HANIF
Journal of Information Systems
https://journal.ilmubersama.com/index.php/hanif

Machine Learning

Flood Prediction Model Using the Random Forest Algorithm in Padangsidimpuan City

Alvi Sahrin Nasution [1], Putri Maulidina Fadilah [2] ; Muhammad Hafiz [3]

[1,2,3] *Department of Mathematic, Faculty of Mathematic and Natural Sciens, Universitas Negeri Medan, Medan, 20221, North Sumatra, Indonesia*

ARTICLE INFORMATION

Received: February 00, 00
Revised: March 00, 00
Available Online: April 00, 00

KEYWORDS

Five Keyword Maximum; banjir,random forest, spasial-temporal, mitigasi

CORRESPONDENCE

Phone: +62 85270115452
E-mail: alvisahrin@unimed.ac.id

A B S T R A C T

Flooding is a complex environmental phenomenon influenced by meteorological, temporal, and urban conditions. This study aims to develop a flood risk prediction model in Padangsidimpuan City using the Random Forest algorithm by integrating meteorological and temporal variables. The data set consists of historical observations from Padangsidimpuan City, including rainfall, temperature, humidity, wind direction, and flood occurrence status. The data were processed and divided into training and testing sets to evaluate model performance. The results indicate that the Random Forest model achieves strong classification performance, with an accuracy of 0.98 and specificity of 0.99, demonstrating a high capability in correctly identifying non-flood conditions. The model also shows good discrimination ability, with a ROC-AUC value of approximately 0.87. However, the recall value of 0.50 suggests that only half of the actual flood events are correctly detected, primarily due to class imbalance between flood and non-flood data. Feature importance analysis reveals that short-term meteorological variables, particularly rainfall and temperature, along with temporal patterns, are the most influential factors in flood prediction. In addition, spatial interpretation shows that flood-prone areas in Padangsidimpuan City are concentrated in densely populated urban zones with limited drainage capacity, highlighting the influence of urban environmental conditions on flood risk.Overall, the Random Forest model provides a strong foundation for flood risk prediction in Padangsidimpuan City. However, further improvements in data balancing and model optimization are required to enhance sensitivity and support a more reliable flood early warning system.

INTRODUCTION

Flooding is a complex hydroelectrically phenomenon that involves multiple interacting dimensions. It is not solely triggered by high rainfall intensity but is also influenced by the interplay between environmental conditions, land-use patterns, and the performance of urban infrastructure systems. This phenomenon has become a critical issue in many developing cities in Indonesia, including Padangsidimpuan City. Recent flood events indicate that existing mitigation systems still have limitations in responding effectively to dynamic spatial and temporal changes [1]. In recent years, the frequency and intensity of flood events in Padangsidimpuan City have shown an increasing trend. One significant event occurred in early 2025, when flash floods affected several districts and impacted thousands of residents. This event was primarily triggered by the overflow of river systems that were unable to accommodate excessive water discharge, coupled with drainage systems that were not functioning optimally. These conditions highlight the vulnerability of urban systems and underscore the need for more adaptive and data-driven flood mitigation strategies [2].

https://doi.org/10.12345/xxxxx

These events indicate that flood control efforts in Padangsidimpuan continue to face substantial challenges, particularly with respect to spatial planning, the effectiveness of drainage systems, and the accuracy of flood risk prediction [3]. Such conditions highlight the critical need for the adoption of data-driven approaches that are capable of capturing both spatial heterogeneity and temporal dynamics within urban environments. By incorporating these aspects, flood mitigation strategies can be formulated in a more adaptive, robust, and responsive manner to ongoing environmental change [4]. From a methodological perspective, several previous studies have demonstrated that the integration of spatial analysis with machine learning techniques can significantly improve the accuracy of identifying flood-prone areas. For instance, a study utilized a geospatial-based Random Forest algorithm to predict flood events and successfully developed an early warning system with a high level of accuracy [5]. Furthermore, a comparative study revealed that the Random Forest algorithm often outperforms other methods in classifying flood occurrences. These findings suggest that such an approach is highly suitable for use as a model in disaster risk analysis, particularly in urban areas [6].

On the other hand, Padangsidimpuan City exhibits relatively complex geographical and hydrological characteristics, as indicated by variations in topography, the presence of river basins, and the dynamic changes in land use. The combination of these factors contributes to an increased potential for flood occurrences in the region [7].

