Journal of Information Technology and Computer Science (INTECOMS)
Volume 8 Nomor 1, Tahun 2025
e-ISSN : 2614-1574
p-ISSN : 2621-3249

HYBRID MACHINE LEARNING PREDIKSI BANJIR MENGGUNAKAN LSTM DAN RANDOM FORESTS PADA GEODATA

HYBRID MACHINE LEARNING PREDICTS FLOODING USING LSTM AND RANDOM FORESTS ON GEODATA

Zakiul Fahmi Jailani[1], Dita Nurmadewi[2]
Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Bakrie[1,2]
dita.nurmadewi@bakrie.ac.id[2]

ABSTRACT

Flood prediction remains a critical concern in Indonesia, a nation frequently affected by seasonal deluges. This research aims to predict flood occurrences in five key provinces by employing a hybrid machine learning approach using Long Short-Term Memory (LSTM) networks and Random Forest models. Leveraging geospatial and temporal data from Petabencana.id, collected between January 2018 and February 2024, the study develops a predictive framework for flood forecasting. The analysis integrates flood depth and historical disaster data to estimate the time to the next flood, with predictions starting after the last data entry in February 2024. The model accurately predicted that Jakarta would experience flooding within 25–50 days post-February, a forecast corroborated by significant floods in April 2024. Other provinces, including Central Java and East Java, displayed longer flood risk windows extending further into the year. With a training accuracy of 99%, the model underscores its reliability in predicting flood events. This study emphasizes the strength of LSTM in capturing temporal patterns and the role of Random Forests in identifying key predictive features. The proposed model offers a valuable tool for disaster management agencies and local governments, enabling them to anticipate and mitigate flood impacts using real-time data from Petabencana.id.

Keywords : *Flood Forecasting, Geospatial Analysis, LSTM, Petabencana.id, Random Forest.*

ABSTRAK

Prediksi banjir tetap menjadi perhatian kritis di Indonesia, negara yang sering terkena banjir musiman.Penelitian ini bertujuan untuk memprediksi kejadian banjir di lima provinsi kunci dengan menggunakan pendekatan hybrid machine learning yang memanfaatkan jaringan Long Short-Term Memory (LSTM) dan model Random Forest. Dengan memanfaatkan data geospasial dan temporal dari Petabencana.id, yang dikumpulkan antara Januari 2018 dan Februari 2024, studi ini mengembangkan kerangka kerja prediktif untuk peramalan banjir. Analisis mengintegrasikan kedalaman banjir dan data bencana historis untuk mengestimasi waktu ke banjir berikutnya, dengan prediksi dimulai setelah entri data terakhir pada Februari 2024. Model ini berhasil memprediksi bahwa Jakarta akan mengalami banjir dalam 25–50 hari pasca-Februari, sebuah ramalan yang dibuktikan oleh banjir besar pada April 2024. Provinsi lain, termasuk Jawa Tengah dan Jawa Timur, menunjukkan jendela risiko banjir yang lebih panjang yang berlangsung lebih jauh ke dalam tahun tersebut. Dengan akurasi pelatihan sebesar 99%, model ini menegaskan keandalannya dalam memprediksi peristiwa banjir. Studi ini menekankan kekuatan LSTM dalam menangkap pola temporal dan peran Random Forest dalam mengidentifikasi fitur prediktif kunci. Model yang diusulkan menawarkan alat berharga bagi agensi manajemen bencana dan pemerintah lokal, memungkinkan mereka untuk mengantisipasi dan memitigasi dampak banjir menggunakan data nyata dari Petabencana.id.

***Kata Kunci**:* Peramalan Banjir, Analisis Geospasial, LSTM, Petabencana.id, Random Forest.

INTRODUCTION

Indonesia, a large archipelago of over 17,000 islands, experiences significant challenges from frequent and severe flooding due to a combination of tropical climate, dense river networks, and rapid urbanization (Blöschl & Montanari, 2010). These factors worsen during the rainy season, causing flooding that not only damages infrastructure but also threatens people's lives and livelihoods.

