International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 9 Issue 9 September - 2025, Pages: 54-64

Application of GIS and Machine Learning to zoning flood risk from historical data in Vietnam

Nguyen Quynh Anh

Lao Cai High School for the Gifted, Viet Nam E-mail: nguyenquynhanh1112008@gmail.com

Abstract: *Viet Nam as the world's 3rd vertex of the pronounced Southeast Asian Water Triangle that feeds the mighty Mekong River experiences major debilitating flooding in the Central, Northwest, and Mekong Delta regions in the winter-spring months. They decided this as one of the first severe climatic hazards of the region. This study focuses on devising geo-hydrometeorological driven flood risk zones as geo-spatial constructs using records of historical flood events. This has been done using Random Forests for Machine Learning. 11 spatial and climatic variables (elevation, slope, TWI, curvature, distance from the river/drainage, river Curve Number, rainfall extremes, etc.) were processed geospatially through a GIS interface. Validation on Berlin's dataset achieved ~80% accuracy (AUC = 0.80) which is within the accepted range for informative filters, thus, confirming the model's reliability. The decentralized trained model has been implemented in an application that creates binary representations of flood events and scores local defense systems corresponding to the level of vulnerability to flooding. Results demonstrate the combined use of GIS and machine learning as a powerful conceptual and practical tool in systematic spatial planning, proactive alert systems, and climatological adaptations for the region. These models will next be exposed to the real-time flux of geo-spatial data from web-observations across the country to minimize disaster risk.*

Keywords: Flood risk zoning; Machine learning; Random Forest; GIS; Topographic factors; Disaster management; Vietnam.

1. Introduction

In Vietnam, the most damaging natural disaster is flooding, exacerbated by climate change, which causes extreme weather events to become more common and more intense. In the Central Region, Northwest Region, and Mekong Delta, heavy rains and rapid water rise result in prolonged flooding, which disrupts economic activity and stifles socioeconomic growth. Predictive models, as well as warning systems, remain underdeveloped due to a lack of digital infrastructure and data on inundation. The study titled 'Flood Risk Zoning Based on Historical Flood Data and Regional Topographic Information' creates GIS maps to subdivide flood zones and classify high- and low- risk zones to aid use in land planning, preemptive flood warning, and local flood management actions. The study combines historical flood data, and terrain factors comprising DEM, Slope, TWI, Drainage, distance to riparian zones, distance to major roads, roughness curvature, and distance to the controlling river, as well as machine learning to create high-resolution district and commune subdivision flood zoning maps. BERLIN Dataset: as an example, this validates the effectiveness of the approach to flooding based on the high-quality datasets available in Berlin. More than this, in other Vietnam studies where the outcome is risk maps, the focus is often on the outcomes of risk mitigation, whereas in other terrains within Vietnam the focus is on improved mitigation planning and the allocation of social factors.

2. Theoretical Background

The incorporation of AI is re-iterated to have been showcased at the Dartmouth Conference in the year 1956. Since that time, the field of AI has seen the refinement of its perception, especially regarding performing functions quintessential to 'life'; be it learning, reasoning or even decision making. Currently, AI in combination with machine learning, has been embraced in a few varying disciplines owing to an array of developed technological devices, along with 'Big Data'. Healthcare, education, and even disaster management, along with numerous other disciplines have seen an incorporation of the tools. With the increasing severity of issues surrounding climate change, the ability to predict and model natural disasters, particularly floods, has turned to be a necessity in improving AI disaster management systems for efficient mitigation and enhancing community resilience.

