International Journal of Informatics Engineering and Computing (IJIMATIC)
Vol. 2, No.1, 2025
E-ISSN: 3090-9112

Efficient Flood Prediction with SVM and RF Algorithm

Juwita Sampe Ruru[1]

Abstract

Flood is a high risk of natural disasters such as floods due to its geological location at the intersection of four major tectonic plates. This study aims to predict flood risks using the Support Vector Machine (SVM) and Random Forest (RF) algorithms, utilizing rainfall, topography, and land use data. Historical rainfall data were obtained from BMKG, topographic data from GIS, and land use data from satellite imagery. The evaluation results show that the RF algorithm outperforms SVM, achieving 92.1% accuracy and an F1-score of 91.8%. RF has proven effective in capturing non-linear relationships between features influencing flood risk. This predictive system is expected to aid disaster mitigation, spatial planning, and the development of an early flood warning system.

Keywords:
Natural disasters, flood prediction, RF, SVM, Rainfall

1. Introduction

Indonesia is a country with a high risk of natural disasters such as tsunamis, floods, forest fires, landslides, and volcanic eruptions. Natural disasters can be interpreted as natural phenomena that result in damage and destruction of the environment so that in the end it causes casualties, losses to property and property and causes damage to buildings in the environment. Its geological location at the confluence of four major plates (Eurasia, Indo-Australia, the Philippines, and the Pacific) makes Indonesia vulnerable to disasters. According to Law No. 24 of 2007, natural disasters are phenomena that interfere with human life, causing material, immaterial, and psychological losses [1],[2].

BNPB data shows an increase in the number of floods, from 525 incidents in 2015 to 1,794 in 2021. From 2019 to 2023, 7,168 flood incidents were recorded, causing fatalities and infrastructure losses. To anticipate the impact of disasters, predictions are needed using algorithms such as Support Vector Machine (SVM) and Random Forest [3]. Lombok is one of the areas in Indonesia that is prone to flooding. due to a combination of high rainfall, land use changes, and topography that supports rapid water flow. To reduce the impact of flooding, an accurate prediction system is needed [4].

Machine learning technologies such as Support Vector Machine (SVM) and Random Forest (RF) have been widely used in natural disaster prediction. SVM works by finding the best hyperplane to separate data based on class, while RF uses an ensemble approach to improve prediction accuracy (Disaster Prediction Technology Literature) [5]. The algorithm-based prediction is expected to help the government and the community in mitigating future flood disasters, by utilizing historical data to improve preparedness [6].

Efficient flood prediction using Support Vector Machines (SVM) and Random Forest (RF) algorithms faces several challenges, primarily related to model accuracy and data

Corresponding Author: Juwita Sampe Ruru1, Universitas Atma Jaya, Yogyakarta, Indonesia

integration. While SVM has shown higher precision in minimizing false positives, its overall accuracy remains limited, as evidenced by a study reporting only 65.61% accuracy for tsunami predictions. Conversely, RF, although exhibiting better recall, struggles with precision, leading to potential misclassifications in flood events. Additionally, the complexity of flash flood dynamics, particularly in regions like the Eastern Mediterranean, necessitates the integration of diverse meteorological data, such as precipitable water vapor and lightning occurrences, to enhance predictive capabilities. Furthermore, the scarcity of observational data in developing regions complicates the implementation of these models, as seen in integrated approaches that combine hydrodynamic modeling with machine learning. Thus, the effective application of SVM and RF in flood prediction requires careful consideration of data quality, model selection, and the specific characteristics of the flooding events being predicted [15],[16],[17],[18].

2. Related Works

Research on flood prediction has evolved significantly over the past decade, moving from traditional statistical methods to advanced machine learning (ML) and deep learning techniques. Traditional approaches typically relied on hydrological and hydraulic modeling, which used empirical equations and historical data to predict flood risks. While these methods provided foundational insights, they often struggled to adapt to changing climate patterns and dynamic land use changes. As a result, there has been a growing emphasis on data-driven approaches that can better capture complex, nonlinear relationships between flood-related variables [19].