Flooding can occur due to several factors, one of which is high rainfall. Heavy rain, especially that which falls in a short

period of time, is often not fully absorbed by the soil. This excess water then flows into rivers and lakes that may not be able to accommodate the additional flow, causing water to overflow into surrounding areas (Rahayu et al., 2024). Other factors include climate change and geography. Rising global temperatures are changing weather patterns, which can cause more intense and frequent rainfall in some areas, while causing drought in others. These changes can increase the frequency and intensity of flooding. Areas located below sea level or areas with soil that is unable to absorb water well are more susceptible to flooding. As well as areas with steep slopes that allow water to flow quickly to lower elevations. In urban and coastal areas, where dense populations and infrastructure are located, the impacts are particularly pronounced, driving demand for more effective and innovative approaches to disaster management (Remondi, Burlando, & Vollmer, 2016).

These increasingly frequent and intense floods pose a serious threat to the country's population, infrastructure, and economy, particularly in coastal and urban areas. This has prompted an urgent need for more accurate and reliable prediction models to mitigate the effects of these natural disasters (Wannewitz & Garschagen, 2021). The increasing incidence of flooding has prompted governments to re-evaluate existing disaster management strategies and seek solutions that can integrate state-of-the-art technology for more accurate predictions (Martel et al., 2024).

Traditional hydrological models, while valuable, often struggle to fully capture the complex interactions of factors that trigger flooding events in Indonesia's diverse geographic landscape (Peel & Blöschl, 2011). These limitations include reliance on static parameters and poor ability to adapt to changing environmental conditions and climate change impacts (Mojaddadi et al., 2017). Alternatively, the use of machine learning techniques, such as Long Short-Term Memory (LSTM) and Random Forest algorithms, offers a more adaptive and dynamic approach (Puspasari et al., 2023). LSTM is very effective for temporal data, while Random Forest is very good at feature selection and structured data management (Priscillia, Schillaci, & Lipani, 2021). This study aims to utilize the sophistication of LSTM and Random Forest to predict flood events in five flood-prone provinces in Indonesia, namely Jakarta, West Java, Central Java, East Java, and Banten. We use comprehensive data from Petabencana.id which includes geospatial and temporal disaster information from January 2018 to February 2024 to develop a predictive framework that can forecast the timing and probability of future floods..

Petabencana.id is an innovative platform that leverages crowdsourcing technology in Indonesia to collect and disseminate real-time disaster information, which is critical in managing emergency response in a country that frequently experiences natural disasters. The platform allows users to report their actual conditions through social media and live applications, which are then integrated and displayed on an interactive map to provide a comprehensive picture of the evolving situation (Hidayat, 2020). With a focus on improving communication and coordination, Petabencana.id supports government agencies and non-governmental organizations in making quick and effective decisions during disasters. The platform not only raises public awareness of disaster risks in their locations but also facilitates better collaboration between communities and emergency management agencies. Petabencana.id has geospatial disaster information. The geospatial data used includes information about specific geographic locations related to disasters, which is critical in disaster analysis and planning. This research aims to produce a robust and real-time predictive model that can help disaster management agencies and

local governments in mitigating flood risks and preparing for early response. By combining historical flood data, environmental variables such as rainfall, and socio-economic factors, our model is expected to not only predict flood events but also when they are likely to occur. The implementation of this model is expected to improve the preparedness and effectiveness of early warning systems, as well as facilitate faster and more accurate decision-making in emergency situations.

The developed model is designed to be integrated into real-time flood monitoring systems, which enhances the capabilities in preparedness, early warning systems, and decision-making. The model seeks to combine various data sources and the latest technologies to form a system that can dynamically adapt to changing conditions. By understanding and analyzing historical patterns and current conditions, the model aims to provide accurate early warnings, allowing for better preventive measures and faster responses in the face of flood disasters, which can ultimately save lives and reduce economic damage.

METHOD

The primary dataset for this study was derived from Petabencana.id, a platform that aggregates crowdsourced disaster reports, with a particular emphasis on floods, which dominate the dataset. Between January 2018 and February 2024, the platform collected crucial geospatial, environmental, and temporal data on various disaster types across Indonesia, including wind events, earthquakes, haze, fire, and volcanic activity. Notably, floods accounted for over 30,000 reports, significantly outnumbering other disaster types, with wind and earthquakes following at just over 4,200 reports each, and lower occurrences of haze, fire, and volcanic activity (Pavani & Malla, 2024). This substantial disparity highlights the prevalence of floods in Indonesia, positioning them as a key focus of this research. Predictive modeling, therefore, centers on flood events for disaster management and preparedness, with particular attention given to the five provinces with the most comprehensive disaster data: Jakarta, West Java, Central Java, East Java, and Bangka Belitung.