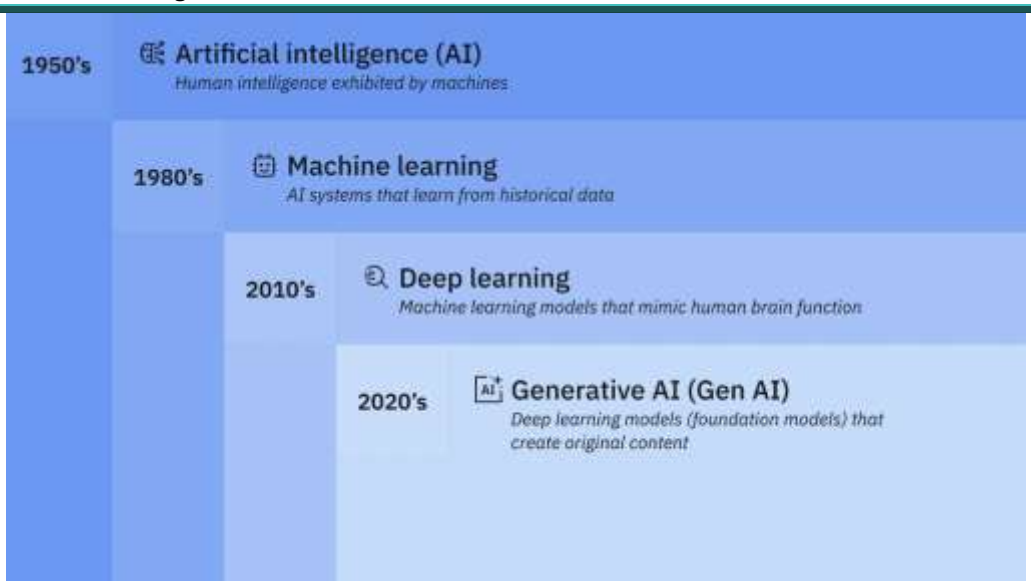

Figure 1. *Introduction to Artificial Intelligence (AI) (GeeksforGeeks.org)*

As a subfield of AI, Machine Learning (ML) systems can enhance their performance through data assimilation and independent functioning, thus eliminating the need for explicit coding. ML now builds accurate risk models to forecast flooding using complex data sets obtained from sensors, satellites, and historical records, thereby improving early warnings, flood zoning, and disaster management.