Among machine learning methods, Support Vector Machines (SVM) and Random Forest (RF) have gained widespread attention for their robustness and adaptability in flood prediction tasks. SVM is effective in handling small datasets and excels at finding optimal decision boundaries, making it suitable for binary flood classification tasks. However, it has limitations in handling large-scale, high-dimensional data. In contrast, RF, as an ensemble learning technique, combines multiple decision trees to improve accuracy and mitigate overfitting. It is particularly effective in modeling nonlinear interactions among diverse geospatial and meteorological variables. Studies have demonstrated that RF often outperforms SVM in flood susceptibility assessments due to its ability to handle heterogeneous datasets [19].

To further enhance predictive performance, researchers have developed hybrid and ensemble models. One notable example is the Intelligent Committee Machine Learning Flood Forecasting (ICML-FF), which integrates outputs from multiple ML models to generate more reliable river level predictions. Such ensemble approaches leverage the strengths of individual models while compensating for their weaknesses, resulting in improved accuracy and robustness. These techniques are especially valuable in regions with complex hydrological patterns where single-model approaches may fall short [20].

Deep learning has introduced a new dimension to flood prediction research. Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), are particularly adept at modeling temporal dependencies in time series data. LSTM models have been shown to outperform traditional and ML-based models in flood forecasting, particularly in scenarios requiring real-time predictions and adaptive learning from evolving datasets. Their ability to capture long-term temporal patterns makes them highly suitable for dynamic flood forecasting systems [21].

Overall, the transition from classical statistical methods to advanced machine learning and deep learning models has significantly improved the accuracy, speed, and scalability of flood prediction systems. These advancements enable more effective flood management and early warning systems, which are crucial for mitigating the impacts of increasingly frequent and severe flood events worldwide. However, challenges remain, particularly regarding data availability, model interpretability, and integration of diverse data sources — areas that continue to drive ongoing research and innovation in this field [22].

3. Proposed Method

3.1 Dataset

This study uses data historical rainfall Rain daily during 10 year final obtained from Body Meteorology, Climatology, and Geophysics (BMKG). This data includes information on rainfall intensity (in millimeters) per month for each region in Lombok. The we collect data from Geographic Information System (GIS) analysis and includes information on elevation, land slope, and water flow patterns. Elevation And slope land influence speed flow water as well as potential for flooding in flat or low areas. Pattern flow water show existence river or flow experience Which can affect the spread of water during heavy rain. We also gathered data from satellite imagery that identifies land use types, such as residential areas, agriculture, and forests.

3.2 Research Flow

Effective flood prediction using Support Vector Machine (SVM) and Random Forest (RF) algorithms relies heavily on the quality, relevance, and diversity of the collected data. The data collection process begins with gathering historical flood occurrence records, including time, location, water level, rainfall intensity, and duration. These data are often sourced from national meteorological and hydrological agencies, such as BMKG (Meteorology, Climatology, and Geophysics Agency) and BNPB (National Disaster Management Agency), which provide comprehensive datasets on precipitation, river discharge, and past flood events.

Additionally, remote sensing data and satellite imagery are utilized to monitor rainfall distribution and land use changes over time, which are critical factors influencing flood risks. Geographic Information System (GIS) data, including elevation, slope, and land cover, further enrich the dataset, helping the models understand the topographical and environmental conditions of flood-prone areas.

The datasets are then preprocessed to ensure consistency, accuracy, and completeness, involving steps such as data cleaning, normalization, and handling missing values. Feature selection is performed to identify the most influential variables, such as rainfall, river water level, soil saturation, and weather parameters. These refined datasets are then used to train SVM and RF models, enabling them to learn patterns and correlations that are indicative of potential flood events. High-quality data collection and preprocessing are therefore essential to achieving accurate, reliable, and timely flood prediction using machine learning approaches.

This study conducts preprocessing Data by conducting Data Cleaning: Removing outliers and filling in missing values to make the data more representative. Data Normalization : All features are normalized so that their distribution is consistent and can be compared between different features. This step is important to improve the performance of machine learning algorithms. Data Formatting: Organizing data in a format suitable for input into a machine learning model, ensuring the data is acceptable to the algorithm used.

Features Selection by Identifying variables or features that can affect the occurrence of flooding. For example, rainfall, elevation, and land use. Conducting a correlation analysis

between variables to find out which features have a significant relationship with the occurrence of flooding. Features that having a high correlation with the label (flooded or not flooded) were selected for use in the model.