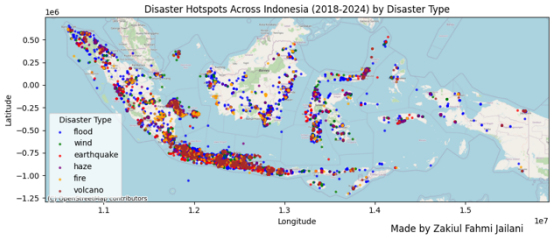

Fig. 1. Spatial Distribution of Disaster Events Across Indonesia (2018–2024)

In addition to the flood reports, Petabencana.id provided flood depth data, which directly relates to the severity of flood events. Flood depth measurements were crucial for understanding the magnitude and potential impact of each flood, and this variable served as the primary environmental factor for modelling flood predictions in this study. While other environmental data, such as air quality, fire distance, volcanic signs, accessibility and structural failures, and visibility and impact data, were available, these variables are not directly linked to flood events. Therefore, they were not incorporated into the flood prediction model. Instead, the focus remained on flood depth as the key environmental input for forecasting the likelihood and timing of future flood events across the top five provinces in Indonesia.

The dataset also incorporated critical temporal features to model the recurrence and cyclic nature of flood events, specifically tailored to the flood occurrences. These features included:

a. Year, month, and day of each reported disaster: By extracting these temporal attributes from the timestamp of each flood event, the model could account for yearly, monthly, and daily variations in flood occurrence.
b. Seasonality: Flood events were categorized based on Indonesia's two primary seasons. The rainy season,

which spans from November to March, typically sees higher flood activity, while the dry season, from April to October, experiences fewer floods. This seasonal categorization provided the model with a deeper understanding of flood patterns relative to the time of year.

These temporal variables were essential for the model to detect and learn patterns associated with Indonesia's distinct wet and dry seasons. By incorporating these, the model's predictive power was enhanced, allowing it to capture not just the geographic but also the temporal dynamics of flood events, crucial for anticipating future occurrences.

In the complex dataset preprocessing process, essential steps were taken to enhance data quality and ensure consistency, crucial for accurate disaster analysis. Initially, rows marked as training data were removed, focusing solely on actual disaster events. To address the issue of missing values that could negatively impact the model, various imputation techniques were tailored to the distribution characteristics of each variable.

Specifically, missing flood depth values were imputed using the median to maintain the integrity of flood severity analysis, as this method is less influenced by outliers. For the categorical variable such as accessibility failure, which indicates access challenges in flood-affected areas, mode imputation was used to accurately reflect the most common category. Other variables describing disaster impact and environmental conditions, such as data condition, structure failure, and data visibility, were also imputed using the mode to fill gaps and enhance model accuracy.

Furthermore, additional environmental and spatial variables like air quality, distance to fires, signs of volcanic activity, and evacuation figures were imputed using the mean, mode, or median, depending on what was most suitable for their specific data distributions. Although these variables do not directly relate to flood prediction, their imputation was vital for maintaining the dataset's overall integrity for broader disaster analysis. Similarly, temporal and regional data such as region code, city, and season were also imputed to maintain consistency across time and different regions.

This meticulous imputation process ensured that the dataset was comprehensive and primed for training machine learning models, preserving the quality of the input features without compromise.

This study employed a hybrid machine learning approach, combining Long Short-Term Memory (LSTM) networks and Random Forest algorithms to predict flood occurrences based on geospatial and temporal data from Petabencana.id. The methodology leveraged the strengths of both models: LSTM networks, known for capturing temporal dependencies in sequential data, and Random Forests, renowned for robust feature importance analysis in structured datasets. The LSTM network was chosen for its capability to learn long-term dependencies in time-series data, making it ideal for modeling the temporal progression of flood events. The architecture featured multiple hidden layers, with each layer comprising LSTM cells, enabling the model to capture both short-term and long-term patterns in the input data. The final output of the LSTM was passed through a fully connected layer, producing a flood likelihood score based on the latest data inputs. Complementing the LSTM, a Random Forest classifier was used to determine the most important features contributing to flood risk. By constructing multiple decision trees, the Random Forest model assessed the relative importance of variables, such as flood depth, accessibility failure, and other environmental data. This provided insights into which features had the highest influence on flood prediction, while also complementing the temporal insights from the LSTM model. The dataset was divided into training (70%) and testing (30%) sets.