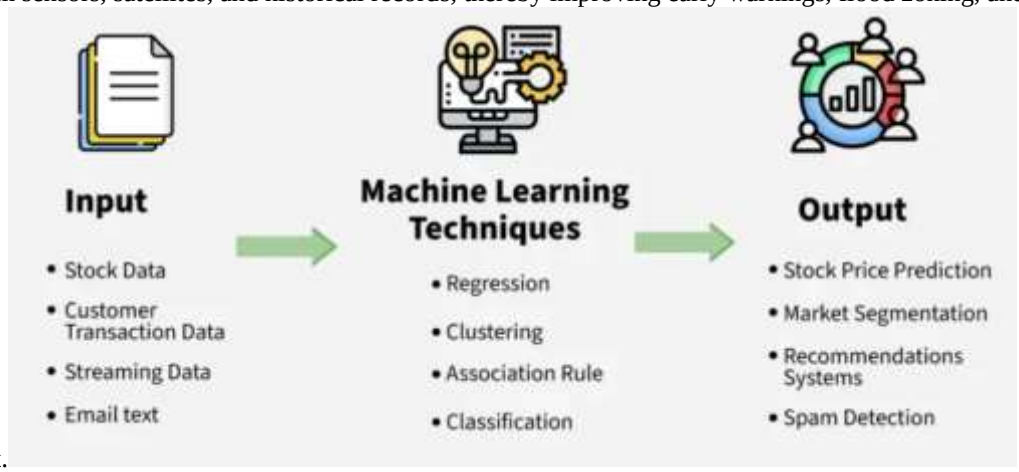

Figure 2*. Machine Learning Techniques (GeeksforGeeks.org)*

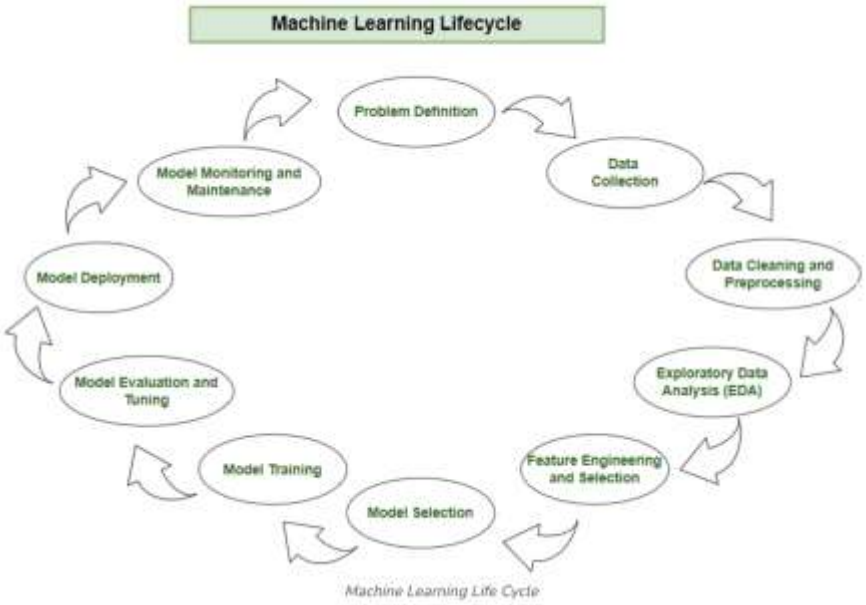

Figure 3. *Machine Learning Life Cycle (GeeksforGeeks.org)*

Supervised learning involves the studying of the associations between the defined attributes and the outcomes of output—usually for classifying (different levels of flood risk) or regressing (estimating economic losses) activities—based on the provided labeled datasets. For instance, on flood risk research, classifiers, including Random Forest, SVM, and Logistic Regression, have managed to high align accuracy in zonal flood forecasting with the provision of elevation, slope, and rainfall (Tehrany et al., 2014). To achieve maximum efficiency, the data must be collected, cleaned, and labeled, which is, often, costly in terms of resources and funds. However, in the case of disaster management, the trained supervised models support in making instantaneous, accurate forecasts accompanied with predictive outcomes.

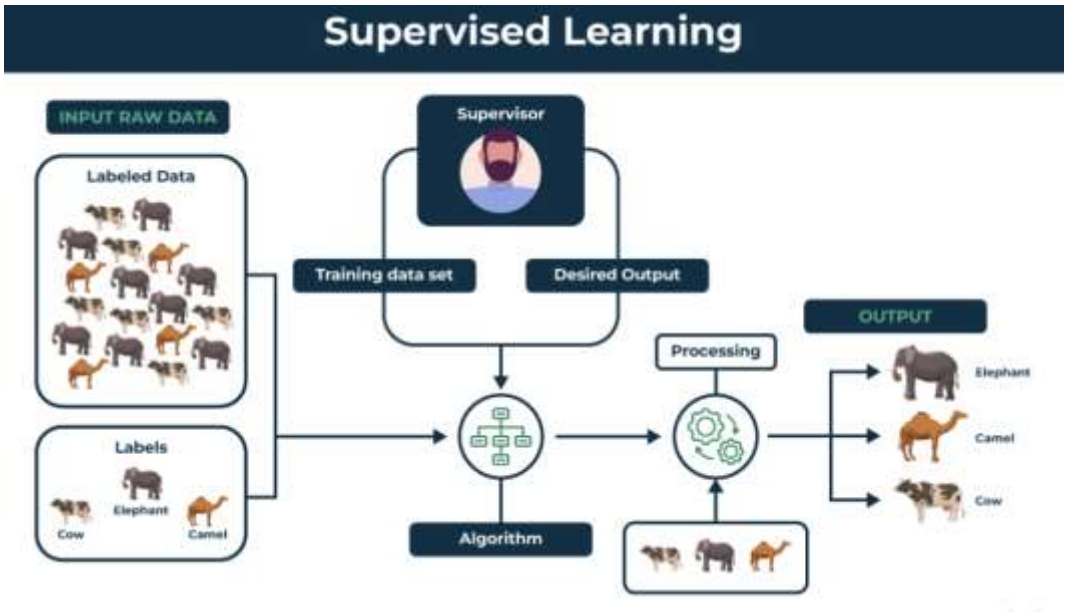

Figure 4. *Supervised Learning (GeeksforGeeks.org)*

Breiman and Cutler's Random Forest tool is a supervised learning method with both classification and regression capabilities. It overcomes overfitting and balances competing objectives by building many decision trees on random subsets and pooling the

predictions. It captures the intricacy, non-linearity, and pointer "Heterophily" features needed for flood risk assessment. Its effectiveness for zoning flood-prone areas, as shown in studies done in China and India (Wang et al., 2020; Li et al., 2021), suggests the still untapped potential in Vietnam for combining relief management with the underlying relief, climatic, and hydraulic data.