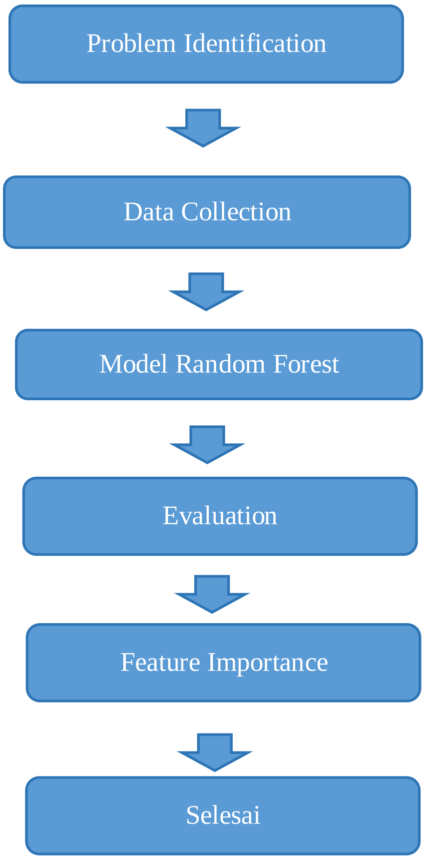
METHOD

This study adopted a qualitative approach integrated with the implementation of the Random Forest algorithm for predictive modeling. The dataset consisted of 730 observational records collected over the period from March 3, 2024, to March 2, 2026. The independent variables included rainfall (RR), maximum temperature (TX), minimum temperature (Tn), average temperature (T avg), relative humidity (RH_avg), and wind direction at maximum speed (ddd_x). Meanwhile, the dependent variable was the flood status, which was defined as a binary outcome (flood and non-flood).

The research process began with exploratory data analysis to identify the characteristics and distribution patterns of each variable. Subsequently, the dataset was separated into predictor variables and the target variable. The data were then divided into training and testing sets using an 80:20 ratio. This data partitioning was intended to evaluate the model’s ability to generalize to unseen data, thereby ensuring the robustness and reliability of the predictive model.

This section systematically described the research methodology employed in developing a flood risk prediction model for Padangsidimpuan City using the Random Forest algorithm. The description encompassed the tools and datasets utilized, the research procedures, data processing techniques, and the methods applied to evaluate the performance of the proposed model. A well-structured methodological framework was essential to ensure transparency and reproducibility, allowing the results to be verified and further developed by other researchers [8], [9]. Furthermore, the selection of the Random Forest algorithm was motivated by its capability to handle non-linear data, reduce the risk of overfitting, and provide reliable estimates of feature importance in predictive analysis [10]. Therefore, the comprehensive presentation of the methodology not only offered a clear understanding of the adopted approach but also strengthened the scientific validity and robustness of the developed model [9].

The research procedure was organized into a series of structured stages, beginning with the collection of historical flood event data and continuing through to the evaluation of the developed prediction model. The overall workflow of the study was illustrated in a flowchart, as presented in Figure 1. Each stage depicted in Figure 1 was further elaborated in detail in the subsequent sections.



Picture 1. Research Flowchart

Based on the conceptual framework presented above, the discussion for each stage can be systematically described as follows:

1. The problem addressed in this study **was** the increasing occurrence of flooding in Padangsidimpuan City, which **had not yet been matched** by an accurate and adaptive flood prediction modeling approach. This condition **indicated** the need for a more effective method to identify flood-prone areas based on complex spatial and temporal characteristics. In addition, although the Random Forest algorithm **had been widely applied** and **had proven** to perform well in various flood prediction studies, its implementation in regions with diverse geographical and hydrological conditions **remained** limited. Therefore, this study **focused** on developing a flood prediction model using the Random Forest algorithm to improve the accuracy of flood risk estimation in Padangsidimpuan City.
2. The data collection technique in this study **employed** a documentation method by utilizing secondary data obtained from official agencies, namely the Meteorology, Climatology, and Geophysics Agency (BMKG) and the Regional Disaster Management Agency (BPBD) of Padangsidimpuan City. The data **were accessed** through the official websites of both institutions, which **provided** systematic records of historical flood events as well as meteorological conditions in the study area. The use of secondary data **was considered** reliable because it **was sourced** from institutions that **implemented** structured, continuous, and accountable data recording systems, ensuring the validity and credibility of the information used in this research.
3. Random Forest is an ensemble learning algorithm in machine learning that integrates the predictions of multiple Decision Trees constructed in a random manner. By leveraging the diversity among individual models and aggregating their outputs, Random Forest can improve prediction accuracy and effectively reduce the risk of overfitting, particularly when handling complex datasets. In general, this algorithm operates by constructing a large number of decision

trees, each trained on randomly selected subsets of the training data. For classification tasks, the final output is determined using a majority voting mechanism across all trees, whereas for regression tasks, the prediction is obtained by averaging the outputs of all trees [11].Mathematically, the classification prediction in Random Forest can be expressed as follows:

$$\hat{y} = mode(\{h_p(x)\}_{b=1}^{B})$$

Where :