The LSTM model was trained using the Adam optimizer with a learning rate of 0.001, while the Random Forest model underwent hyperparameter tuning through grid search to achieve optimal performance. Several evaluation metrics, including accuracy, precision, recall, and the F1-score, were used to assess the models. Additionally, the model's predictions were cross-validated by comparing them to actual flood events, particularly the significant flooding in Jakarta in April 2024, following the data cutoff of February 2024. This validation demonstrated the model's ability to predict flood occurrences with strong accuracy, especially in high-risk areas.

RESULTS AND DISCUSSION

Random forest feature importance analysis for flood prediction. In this study, the Random Forest algorithm was employed to analyze feature importance, aiding in the identification of the most critical variables contributing to flood prediction. Random Forest's inherent ability to handle high-dimensional data and its capability to rank features based on their significance made it an ideal choice for this task. The right panel of Figure 2 showcases the feature importance rankings, with flood_depth emerging as the most influential variable. This is consistent with the understanding that flood depth directly correlates with flood severity, making it the most important predictor in flood occurrence modelling.

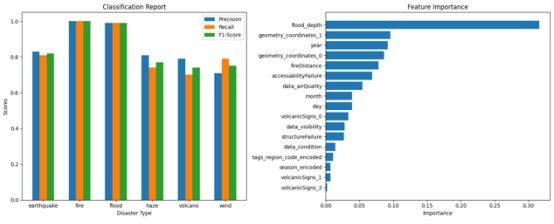

Fig. 1 Classification Performance and Feature Importance for Flood Prediction Model

Machine learning prediction of flood occurrences. The hybrid machine learning model combining LSTM and Random Forest was used to predict future flood occurrences in the five key provinces based on geospatial and temporal data. The LSTM model effectively captured the temporal dependencies in flood occurrences by leveraging year, month, day, and seasonal data. Random Forest complemented this by identifying critical features, such as flood depth, to assess flood risk.

The model achieved a prediction accuracy of 99% for training data, and for testing data, it predicted flood occurrences with an accuracy of 63.6%, as indicated by the test evaluation metrics. The model's performance was validated by predicting the likelihood of flood events within a 25–50 day window after February 2024. This was corroborated by real-world observations when Jakarta experienced significant flooding in April 2024, showcasing the practical utility and reliability of the model. Predictions for other provinces indicated varying risk windows, with longer timeframes predicted for Central Java and East Java, matching local weather conditions and flood patterns.

Random Forest analysis highlighted flood depth as the most significant predictor of future flood events, followed by geographical and environmental factors such as longitude, latitude, and fire distance. These findings reinforce the importance of monitoring flood depth as a key indicator for early flood warning systems. Temporal variables, including seasonality, also played a significant role in the model's predictive performance, demonstrating how seasonal patterns contribute to flood risks in different regions.

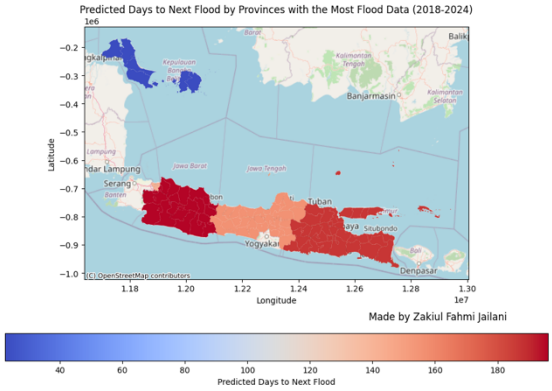

Fig. 2 Choropleth map depicting the predicted number of days to the next flood in provinces with the most flood data recorded between 2018 and 2024.

Figure 3 presents a choropleth map representing the predicted number of days until the next flood event in the top five provinces. The map reveals clear differences in flood risk timelines across these provinces. Bangka Belitung shows the highest risk of flooding, with predictions indicating that floods are likely to occur within the next 40-100 days. Jakarta and Central Java are expected to see floods in the intermediate range of 120-160 days. Finally, West Java and East Java exhibit the longest flood risk timelines, with predictions showing that floods may occur between 170-200 days, possibly during the later months of the year.

CONCLUSION

This study explored the potential of a hybrid machine learning approach, combining Long Short-Term Memory (LSTM) networks and Random Forest algorithms, to predict flood occurrences in Indonesia using comprehensive flood data from Petabencana.id. The LSTM model effectively captured temporal patterns, while the Random Forest model provided valuable insights into the most influential factors in flood prediction, such as flood depth and accessibility failure. This combined approach offers a practical method to estimate when and where floods are likely to occur, contributing to disaster management efforts across key provinces.