3. Materials and Methods

This study concerns itself with the rising flood risks on Vietnam which is mapped under a historical and topographic data zoning system. The Domestic and Foreign Works on GIS, Machine Learning, and Spatial Analysis provided literature that was foundational to the methodology. The data public and government records on DEMs, rainfall, water levels, and historic floods. A pipeline was designed around the predominantly standard procedures to train a classification model to geographically and climatically derived attributes while experimental testing was conducted to gauge precision. A prototype was also created that helps support physical frameworks, enabling risk visualization, local planning, and community data-enhanced adaptability designed for provisional studies.

3.1Design and Implementation of the Machine Learning Model

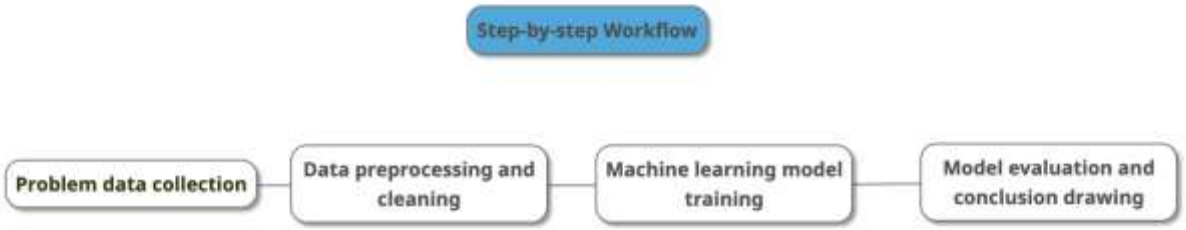

Figure 3.1. Methodology Proposal and Workflow

Step 1: Compilation of Flood Data: Every team member collected and constructed topologically diverse datasets of flash floods whose geographical location is publicly available and contains altimetry, rainfall, slope, land cover, and hydro features. First, out of 84 topographic survey points, individual datasets were created containing attributes to describe floods as occurrences (yes/no) such as elevation, slope, TWI, curvature, aspect, and rainfall extremes, as well as distances to rivers, roads, and drainage systems. Data were kept under Geospatial and Raster data formats to facilitate model training.

Step 2: Data Preprocessing and Cleaning: There was no evidence of duplicates of the dataset, missing values, and class balance. Data augmentation created new points near to the originals. Visualization techniques supplemented by distribution plots and a correlation heatmap clarified variable and inter-variable friendships. These steps ensured that the model feature supports were high quality and devoid of noise.