After data ready, next step is share data into two sets: Training Data (70%): Data used to train the machine learning model. Training data will be used to teach the model to recognize patterns in the data. Test Data (30%): Data used to test the model's ability to predict floods on previously unseen data. This division is done randomly so that the model can be tested under varying conditions

3.3 Proposed Algorithm

This research uses the SVM algorithm. In this case, the main focus is on the advantages of SVM regarding complex patterns in data and being able to design and analyze flood detection systems by implementing the SVM, making it easier to monitor or monitor floods. This research will focus on the implementation of SVM to detect floods in artificial simulations, without considering geography and climate. The SVM is a supervised learning algorithm primarily used for classification tasks, although it can also be adapted for regression. Mathematically, SVM aims to find the optimal hyperplane that best separates two classes of data points with the maximum possible margin. Given a training dataset $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where each x_i is a feature vector in $\mathbb{R}^d$and each y_i is a label in $\{-1, +1\}$, the linear SVM formulation seeks to minimize the norm of the weight vector w,

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{Eq. (1)}$$

Subject to constraint $y_i(w^T x_i + b) \geq 1$. This ensures that each data point is correctly classified and lies on the correct side of the margin. However, real-world data is often not perfectly separable. To accommodate this, SVM introduces slack variables $\xi_i \geq 0$and modifies the optimization to allow for some misclassifications. The soft margin SVM objective becomes equation 2:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{Eq. (2)}$$

subject to $y_i(w^T x_i + b) \geq 1 - \xi_i$. The parameter C controls the trade-off between maximizing the margin and minimizing classification errors. For data that is not linearly separable even with slack variables, SVM employs the "kernel trick" to implicitly map data into a higher-dimensional feature space where linear separation is possible. Instead of computing dot products directly in high-dimensional space, a kernel function K $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is used to compute inner products efficiently. In the dual form of the optimization, this leads to the objective in Equation 3:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad \text{Eq. (3)}$$

with constraints $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$. The final decision function for classifying a new data point x is then in Equation 4:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad \text{Eq. (4)}$$

which uses only a subset of the training data called support vectors. SVMs are powerful classifiers that maximize the margin between classes and can be extended to non-linear problems using kernel functions.

4. Result and Analysis

Model Performance
Random Forest superior in all metric evaluation, with accuracy And overall performance is better compared to SVM. Random Forest is more effective in capturing non-linear relationships between features, which plays an important role in flood prediction. SVM shows quite good performance, but is limited in handling data with more complex relationships.

Analysis Factor
Rainfall Rain is a factor dominant in prediction flood. Region with rainfall High rainfall, low elevation, and less steep land slopes are at greater risk of flooding. Areas with land use dominated by settlements have higher risk of flooding, because land conversion reduces water absorption capacity.

Prediction Example
Prediction result:

Rainfall (mm)	Elevation (m)	Slope (°)	Land Use	Current Labels	SVM Prediction	RF Prediction
150	10	2.0	Settlement	1 (Flood)	1 (Flood)	1 (Flood)
30	80	12.5	Forest	0 (No Flood)	0 (No Flood)	0 (No Flood)
100	25	3.5	Ricefield	1 (Flood)	0 (No Flood)	1 (Flood)
20	150	20.0	Forest	0 (No Flood)	0 (No Flood)	0 (No Flood)

Random Forest model can be used to develop a flood early warning system in Lombok. Spatial Planning: Flood risk maps can help in planning better land use to reduce flood risk. Mitigation Disaster: Public can use results prediction For mitigate and raise awareness of the potential for flooding in their area.

Evaluation Results:

Metric	SVM	Random Forest

Accuracy	85.4%	92.1%
Precision	80.2%	90.3%
Recall	88.7%	93.5%
F1-Score	84.2%	91.8%

6. Conclusion

This study shows that the Random Forest algorithm is superior to SVM in predicting floods in Lombok, with higher accuracy and the ability to capture non-linear relationships between factors affecting flood risk. Further research can consider real-time data integration to improve the accuracy and responsiveness of the flood early warning system. Flood is a high risk of natural disasters such as floods due to its geological location at the intersection of four major tectonic plates. The evaluation results show that the RF algorithm outperforms SVM, achieving 92.1% accuracy and an F1-score of 91.8%. RF has proven effective in capturing non-linear relationships between features influencing flood risk. This predictive system is expected to aid disaster mitigation, spatial planning, and the development of an early flood warning system.