Although the model achieved high accuracy on training data (99%) and moderate accuracy on test data (63.6%), there are still opportunities for improvement. The model's predictions aligned closely with real-world events, as demonstrated by its successful prediction of the April 2024 floods in Jakarta. However, predictions for other regions, such as West Java and East Java, indicated longer-term flood risk windows, highlighting the need for further model refinement. The variability of flood patterns across different provinces and environmental conditions presents a significant challenge that requires additional fine-tuning.

Future research could enhance the model's predictive performance by integrating more granular environmental data, such as localized rainfall patterns, river discharge, and land-use changes. Additionally, exploring alternative machine learning algorithms might capture more complex relationships between environmental and socio-economic factors. Expanding the geographical scope of the analysis beyond the five provinces with the most data would provide a more comprehensive national flood risk model.

Moreover, there is potential for real-time integration of this model into flood early warning systems. By incorporating live data streams from platforms like Petabencana.id, the model could be scaled for dynamic updates, enabling disaster management agencies to issue timely warnings and make informed decisions. By addressing these areas for improvement, this approach can better support flood prediction, early warning systems, and disaster management initiatives across Indonesia, ultimately enhancing national preparedness against future flood risks.

BIBLIOGRAPHY

Blöschl, G., & Montanari, A. (2010). Climate change impacts—throwing the dice? *Hydrological Processes, 24*(3),

374-381. https://doi.org/10.1002/hyp.7574

Hidayat, Y. H. (2020). Petabencana.id in Flood Disaster Management: An Innovation in Collaborative Governance-based Early Warning System in Indonesia. *Jurnal Kebijakan dan Administrasi Publik, 24*(1), 2477-4693. https://doi.org/10.22146/jkap.53167

Martel, J.-L., et al. (2024). Assessing the adequacy of traditional hydrological models for climate change impact studies: A case for long-short-term memory (LSTM) neural networks. *EGUsphere, 2024*, 1-44. https://doi.org/10.5194/egusphere-2024-2133

Mojaddadi, H., Pradhan, B., Nampak, H., Ahmad, N., & Ghazali, A. H. bin. (2017). Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. *Geomatics, Natural Hazards and Risk, 8*(2), 1080-1102. https://doi.org/10.1080/19475705.2017.1294113

Pavani, T. D. N., & Malla, S. (2024). A review of deep learning techniques for disaster management in social media: trends and challenges. *European Physical Journal Special Topics*. https://doi.org/10.1140/epjs/s11734-024-01172-9

Peel, M. C., & Blöschl, G. (2011). Hydrological modelling in a changing world. *Progress in Physical Geography: Earth and Environment, 35*(2), 249-261. https://doi.org/10.1177/0309133311402550

Priscillia, S., Schillaci, C., & Lipani, A. (2021). Flood susceptibility assessment using artificial neural networks in Indonesia. *Artificial Intelligence in Geosciences,* 2, 215-222.https://doi.org/10.1016/j.aiig.2022.03.002

Puspasari, R. L., Yoon, D., Kim, H., & Kim, K.W. (2023). Machine Learning for Flood Prediction in Indonesia: Providing Online Access for Disaster Management Control. *Economic and Environmental Geology,* 56(1), 65-73. https://doi.org/10.9719/EEG.2023.56.1.65

Rahayu, H. P., Zulfa, K. I., Nurhasanah, D., Haigh, R., Amaratunga, D., & Wahdiny, I. I. (2024). Unveiling transboundary challenges in river flood risk management: learning from the Ciliwung River basin. *Natural Hazards and Earth System Sciences, 24*(6), 2045-2064. https://doi.org/10.5194/nhess-24-2045-2024

Remondi, F., Burlando, P., & Vollmer, D. (2016). Exploring the hydrological impact of increasing urbanisation on a tropical river catchment of the metropolitan Jakarta, Indonesia. *Sustainable Cities and Society, 20*, 210-221. https://doi.org/10.1016/j.scs.2015.10.001

Wannewitz, M., & Garschagen, M. (2021). Review article: Mapping the adaptation solution space – lessons from Jakarta. *Natural Hazards and Earth System Sciences, 21*(11), 3285-3322. https://doi.org/10.5194/nhess-21-3285-2021