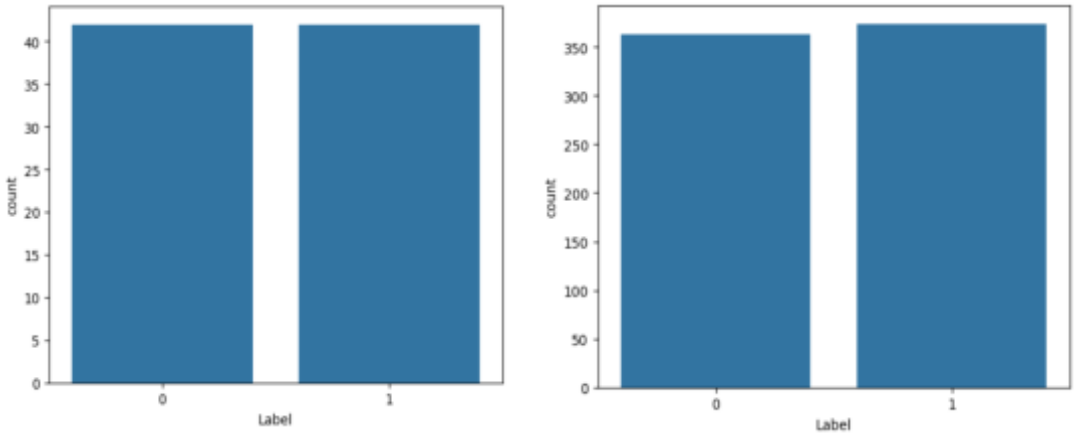

Figure 3.2 Data Distribution Before and After Data Augmentation

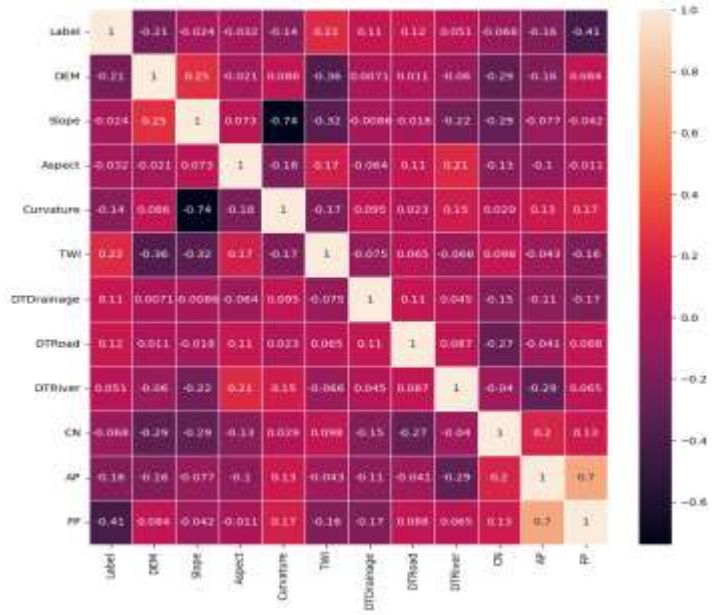

Figure 3.3 Correlation matrix of the dataset before data augmentation

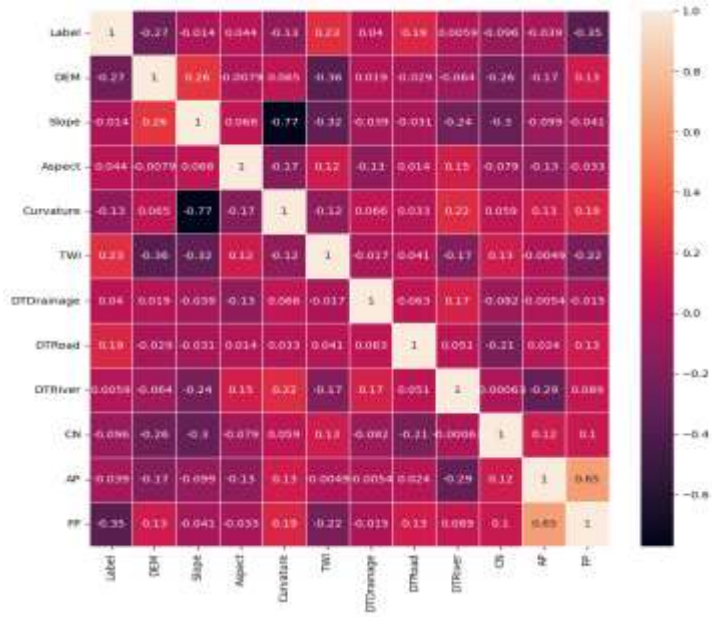

Figure 3.4 Correlation matrix of the dataset after data augmentation

Step 3: Machine Learning Model Training: With respect to the class requirements, the dataset divided was 60-20 training and testing and further divided to 20 validation for the purposes of this class. For class validation, the model was a Random Forest Classifier (i.e. 100 trees, Gini criterion, random_state=42) executed on Google CoLab. The ability to predict flash flood risks based upon geospatial data was validated, and high accuracy on validation was achieved.

Step 4: Model Evaluation and Conclusion: The trained model was tested using accuracy, precision, recall, and F1-score. Results showed 80% accuracy, with ~81% precision and ~79% recall, indicating good effectiveness in identifying flash flood risk.

Misclassifications occurred mostly in areas with intermediate features, highlighting the need for more diverse data to improve performance.

3.2. Software Application Deployment

To translate research outcomes into practical tools for disaster mitigation, an online software application was created to determine the potential of flash floods with the trained machine learning model. The software interface is designed such that it can be used without any training in IT or Machine Learning. The system was implemented using Flask API, which is a lightweight Python framework. Flask, as unbundled for larger systems, requires additional modules, lacks built-in protective mechanisms, and possesses tradeoffs for larger systems. Because of these tradeoffs, Flask was considered reasonable for this project. The interface of the system can be used for the illustration and testing of tools for local disaster mitigation to enhance the 'ground-up' integration approach using HTML, CSS, and JavaScript frameworks for web programming.