B is number of tree

$h_b(x)$ is prediction from trees to B

$\hat{y}$ is final prediction result

4. A confusion matrix was a widely used evaluation metric in classification tasks within machine learning, designed to provide a detailed representation of model performance based on predicted outcomes for each target class. This matrix presented information on the number of true positives, true negatives, false positives, and false negatives, thereby enabling a more comprehensive analysis of the model's ability to accurately distinguish between classes [12]. Consequently, the confusion matrix served as a crucial evaluation tool for assessing model reliability and identifying areas that required improvement in order to enhance predictive performance. Table 1 below presented the confusion matrix [13].

Table 1. Confusion Matrix

Actual	**Prediction**	
	TRUE	**FALSE**
TRUE	TP	FN
FALSE	FP	TN

The performance of the proposed model is evaluated using several standard metrics, including accuracy, precision, recall, and F1-score. These metrics are employed to comprehensively assess the classification performance.

RESULTS AND DISCUSSION

This study aims to develop a flood risk prediction model in disaster-prone areas of Padangsidimpuan City by utilizing the *Random Forest* algorithm. The research process is carried out through a series of structured stages, beginning with the collection of relevant secondary data, followed by a prepossessing stage to ensure data quality and readiness. Subsequently, the dataset is divided into training and testing sets prior to model development. The trained model is then tested and evaluated to assess its performance in accurately and consistently classifying flood risk.
The datasets employed in this study consists of flood event records in Padangsidimpuan City during the period 2024–2026, obtained from secondary sources. The data were collected from official institutions, namely the Meteorological, Climatology, and Geophysical Agency (BMKG) and the Regional Disaster Management Agency (BPBD) of Padangsidimpuan City. This datasets comprises seven historical records of flood occurrences, incorporating several key variables, including location, temporal information (month and year), rainfall, season, and flood status. These variables serve as the basis for the analysis to identify patterns and factors influencing flood events.

 ISSN 3025-2342 (ONLINE)

Figure 2. Flood Susceptibility Map of Padangsidimpuan City

Fig. 2 illustrates the spatio-temporal distribution of flood susceptibility in Padangsidimpuan City derived from the Random Forest model. The results indicate that areas with high flood risk are consistently concentrated in the urban core, whereas peripheral regions exhibit lower risk levels and remain relatively stable throughout the observation period. The machine learning-based flood susceptibility map further reveals that high-risk zones are predominantly located in districts characterized by intense urbanization and limited drainage capacity. This finding suggests that anthropogenic factors, particularly land-use changes, play a significant role in influencing flood occurrences. Moreover, areas classified under moderate risk form transitional zones shaped by the interaction between climatic conditions and geographical characteristics. In contrast, low-risk areas are generally situated in regions with higher elevation and lower population density.

The data preparation phase was rigorously conducted to ensure the integrity, consistency, and analytical readiness of the datasets for the development of a flood risk prediction model based on the *Random Forest* algorithm. Initially, heterogeneous data from multiple sources were systematically imported and consolidated within the Python environment to construct a comprehensive and coherent datasets representing both temporal and spatial dimensions. Subsequent prepossessing procedures involved data cleansing operations, including the identification and treatment of missing values as well as the elimination of duplicate records to prevent bias in model training. For continuous variables, missing data were addressed using interpolation techniques in the case of time-series observations or through mean imputation. For categorical features, the most frequent category was employed to replace missing entries, thereby preserving the inherent distribution of the data. Subsequent prepossessing procedures involved data cleansing operations, including the identification and treatment of missing values as well as the elimination of duplicate records to prevent bias in model training. For continuous variables, missing data were addressed using interpolation techniques in the case of time-series observations or through mean imputation. For categorical features, the most frequent category was employed to replace missing entries, thereby preserving the inherent distribution of the data.

