning Rate (learning_rate)	0.005–0.015 (log scale)
Subsample (subsample)	0.85–1.00
Column Sampling (colsample_bytree)	0.75–0.95
Minimum Child Weight (min_child_weight)	4–8
Gamma (gamma)	0.05–0.20
L1 Regularization (reg_alpha)	0.70–1.20
L2 Regularization (reg_lambda)	0.80–2.50
Sample Type	{uniform, weighted}
Normalize Type	{tree, forest}
Rate Drop	0.00–0.30
Skip Drop	0.00–0.30

The selected search spaces were designed to balance model complexity and generalization performance. The optimization focused on controlling tree depth, learning rate, sampling strategy, and regularization parameters to reduce overfitting while maintaining predictive accuracy. Additional hyperparameters, including sample_type, normalize_type, rate_drop, and skip_drop, were optimized only when the DART booster was selected to exploit its dropout mechanism for improving model robustness.

During the optimization process, each trial represented a unique combination of hyperparameter values generated by Optuna. For every trial, the XGBoost model was trained and evaluated using TimeSeriesSplit cross-validation, and the average RMSE across all validation folds was computed as the objective value. The hyperparameter combination producing the lowest average RMSE was selected as the optimal configuration for the final Hybrid TabNet–XGBoost model, as expressed in Equation 4.

$$\theta^* = \arg\min_{\theta} \frac{1}{k} \sum_{i=1}^{K} RMSE_i \qquad \text{Equation 4}$$

where θ^* denotes the optimal hyperparameter combination, θ represents the candidate hyperparameter set evaluated in each trial, $RMSE_i$ is the RMSE obtained from the i-th cross-validation fold, and K is the total number of folds [34]. In this study, the optimization was conducted over 50 trials, and the best hyperparameter combination was selected for model training [34].

The hyperparameter optimization procedure was implemented using the *Optuna* framework. During each optimization trial, a new set of *XGBoost* hyperparameters was generated and evaluated using TimeSeriesSplit with RMSE as the objective function.

G. Hybrid TabNet-XGBoost

The Hybrid TabNet–XGBoost model integrates the feature selection capability of TabNet with the predictive capability of XGBoost within a unified framework [35]. In this approach, TabNet is employed to identify and learn the most informative feature representations through a sequential attention mechanism, while XGBoost utilizes the selected feature representation to construct the prediction model using the gradient boosting algorithm [35]. This integration enables the model to reduce input dimensionality while preserving the most relevant information for the subsequent learning process [35].

Unlike conventional machine learning approaches that directly utilize all input variables, the proposed hybrid framework first performs an adaptive feature learning process to determine which variables contribute the most to the prediction task. Through multiple decision steps, TabNet progressively learns the relationships among input variables and produces a compact feature representation that retains the most informative characteristics of the original data. This representation serves as a refined input for the subsequent prediction model, allowing redundant or less informative features to have a smaller influence on the learning process [35].

The generated feature representation is then transferred directly to the XGBoost model, where the prediction process is performed through an ensemble of gradient-boosted decision trees. During training, each tree iteratively improves the prediction produced by the previous trees, enabling the model to capture nonlinear relationships among the selected features while reducing prediction errors. The combination of adaptive feature learning and gradient boosting allows the hybrid framework to improve learning efficiency while maintaining strong predictive performance for complex tabular datasets [35]. The overall flowchart of the proposed Hybrid TabNet–XGBoost model is illustrated in Figure 2.