4. Results

a. Introduction to the Dataset Used in the Study

Data Source and Scope

For this research project, the dataset was obtained from Berliner Wasserbetriebe, the city-owned corporation responsible for the collection and treatment of water and wastewater. It is recorded in the dataset and is based on fire brigade documents, social media activity, and community-driven information forms. It documents flooding from the years 2005-2017 specifically in the residential and transit zones, also covering the offline and online social media tools. The integrated community and institutional approaches foster the precision of the dataset and add great value to the dataset, especially for the community and for the urban flooding management, considering the rapid urbanization. Their input is vital from both the urban institutional and community pillars to ensure proper governance in the context of rapid tier city modernization/restoration.

Flood Influencing Factors

From the dataset and previous research, the following 11 most relevant factors were obtained and divided into three categories: topography, infrastructure, and hydrometeorology.

- Elevation & Slope: Low-lying and flat areas tend to increase flood risk, while dip slope can influence the speed of the resultant and the stagnation.
- Topographic Wetness Index (TWI) & Curvature: Increments that assess the accumulation of water, and concave forms of land that are capable of being flooded.
- Aspect: An indirect way to assess the flood risk, assists in the evaluation of flow direction.
- Distance to River, Roads, Drainage (DTRiver, DTRoad, DTDrainage): The vicinity to both natural and artificial drainage systems.
- Curve Number (CN): Index that measures the runoff potential through specific soils and land cover.
- Annual Maximum Precipitation (AP) & Frequency of Extreme Precipitation (FP): assess the strength and frequency of rainfall usually derived from the German Railway Metreoric Service radar and stored for multifunction scale and extreme precipitation data.
- The above cited elements work as input variables for the necessary training and 评估 of the artificial intelligence models concerning flood risk zoning.

Evaluation Results of the Trained Model's Performance

Model evaluation indicated that the Random Forest could classify the cleaned and enriched datasets with an accuracy in the region of 80%, satisfying the project objectives. Further evaluation was done using the ROC curve and the AUC index. The ROC curve shows the partitioning of the true positive rate and the false positive rate across varying threshold levels and the TPR and FPR associated with each threshold, with the AUC providing a single value measure of a model's classification ability (0.5 is considered random, >0.7 is acceptable, >0.8 is good, >0.9 is considered excellent). In this case, the model AUC was upward of 0.80 which indicated a good performance with clear class separation. There was a slight augmentation in AUC with data augmentation (0.80→0.81) and a data augmentation shifted ROC curve which was nearer the upper left of the graph, indicative of improved classification performance.