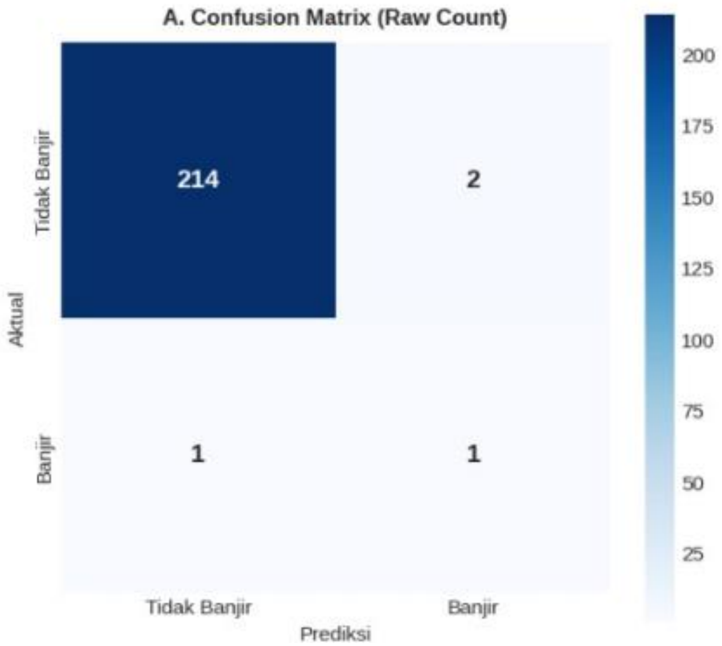
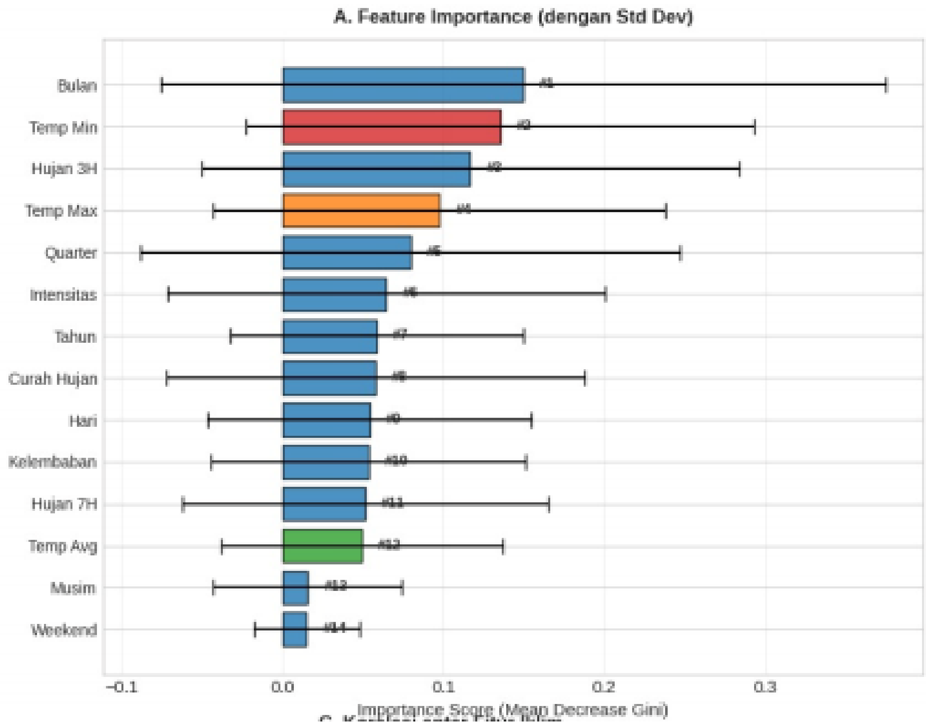