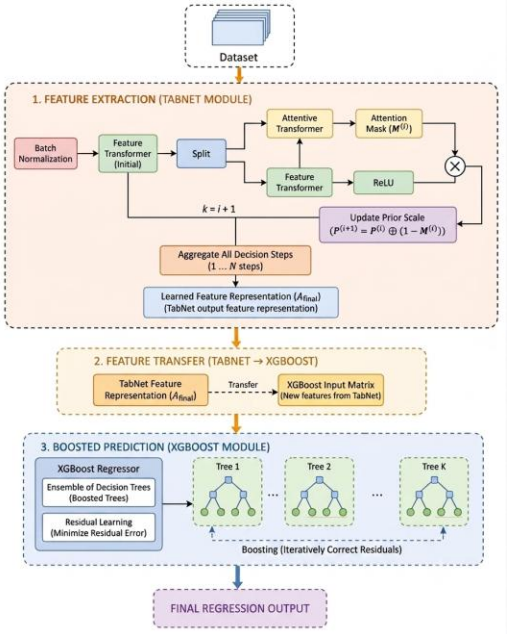

Figure 2 Flowchart Hybrid Tabnet-XGBoost

As described in Equation 3, TabNet generates the final feature representation, a_{final}, by aggregating the activations obtained from all decision steps [35]. The resulting feature representation is subsequently transferred as the input feature matrix for the XGBoost model without additional transformation [35]. XGBoost then constructs an ensemble of decision trees iteratively, where each tree is trained to minimize the prediction error while improving the performance of the previously generated trees [35]. Through this boosting mechanism, the prediction model is progressively refined by incorporating information learned from earlier iterations [35]. The learning process of XGBoost is formulated as an optimization problem that minimizes an objective function consisting of a loss function and a regularization term, as expressed in Equation 5.

$$L(\theta) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{K=1}^{K} \Omega(f_k) \qquad \text{Equation 5}$$

Where L(θ) denotes the objective function of the model, $l(\hat{y}_i, y_i)$ represents the loss function that measures the difference between the predicted and observed values, $\Omega(f_k)$ is the regularization function for the k-th decision tree, and K denotes the total number of decision trees in the ensemble [36]. The regularization term is defined as Equation 6, which controls model complexity and reduces the risk of overfitting during the learning process.

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w^2{}_j \qquad \text{Equation 6}$$

Where T denotes the number of leaf nodes in a decision tree, w_j represents the weight assigned to the j-th leaf node, γ is the penalty parameter controlling the number of leaf nodes, and λ is the L2 regularization parameter that penalizes large leaf weights [36]. By utilizing the feature representation generated by TabNet together with the objective and regularization functions of XGBoost, the hybrid framework performs feature learning and prediction sequentially within a single modeling pipeline [36].

After obtaining the optimal hyperparameters, the selected lag features were used to train the Hybrid TabNet–XGBoost model.

H. Model Evaluation

The trained model was evaluated using an independent testing dataset to assess its prediction performance [37]. The evaluation was conducted by comparing the predicted rainfall values with the corresponding observed values using four performance metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These metrics provide complementary information regarding prediction accuracy, where RMSE and MAE measure the magnitude of prediction errors, MAPE quantifies the relative percentage error, and R^2 evaluates the proportion of variance in the observed rainfall explained by the model [37]. Lower values of RMSE, MAE, and MAPE indicate better predictive performance, whereas an R^2 value closer to one indicates stronger agreement between the predicted and observed rainfall. Since rainfall observations frequently contain zero or near-zero values, MAPE was interpreted with caution because it is sensitive to small actual values. The Root Mean Square Error (RMSE) measures the average magnitude of prediction errors while assigning greater penalties to larger errors, as expressed in Equation 7.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{Equation 7}$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, and n represents the number of test data points. The smaller the RMSE value, the better the model's predictive ability [39]. The Mean Absolute Error (MAE) measures the average absolute difference between the predicted and observed values, as expressed in Equation 8.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{Equation 8}$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, and n represents the number of test data points. A smaller MAE value indicates that the predicted results are closer to the observed values[39]. The Mean Absolute Percentage Error (MAPE) measures the average relative prediction error expressed as a percentage of the observed values, as expressed in Equation 9.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad \text{Equation 9}$$

Where, y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n represents the number of test data points. A smaller MAPE value indicates that the prediction results have a lower relative error compared to the actual value [39]. he coefficient of determination (R^2) measures the proportion of variance in the observed rainfall explained by the prediction model, as expressed in Equation 10.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{Equation 10}$$

Where y_i represents the observed rainfall, $\hat{y}_i$ denotes the predicted rainfall, $\bar{y}$ is the mean of the observed rainfall, and n is the total number of observations. The numerator represents the residual sum of squares (RSS), while the denominator represents the total sum of squares (TSS). Consequently, R^2 quantifies the proportion of the variability in the observed rainfall that is explained by the prediction model. An R^2 value closer to one indicates better agreement between the predicted and observed rainfall, whereas values closer to zero indicate limited explanatory capability [41]. The evaluation results obtained from these four metrics were used to comprehensively assess and compare the predictive performance of the proposed model.