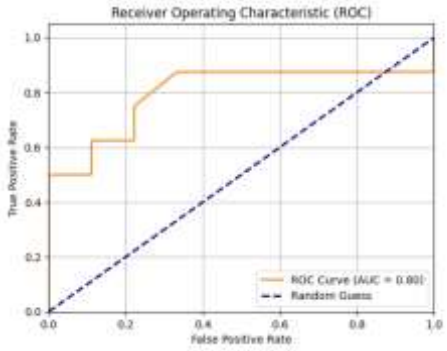

Figure 4.2.1. Model after trained and test on original dataset

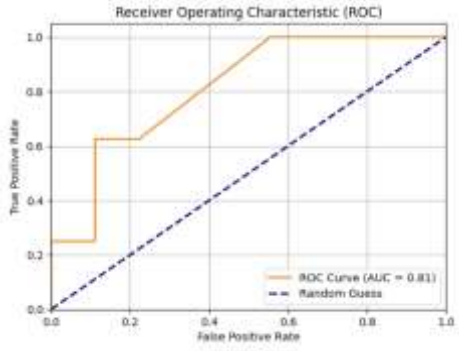

Figure 4.2.2 Model after trained and test on augmented datasets

b. Results of Application Platform Setup and Tuning

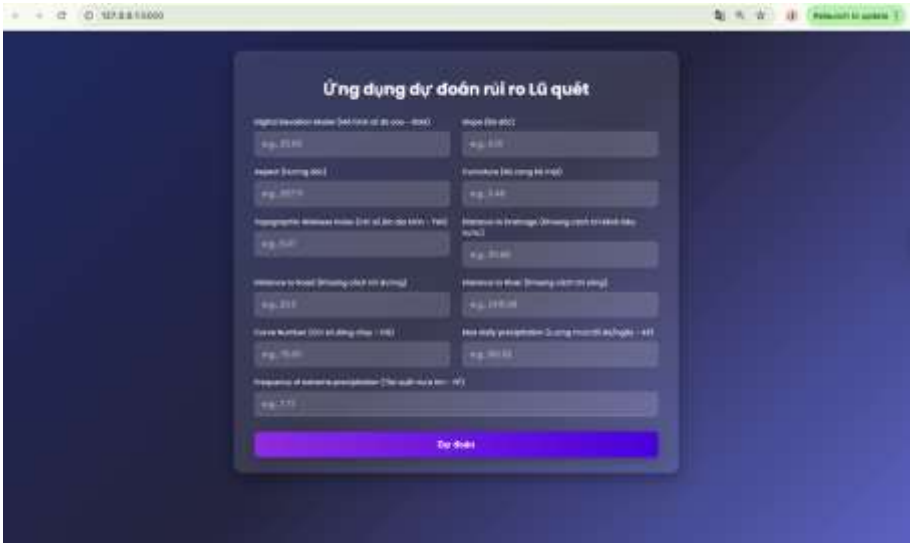

Figure 4.3.1. The Platform Setup and Interface

The software interface is designed to be simple and intuitive, including input fields corresponding to the parameters selected during the model development process. Specifically, the application requires users to provide 11 input data layers that influence the risk of flash floods in the study area, including:

- **Digital Elevation Model (DEM):** a base data layer that reflects the basic topography of the region.

Figure 4.3.2. Digital Elevation Model

- **Slope:** represents the steepness of the terrain, which is a critical factor influencing surface runoff.

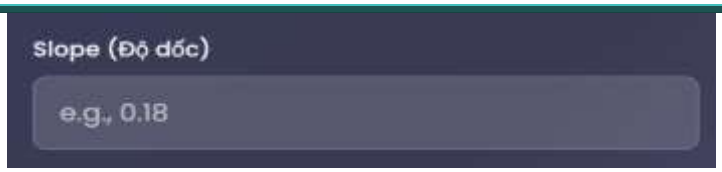

Figure 4.3.3. Slope

- **Aspect:** the main orientation of the terrain's slope, which influences moisture levels and flow direction.

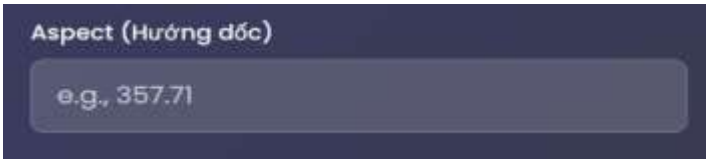

Figure 4.3.4. Aspect

- **Curvature:** indicates the shape of the terrain surface, whether it is concave, convex, or flat.

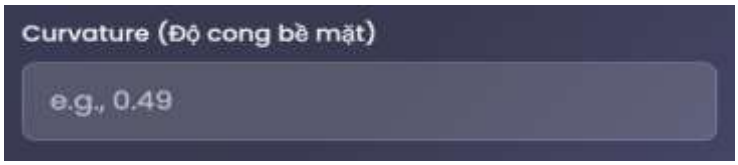

Figure 4.3.5. Curvature

- **Topographic Wetness Index (TWI):** measures the tendency of water to accumulate on the terrain surface.

Figure 4.3.6. Topographic Wetness Index

- **Distance to Drainage:** indicates the proximity to natural drainage channels or streams.

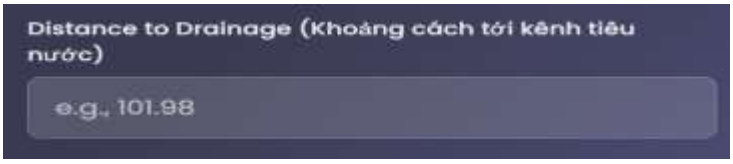

Figure 4.3.7. Distance to Drainage

- **Distance to Road:** reflects the impact of infrastructure on surface runoff and drainage.