Figure 3.Convution Matrix

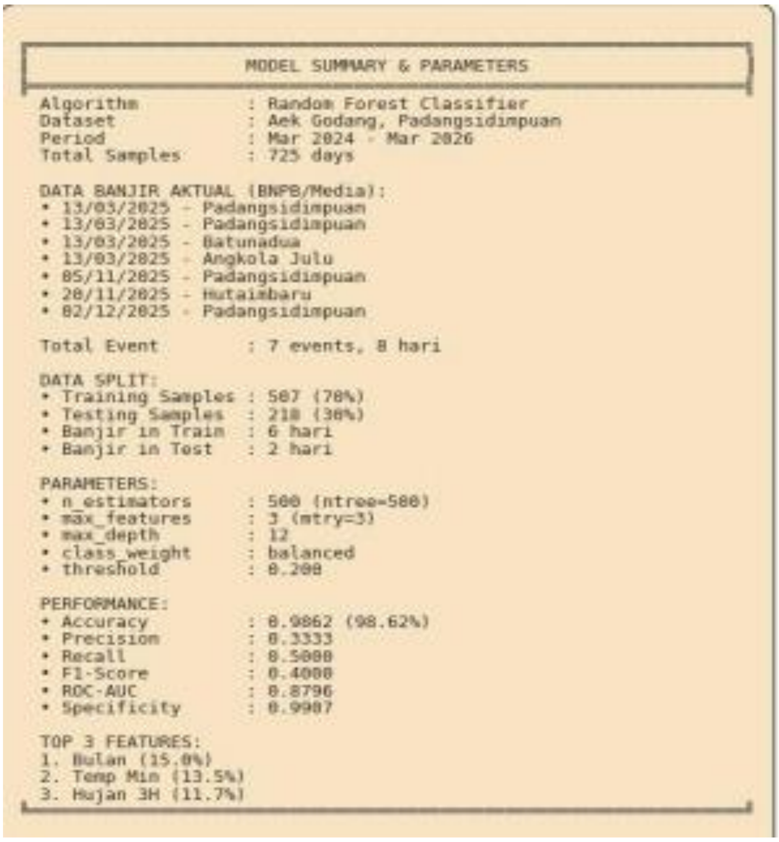
Fig 3. The performance of the proposed Random Forest model was evaluated using a confusion matrix. The results indicate that the model correctly classified 214 non-flood instances as true negatives and successfully identified 1 flood instance as a true positive. However, several misclassifications were observed, including 1 flood instance incorrectly predicted as non-flood (false negative) and 2 non-flood instances misclassified as flood events (false positives).The presence of false negatives is of particular concern, as it may lead to undetected flood events, which could have serious implications in real-world applications. On the other hand, false positives may result in unnecessary alerts, potentially reducing the reliability of the system. These findings suggest that while the model demonstrates strong performance in recognizing the majority class (non-flood), its ability to detect the minority class (flood) remains limited.

This limitation is likely due to the imbalanced nature of the dataset, where non-flood instances significantly outnumber flood instances. Despite this, the recall value of 0.50 for the flood class indicates that the model is still capable of identifying a portion of actual flood events. Overall, the Random Forest model achieves high overall accuracy; however, its sensitivity to flood occurrences requires further improvement. To enhance model performance, particularly in detecting flood events, future work may consider the implementation of data balancing techniques such as oversampling, undersampling, or Synthetic Minority Over-sampling Technique (SMOTE), as well as hyperparameter optimization to improve classification performance on the minority class.



Picture 4.Feature Importance

The feature importance analysis reveals that temporal and meteorological variables—particularly month, minimum temperature, maximum temperature, and short-term rainfall—serve as the primary determinants influencing the predictions of the Random Forest model. These variables contribute most significantly to enhancing the model's ability to distinguish between flood and non-flood events.In contrast, other variables such as humidity, average temperature, and time-related indicators (e.g., weekends) exhibit relatively minor contributions, indicating that their influence on the model can be considered limited. Furthermore, the variability in importance scores, as reflected by the standard deviation, suggests that certain variables demonstrate differing levels of consistency across individual decision trees within the ensemble.Overall, these findings suggest that the model is more responsive to short-term weather dynamics and temporal patterns, which aligns with the inherent nature of flood events that are strongly driven by meteorological conditions.



```
MODEL SUMMARY & PARAMETERS
Algorithm          : Random Forest Classifier
Dataset            : Aek Godang, Padangsidimpuan
Period             : Mar 2024 - Mar 2026
Total Samples      : 725 days

DATA BANJIR AKTUAL (BNPB/Media):
• 13/03/2025 - Padangsidimpuan
• 13/03/2025 - Padangsidimpuan
• 13/03/2025 - Batunadua
• 13/03/2025 - Angkola Julu
• 05/11/2025 - Padangsidimpuan
• 20/11/2025 - Hutaimbaru
• 02/12/2025 - Padangsidimpuan

Total Event        : 7 events, 8 hari

DATA SPLIT:
• Training Samples : 507 (70%)
• Testing Samples  : 218 (30%)
• Banjir in Train  : 6 hari
• Banjir in Test   : 2 hari

PARAMETERS:
• n_estimators     : 500 (ntree=500)
• max_features     : 3 (mtry=3)
• max_depth        : 12
• class_weight     : balanced
• threshold        : 0.200

PERFORMANCE:
• Accuracy         : 0.9862 (98.62%)
• Precision        : 0.3333
• Recall           : 0.5000
• F1-Score         : 0.4000
• ROC-AUC          : 0.8796
• Specificity      : 0.9907

TOP 3 FEATURES:
1. Bulan (15.0%)
2. Temp Min (13.5%)
3. Hujan 3H (11.7%)
```