I. Prediction

The trained model was subsequently used to generate predictions on the testing dataset using the optimized model parameters [41]. During the prediction stage, each testing observation was processed to produce the corresponding output based on the patterns learned during the training process [41]. The generated predictions were then compared with the corresponding observed values to assess the model performance. This procedure enables the evaluation of the model's ability to generalize to previously unseen data [41].

The prediction performance was evaluated using the quantitative metrics described in the previous subsection [41]. In addition, the prediction results were visualized by comparing the predicted and observed values over the testing period to facilitate performance interpretation [41]. This visualization provides an overview of the agreement between the predicted and observed values while highlighting potential deviations throughout the prediction period [41]. The combination of quantitative evaluation and graphical visualization provides a comprehensive assessment of the model performance [41].

III. Result and Discussion

A. Dataset

The dataset used in this study consists of atmospheric observation data obtained from radiosonde observations integrated with surface meteorological observations collected at the BMKG Juanda Class I Meteorological Station. The observation period covers January 2019 to December 2025, representing variations in atmospheric conditions across different seasons. After the preprocessing stage, a total of 2,551 observation records were obtained and prepared for the model development process.

The dataset consists of upper-air atmospheric parameters, atmospheric stability indices, and surface meteorological variables. The upper-air variables include 850 hPa, 700 hPa, 500 hPa, CAPE, CIN, KO Index, Lifted Index (LI), K Index (KI), Total Totals (TT), Showalter Index (SI), SWEAT Index, Lifting Condensation Level (LCL), Convective Condensation Level (CCL), Level of Free Convection (LFC), Total Precipitable Water (TPW), Maximum Vertical Velocity (MVV), and Boyden Index. In addition, surface meteorological observations consisting of Height, Press0, Temp0, Humi0, WS, and WD were incorporated into the dataset to enrich the atmospheric information used during model development. Daily precipitation (CH) was used as the target variable. A summary of the dataset used in this study is presented in Table 3.

TABLE 3
DATASET

Tanggal	LI	KI	.	.	Humi0	WS	WD
01/01/19	-2.50	18.25	.	.	55.32	10.86	168.55
02/01/19	-2.50	36.55	.	.	59.16	9.11	136.16
.	.	.			.	.	.
.	.	.	.	.	.	.	.
30/12/25	-3.00	30.55	.	.	22.38	11.41	111.43
31/12/25	-4.00	36.40	.	.	26.68	13.37	89.94

To preserve the temporal characteristics of the rainfall time series, the dataset was partitioned chronologically into training and testing sets rather than using random sampling. The observations from January 2019 to November 2025 were used for model training, while the observations from December 2025 were reserved exclusively for model testing. This chronological partitioning prevents information leakage between the training and testing datasets and provides a more realistic evaluation of the model's predictive capability for future rainfall events. The rainfall distribution during the observation period is illustrated in Figure 3.

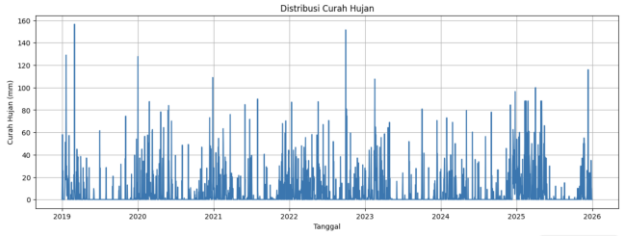

Figure 3 Rainfall Distribution

Most observations correspond to low or zero rainfall, whereas high-intensity rainfall events account for only a small proportion of the dataset. This distribution indicates

that the dataset is imbalanced, with the majority of observations representing non-rainfall or low-rainfall conditions. Such data imbalance increases the difficulty of rainfall prediction because the model is primarily exposed to low-intensity events during training, while extreme rainfall events occur relatively infrequently. Consequently, the proposed Hybrid TabNet–XGBoost model was expected to learn the dominant rainfall patterns while maintaining its ability to identify high-intensity rainfall events despite their limited representation in the dataset.