References

[1] Rahmaniah, "Analysis of the Causes of Flood Natural Disasters in the Region Indonesia," 2012. [Online]. Available: https://jurnal.habi.ac.id/index.php/Pendikdas

[2] A. Purnomo, J. Indra, E. E. Awal, and T. Rohana, "Analysis of Flood Prediction in Indonesia Using the Support Vector Machine and Random Forest Algorithms," *J. of Science and Humanity*, vol. 6, no. 1, pp. 230–239, 2026, doi: 10.47065/josh.v6i1.5958.

[3] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[6] Meteorology, Climatology, and Geophysics Agency, "Daily Rainfall Data in Indonesia," 2023.

[7] National Disaster Management Agency (BNPB), "Statistics of Natural Disasters in Indonesia," 2023.

[8] OpenStreetMap Contributor, "Land Use and Topographic Data through GIS," 2023.

[9] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2001.

[10] L. Mason and C. Perlich, "Comparison of tree-based ensemble algorithms," *Machine Learning*, vol. 40, no. 4, pp. 623–640, 2009.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[12] D. Jensen and T. Coach, *Introduction to Data Science: Python's Approach to Concepts, Techniques, and Applications*, Jumper, 2017.

[13] Meteorology, Climatology, and Geophysics Agency (BMKG), "Standard Operating Procedures for Rainfall Data Collection in Indonesia," 2023.

[14] West Southeast Island Provincial Government, "Topographic Data and Spatial Layout of the Region," 2023.

[15] H. T. Sukmana, Y. Durachman, A. Amri, and S. Supardi, "Comparative Analysis of SVM and RF Algorithms for Tsunami Prediction: A Performance Evaluation Study," *J. Appl. Data Sci.*, vol. 5, no. 1, Jan. 2024, doi: 10.47738/jads.v5i1.159.

[16] S. Asaly, L.-A. Gottlieb, Y. Yair, C. Price, and Y. Reuveni, "Predicting Eastern Mediterranean Flash Floods Using Support Vector Machines with Precipitable Water Vapor, Pressure, and Lightning Data," *Remote Sens.*, vol. 15, no. 11, 2023, doi: 10.3390/rs15112916.

[17] J. Sampurno, V. Vallaeys, R. Ardianto, and E. Hanert, “Integrated hydrodynamic and machine learning models for compound flooding prediction in a data-scarce estuarine delta,” *Nonlinear Process. Geophys.*, vol. 29, no. 3, pp. 301–315, 2022, doi: 10.5194/npg-29-301-2022.

[18] S. S. Band et al., “Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms,” *Remote Sens.*, vol. 12, no. 21, 2020, doi: 10.3390/rs12213568.

[19] N. Mahdizadeh Gharakhanlou and L. Perez, “Flood susceptible prediction through the use of geospatial variables and machine learning methods,” *J. Hydrol.*, vol. 617, p. 129121, Feb. 2023, doi: 10.1016/j.jhydrol.2023.129121.

[20] Faruq, S. F. M. Hussein, A. Marto, and S. S. Abdullah, “Flood River Water Level Forecasting using Ensemble Machine Learning for Early Warning Systems,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1091, no. 1, p. 012041, Nov. 2022, doi: 10.1088/1755-1315/1091/1/012041.

[21] Q. Zhou et al., “A deep-learning-technique-based data-driven model for accurate and rapid flood predictions in temporal and spatial dimensions,” *Hydrol. Earth Syst. Sci.*, vol. 27, no. 9, pp. 1791–1808, 2023, doi: 10.5194/hess-27-1791-2023.

[22] W.-D. Guo, W.-B. Chen, S.-H. Yeh, C.-H. Chang, and H. Chen, “Prediction of River Stage Using Multistep-Ahead Machine Learning Techniques for a Tidal River of Taiwan,” *Water*, vol. 13, no. 7, 2021, doi: 10.3390/w13070920.