Figure 5. Model Summary Algoritma Random Forest

Figure 5. The Random Forest model developed in this study demonstrated strong overall classification performance. It achieved an accuracy of 0.98, with a sensitivity (recall) of 0.50, indicating that the model detected half of the flood events. The specificity reached 0.99, reflecting a high capability in correctly identifying non-flood conditions. These results suggest that the model was able to reliably distinguish flood occurrences despite the relatively small number of positive (flood) cases compared to non-flood cases.

The model's discriminative capability is further supported by the Receiver Operating Characteristic (ROC) curve, with an area under the curve (AUC) of approximately 0.87. The curve's clear deviation from the diagonal line indicates that the Random Forest model performs substantially better than random classification. These findings suggest that the model maintains a reasonable balance between the true positive rate (sensitivity) and the false positive rate.

CONCLUSION

This research shows that the Random Forest algorithm is a suitable and effective method for estimating flood risk in Padangsidimpuan City by incorporating both meteorological and temporal variables. The model demonstrates excellent overall classification performance, with an accuracy of 0.98 and a specificity of 0.99, indicating a strong capability in correctly identifying non-flood conditions. In addition, the ROC-AUC value of around 0.87 confirms that the model possesses good discriminative power and performs substantially better than random guessing. Nevertheless, the model exhibits a recall value of 0.50, meaning that only half of the actual flood events are correctly detected. This weakness is largely attributed to the imbalanced distribution of the dataset, where non-flood cases dominate over flood occurrences. As a result, although the model performs well in general classification tasks, its reliability for early flood warning applications is still limited and requires further enhancement. The analysis of feature importance indicates that short-term meteorological variables, especially rainfall and temperature-related factors, together with temporal characteristics, are the most influential predictors of flood events. Furthermore, the spatial interpretation suggests that areas with high flood susceptibility are mainly located in densely urbanized zones with limited drainage capacity, highlighting the important role of human-induced environmental changes in increasing flood risk. In summary, while the proposed model provides a strong baseline for flood risk prediction, future studies should focus on improving model sensitivity through data balancing strategies, feature engineering, and parameter optimization. These improvements are necessary to develop a more accurate and reliable early warning system that can better support flood disaster mitigation efforts in Padangsidimpuan City.

REFERENCES

Electronic Document

[1] Nasional TVRINews. (2025). *Banjir dan Longsor Melanda Kota Padangsidimpuan, 1.504 Warga Terkena Dampak*.

[2] Detik.com. (2025). *Banjir Terjang Padangsidimpuan, Satu Orang Diduga Hanyut.*

[3] Hariantabagsel.com. (2024). *Drainase Tidak Berfungsi, Puluhan Rumah Warga di Kota Padangsidimpuan Kebanjiran.*

[4] Pemerintah Kota Padangsidimpuan. (2025). *Plt. Ketua Dharma Wanita Serahkan Bantuan untuk Korban Banjir dan Longsor*.

Jurnal

[5] Putra, M.R.P., Ashari, R., & Muhirin. (2025). *Flood Prediction Using Machine Learning Model Integrated with Geographical Information System*. Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika.

[6] Maulita, I., Widiawati, C.R.A., & Wahid, A.M. (2025). *Analisis Komparatif Linear Regression, Random Forest, dan Gradient Boosting untuk Prediksi Banjir*. Jurnal Pendidikan dan Teknologi Indonesia (JPTI).

[10] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] S. Chen, Y. Wang, and L. Zhang, “Flood susceptibility mapping using Random Forest model and GIS,” *Natural Hazards*, vol. 92, pp. 1–20, 2018.

[12] Düntsch, I., & Gediga, G. (2020). *Confusion matrices and rough set data analysis*. International Journal of Approximate Reasoning

[13] Widjiyati, “Analisis risiko banjir menggunakan Random Forest,” *Jurnal Hidrologi Indonesia*, vol. 5, no. 2, pp. 45–52, 2021.

Book

[8] W. K. Trochim, J. P. Donnelly, and K. Arora, *Research Methods: The Essential Knowledge Base*, 2nd ed. Boston, MA, USA: Cengage Learning, 2016.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.