B. Pre-Processing

The preprocessing stage was carried out to improve data quality before model development. This stage consisted of data cleaning, data type conversion, chronological ordering of observations, and missing value handling using linear interpolation. In addition, radiosonde observations were integrated with surface meteorological observations to produce a unified dataset suitable for subsequent analysis. After preprocessing, all variables were successfully transformed into numerical format and no missing values remained in the dataset. The resulting dataset preserves the chronological order of observations and provides consistent input for the feature selection and model development stages. These preprocessing steps ensure that both upper-air and surface meteorological information can be utilized simultaneously throughout the modeling process.

The preprocessing results indicate that the integrated dataset was successfully standardized and prepared for subsequent analysis. Data cleaning and interpolation effectively eliminated incomplete observations while preserving the temporal continuity of the rainfall time series. Consequently, the final dataset contained complete numerical records without missing values, enabling reliable feature selection using TabNet and subsequent rainfall prediction using the Hybrid TabNet–XGBoost model. These improvements ensure that the modeling process is based on consistent and high-quality meteorological observations.

C. Feature Selection Using TabNet

Following preprocessing, feature selection was performed using the TabNet model to identify the variables that contributed most significantly to the learning process. The TabNet model was trained using all atmospheric variables as input features and precipitation (CH) as the target variable. The resulting feature importance values indicate the contribution of each variable to the learning process, where higher feature importance values correspond to greater influence on model learning. To reduce feature redundancy while preserving the most informative atmospheric variables, the fifteen predictors with the highest feature importance values were retained for subsequent lag feature construction and Hybrid TabNet–XGBoost model development. The feature importance results are presented in Table 4.

TABLE 4
FEATURE IMPORTANCE

No.	Variable	Feature Importance
1	LCL	0.197996
2	500	0.090762
3	WS	0.076466
4	KO	0.073724
5	LI	0.073455
6	Humi0	0.071792
7	850	0.069097
8	700	0.063985
9	TPW	0.056000
10	Press0	0.038424
11	WD	0.037330
12	TT	0.035246
13	SI	0.027897
14	MVV	0.016189
15	SWEAT	0.015742

Based on Table 4, LCL achieved the highest feature importance value (0.197996), followed by the 500 hPa pressure level (0.090762), wind speed (WS) (0.076466), KO Index (0.073724), and Lifted Index (LI) (0.073455). The dominance of LCL indicates that the lifting condensation level contributes substantially to the feature representation learned by TabNet. Likewise, the relatively high importance values obtained by the atmospheric stability indices and upper-air parameters indicate that these variables provide important information for capturing atmospheric conditions associated with rainfall formation.

The initial dataset consisted of 23 atmospheric predictor variables, including upper-air atmospheric parameters, atmospheric stability indices, and surface meteorological observations. Based on the feature importance ranking generated by TabNet, the 15 highest-ranked variables were selected because they provided the greatest contribution to the prediction task, whereas the remaining variables exhibited relatively low importance values and were excluded to reduce feature redundancy and computational complexity. This selection strategy enables the subsequent prediction model to focus on the most informative atmospheric variables while minimizing the influence of less relevant predictors.

The selected variables are consistent with the physical processes governing rainfall formation. For instance, LCL represents the altitude at which an air parcel becomes saturated and cloud formation begins, making it highly relevant to precipitation development. Similarly, atmospheric stability indices such as the KO Index, Lifted Index (LI), and Total Totals (TT), together with upper-air parameters at the 850, 700, and 500 hPa pressure levels, characterize atmospheric instability and moisture distribution that strongly influence convective rainfall. The inclusion of surface variables, including wind speed (WS), relative humidity (Humi0), surface pressure (Press0), and wind direction (WD), further indicates that integrating upper-air and surface

observations provides complementary information for rainfall prediction.

D. Lag Feature Construction Using ACF and PACF

After obtaining the fifteen variables with the highest feature importance values, lag feature generation was performed using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analyses. This analysis aims to determine the optimal lag of each variable according to its temporal characteristics so that historical information can be incorporated into the prediction process. The selected lag values are presented in Table 5.

TABLE 5
SELECTED LAG

No.	Variable	Lag
1	LCL	1
2	500	1
3	WS	1
4	KO	1
5	LI	3
6	Humi0	1
7	850	1
8	700	1
9	TPW	1
10	Press0	1
11	WD	2
12	TT	2
13	SI	2
14	MVV	7
15	SWEAT	4
16	CH	2

Based on Table 5, most variables exhibit the strongest temporal relationship at lag 1, including LCL, 500, WS, KO, Humi0, 850, 700, TPW, and Press0. These results indicate that atmospheric conditions observed on the previous day provide the most relevant information for predicting current rainfall. In contrast, several variables exhibit stronger temporal dependencies at longer lag periods, including LI (lag 3), WD (lag 2), TT (lag 2), SI (lag 2), SWEAT (lag 4), and MVV (lag 7). This finding suggests that different atmospheric variables influence rainfall over different temporal scales, reflecting the diverse evolution of atmospheric processes.

Furthermore, the target variable (CH) achieved the highest autocorrelation at lag 2, indicating that previous rainfall observations also contribute to the current prediction process. The selected lag values were subsequently incorporated during feature engineering to preserve the temporal dependency of each predictor before model training. By assigning lag values individually rather than applying a uniform lag to all variables, the proposed approach better captures the unique temporal characteristics of each atmospheric parameter, thereby improving the representation of historical information in the Hybrid TabNet–XGBoost model.

E. Data Splitting

After the feature engineering stage was completed, the dataset was divided into training and testing subsets using the chronological split approach. Data collected from January 2019 to December 2024 were used for model training, whereas observations from January 2025 to December 2025 were reserved for testing. This partition preserves the temporal order of observations and prevents future information from being introduced during model training. The chronological split ensures that the training and testing datasets preserve identical feature representations while maintaining complete temporal independence for model evaluation. Consequently, the testing dataset provides a realistic benchmark for evaluating the generalization capability of the proposed Hybrid TabNet–XGBoost model.

This chronological partitioning enables the model to be evaluated under conditions that closely resemble real-world forecasting scenarios, where future rainfall observations are unavailable during model training. By preserving the temporal sequence of the data, the evaluation results provide a more reliable assessment of the model's ability to generalize to unseen observations. Consequently, the reported prediction performance reflects the practical applicability of the proposed Hybrid TabNet–XGBoost model for operational rainfall prediction.

F. Hyperparameter Optimization Using Optuna

After the dataset was divided into training and testing subsets, hyperparameter optimization was performed using Optuna combined with TimeSeriesSplit cross-validation. The optimization process evaluated fifty parameter combinations using RMSE as the optimization objective. During each trial, Optuna searched for the parameter configuration that minimized the average validation error. The best hyperparameter configuration is presented in Table 6.

TABLE 6
BEST VALUE HYPERPARAMETER

Parameter	Best Value
Number of Estimators	375
Maximum Depth	3
Learning Rate	0.010798
Subsample	0.862855
Column Sample by Tree	0.796719
Minimum Child Weight	8
Gamma	0.173777
Regularization Alpha	0.856228
Regularization Lambda	2.241035
Sample Type	Uniform
Normalize Type	Tree
Rate Drop	0.277368

Based on the optimization results, the lowest average cross-validation error achieved was 14.2949 RMSE,

indicating that the selected hyperparameter configuration provided the best balance between prediction accuracy and model generalization during the optimization process. Consequently, this parameter combination was adopted for training the final Hybrid TabNet–XGBoost model. The optimization process selected the DART booster with 375 estimators, a maximum tree depth of 3, and a relatively low learning rate (0.010798), suggesting that the model favors a gradual learning process while maintaining moderate tree complexity to reduce the risk of overfitting.

In addition, the optimized sampling parameters (subsample = 0.862855 and colsample_bytree = 0.796719) indicate that only a subset of training samples and predictor variables was utilized during the construction of each tree, thereby improving model diversity and enhancing generalization performance. Furthermore, the relatively large values of min_child_weight, gamma, and the regularization parameters (reg_alpha and reg_lambda) demonstrate that the optimization process favored a more regularized model capable of reducing unnecessary model complexity while preserving predictive accuracy. These findings indicate that the optimized hyperparameter configuration effectively balances model flexibility and robustness for rainfall prediction.

G. Hybrid TabNet-XGBoost Performance

After obtaining the optimal hyperparameter configuration, the Hybrid TabNet–XGBoost model was trained using the training dataset and evaluated using the testing dataset. Figure 2 illustrates the comparison between the observed and predicted rainfall values, while the quantitative evaluation results are summarized in Table 7.

TABLE 7
EVALUATION RESULT

Metric	Value
RMSE	19.2525
MAE	12.0255
MAPE	17.28%
R^2	0.0449

Based on Table 7, the Hybrid TabNet–XGBoost model achieved an RMSE of 19.1347 mm, MAE of 11.9742 mm, MAPE of 17.62%, and an R^2 of 0.0449 on the testing dataset. The RMSE and MAE values indicate that the model is capable of maintaining relatively low prediction errors for most daily rainfall observations, while the MAPE value suggests satisfactory relative prediction accuracy. However, the relatively low R^2 value indicates that only a small proportion of the variability in the observed rainfall is explained by the model. This result suggests that although the proposed model can capture the general rainfall trend, considerable variability remains unexplained, particularly during highly variable rainfall events. Since daily rainfall observations frequently contain zero or near-zero values, RMSE, MAE, and R^2 provide a more reliable assessment of model performance than MAPE alone.

After evaluating the model using quantitative performance metrics, the prediction results were further analyzed by comparing the observed and predicted rainfall values throughout the 2025 testing period. The comparison is illustrated in Figure 4, which provides a visual representation of the model's capability to capture temporal rainfall variations.

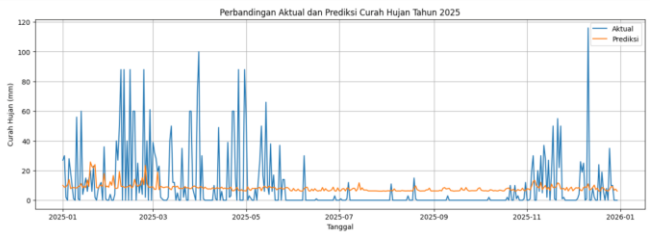

Figure 4 Actual vs Predict Plot

As shown in Figure 4, the Hybrid TabNet–XGBoost model is able to follow the overall rainfall trend during most of the testing period. The predicted values generally remain close to the observed rainfall during low to moderate rainfall conditions, indicating that the model successfully captures the dominant temporal patterns learned from the training data. However, noticeable differences are observed during several high-intensity rainfall events. When the observed rainfall reaches approximately 80–120 mm, the predicted values remain substantially lower, generally ranging between 7–15 mm. This behavior indicates that the proposed model tends to underestimate extreme rainfall events. The comparison also shows that prediction errors become more pronounced as rainfall intensity increases, whereas relatively small deviations are observed during low and moderate rainfall events.

The relatively low R^2 value can be attributed to several characteristics of the dataset. First, the rainfall observations exhibit a highly imbalanced distribution, where most observations correspond to zero or low rainfall, while high-intensity rainfall events occur only occasionally. Second, rainfall is influenced by complex atmospheric processes that involve interactions among numerous meteorological variables, making it difficult to fully represent the observed variability using the available predictors. Finally, although the integration of radiosonde and surface meteorological observations provides valuable atmospheric information, local-scale phenomena such as convective cloud development and mesoscale weather systems may not be completely captured by the available observations. Consequently, the model demonstrates good performance under normal rainfall conditions but experiences reduced accuracy when predicting rare extreme rainfall events.

Overall, the proposed Hybrid TabNet–XGBoost model demonstrates the ability to capture the general temporal

pattern of daily rainfall and provides stable predictions for the dominant rainfall conditions observed in the study area. Nevertheless, the model exhibits reduced performance during high-intensity rainfall events, leading to an underestimation of extreme rainfall and a relatively low R^2 value. Therefore, the proposed model should be considered effective for representing general rainfall variability within the Surabaya region and the characteristics of the available dataset, while further improvements through the incorporation of additional atmospheric observations, spatial information, or remote sensing data may enhance its capability to predict extreme rainfall events.

H. Prediction

After the Hybrid TabNet–XGBoost model was evaluated using the 2025 testing dataset, the optimized model was subsequently applied to generate daily rainfall predictions for the January–December 2026 period. The prediction process was performed using the processed atmospheric variables together with the selected lag features generated during feature engineering. A sample of the prediction results is presented in Table 8.

TABLE 8
PREDICTED RAINFALL 2026

Date	Predicted Rainfall (mm)
2026-01-01	11.198052
2026-01-02	12.151113
2026-01-03	12.151113
...	...
2026-12-29	12.151113
2026-12-30	12.151113
2026-12-31	11.310021

The prediction results indicate that the optimized Hybrid TabNet–XGBoost model successfully generated continuous daily rainfall estimates throughout the 2026 prediction period. The predicted rainfall values generally range between 11 mm and 12 mm, resulting in relatively stable prediction outputs over time. This relatively narrow prediction range reflects the dominant rainfall patterns learned from the historical atmospheric observations used during model training.

The relatively stable prediction values also indicate that the proposed model tends to generate predictions that are dominated by the prevailing rainfall conditions represented in the training dataset. This behavior is consistent with the evaluation results presented in the previous section, where the model demonstrated satisfactory performance for low-to-moderate rainfall events but exhibited reduced sensitivity when predicting high-intensity rainfall. Consequently, the generated predictions should be interpreted as model projections based on historical atmospheric relationships rather than as precise estimates of future rainfall variability.

Furthermore, the tendency to produce nearly constant rainfall predictions suggests that the model has not fully captured the complex variability associated with extreme rainfall events. This limitation is likely related to the imbalanced distribution of the training dataset, where observations of heavy rainfall occur much less frequently than low-rainfall events. As a result, the model primarily learns the dominant rainfall patterns while the representation of rare extreme events remains limited. Future studies may improve prediction variability by incorporating additional meteorological observations, spatial information, satellite-based data, or more balanced training datasets to better represent extreme rainfall conditions.

It should be noted that the prediction results for 2026 have not yet been validated against actual rainfall observations because the corresponding observational data were unavailable at the time of this study. Therefore, the generated rainfall values should be regarded as forecast projections illustrating the practical application of the proposed Hybrid TabNet–XGBoost model rather than as evidence of forecasting accuracy. A comprehensive evaluation of the model's predictive performance for 2026 can only be conducted after the corresponding observational data become available.

IV. CONCLUSION

This study proposed a Hybrid TabNet–XGBoost model for daily rainfall prediction by integrating radiosonde observations with surface meteorological data. The proposed framework combines TabNet for feature selection with XGBoost for rainfall prediction, while temporal dependencies are incorporated through lag feature engineering based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analyses. Hyperparameter optimization using Optuna further improved the XGBoost configuration prior to model training.

The feature selection process identified fifteen atmospheric variables with the highest contribution to the learning process, including LCL, 500 hPa, WS, KO, LI, Humi0, 850 hPa, 700 hPa, TPW, Press0, WD, TT, SI, MVV, and SWEAT. The lag analysis indicated that most variables exhibited the strongest temporal dependency at lag 1, whereas several atmospheric parameters required longer lag periods to better represent their temporal relationships.

Using the optimized hyperparameter configuration, the proposed Hybrid TabNet–XGBoost model achieved an RMSE of 19.1347 mm, an MAE of 11.9742 mm, a MAPE of 17.62%, and an R^2 of 0.0449 on the testing dataset. These results indicate that the proposed model is capable of capturing the general temporal pattern of daily rainfall and provides satisfactory prediction performance for the dominant rainfall conditions represented in the dataset. However, the relatively low R^2 value and the comparison between observed and predicted rainfall indicate that the model tends to underestimate high-intensity rainfall events, primarily due to the imbalanced distribution of rainfall observations and the complex nature of extreme rainfall processes.

The optimized model was subsequently applied to generate daily rainfall projections for the 2026 period using historical atmospheric observations. Since the corresponding observational data were unavailable at the time of this study, these results should be interpreted as forecast projections rather than validated forecasting performance. Therefore, the applicability of the proposed framework is currently limited to the characteristics of the Surabaya study area and the available observational dataset. Future research should investigate the integration of additional meteorological variables, satellite and radar observations, longer observation periods, and multi-station datasets to improve model generalization and enhance the prediction of extreme rainfall events.

BIBLIOGRAPHY

[1] Intergovernmental Panel on Climate Change (IPCC), Climate Change 2023: Synthesis Report. Geneva, Switzerland: IPCC, 2023.

[2] World Meteorological Organization (WMO), State of the Global Climate 2023. Geneva, Switzerland: WMO, 2024.

[3] Badan Nasional Penanggulangan Bencana (BNPB), Indeks Risiko Bencana Indonesia 2024. Jakarta, Indonesia: BNPB, 2024.

[4] Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), Prakiraan Musim Hujan Indonesia 2024/2025. Jakarta, Indonesia: BMKG, 2024.

[5] R. R. Rogers and M. K. Yau, A Short Course in Cloud Physics, 3rd ed. Oxford, U.K.: Butterworth-Heinemann, 1989.

[6] T. N. Krishnamurti, L. Stefanova, and V. Misra, Tropical Meteorology: An Introduction. New York, NY, USA: Springer, 2013.

[7] Pemerintah Kota Surabaya, Laporan Penanggulangan Banjir Kota Surabaya Tahun 2024. Surabaya, Indonesia, 2024.

[8] A. Abbot and J. Marohasy, "Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks," Atmospheric Research, vol. 138, pp. 166–178, 2014.

[9] H. Hersbach et al., "The ERA5 global reanalysis," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020.

[10] World Meteorological Organization (WMO), Guide to Instruments and Methods of Observation (WMO-No. 8), 2021 ed. Geneva, Switzerland: WMO, 2021.

[11] R. A. Anthes, "Data assimilation and radiosonde observations in numerical weather prediction," Bulletin of the American Meteorological Society, vol. 100, no. 7, pp. 1285–1301, 2019.

[12] C. A. Doswell III, H. E. Brooks, and R. A. Maddox, "Flash flood forecasting: An ingredients-based methodology," Weather and Forecasting, vol. 11, no. 4, pp. 560–581, 1996.

[13] G. J. Haklander and A. Van Delden, "Thunderstorm predictors and their forecast skill for the Netherlands," Atmospheric Research, vol. 67–68, pp. 273–299, 2003.

[14] P. Markowski and Y. Richardson, Mesoscale Meteorology in Midlatitudes. Chichester, U.K.: Wiley-Blackwell, 2010.

[15] H. V. Gupta, S. Sorooshian, and P. O. Yapo, "Toward improved calibration of hydrologic models," Water Resources Research, vol. 34, no. 4, pp. 751–763, 1998.

[16] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, 2021.

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[18] A. Borisov, I. Seßler, T. Leemann, K. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7499–7519, 2024.

[19] Mosavi, A., Ozturk, P., and K. W. Chau, "Flood prediction using machine learning models: Literature review," Water, vol. 10, no. 11, Art. no. 1536, 2018.

[20] Ke, G., et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[21] Hochreiter, S., and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] H. Hersbach et al., "The ERA5 global reanalysis," Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020.

[23] A. Borisov, I. Seßler, T. Leemann, K. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7499–7519, 2024.

[24] X. Liu, Y. Zhang, and J. Wang, "A Hybrid TabNet-XGBoost Model for Tabular Data Prediction," Applied Soft Computing, vol. 145, Art. no. 110560, 2023.

[25] C. J. Willmott and K. Matsuura, "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE)," Climate Research, vol. 30, pp. 79–82, 2005.

[26] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[27] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.

[28] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[29] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[30] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, 2021.

[31] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, 2021.

[32] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[33] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, Applied Linear Statistical Models, 5th ed. New York, NY, USA: McGraw-Hill, 2005.

[34] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.

[35] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Anchorage, AK, USA, 2019, pp. 2623–2631.

[36] X. Liu, Y. Zhang, and J. Wang, "A Hybrid TabNet-XGBoost Model for Tabular Data Prediction," Applied Soft Computing, vol. 145, Art. no. 110560, 2023.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[38] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[39] C. J. Willmott and K. Matsuura, "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE)," Climate Research, vol. 30, pp. 79–82, 2005.

[40] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[41] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.