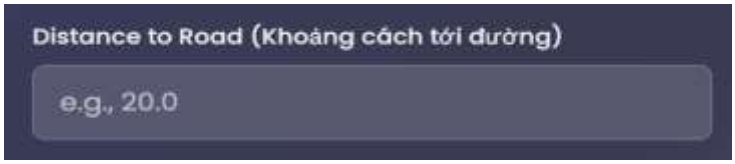

Figure 4.3.8. Distance to road

- **Distance to River:** a crucial factor in flood risk zoning.

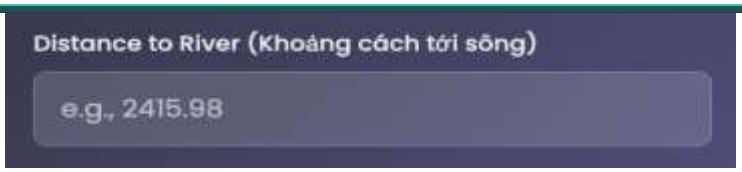

Figure 4.3.9. Distance to River

- **Curve Number (CN):** indicates the soil's infiltration capacity, derived from a combination of soil type and land cover.

Figure 4.3.10. Curve Number

- **Annual Peak Rainfall (AP):** represents the potential for extreme rainfall events in the area.

Figure 4.3.11. Annual Peak Rainfall

- **Frequency of Peak Rainfall (FP):** measures how often extreme rainfall events have occurred historically.

Figure 4.3.12. Frequency of Peak Rainfall

After users input all required data into the corresponding fields, typically as specific values, they can click the “Predict” button. The software will then automatically invoke a pre-trained machine learning model (e.g., using algorithms such as Random Forest) to perform inference. Based on the input data, the model will generate a prediction regarding the likelihood of flash flood occurrence in the specified area.

The results can be displayed in two formats:

- **Classification output** indicating whether the input data corresponds to a flood-prone area (labeled as “Yes” or “No”)
- **Prediction probability** showing the model’s confidence level for the result

This application holds significant potential in supporting government agencies and provincial disaster prevention committees in mountainous areas such as Lào Cai, Yên Bái, and Hà Giang in regularly mapping flash flood risk zones. Additionally, it serves as an intuitive tool for raising public awareness about local disaster risk levels.

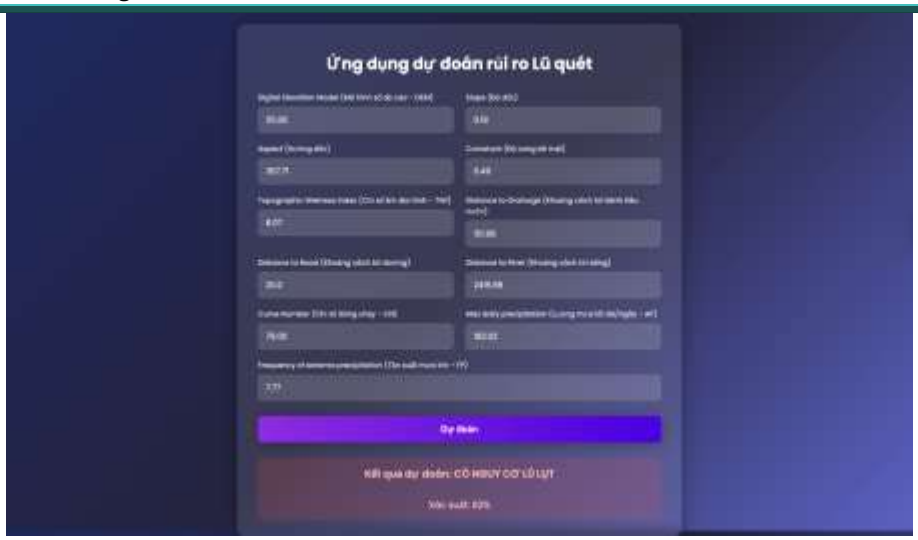

Figure 4.3.13. Software interface after entering all data and generating results

5. Discussion

The Random Forest classification model approximated both the accuracy and the classification AUC metrics at 80% and 0.80 respectively. Support for the model's class discriminating capability was further evidenced through the analysis conducted on both the original and the augmented datasets. These results further enhance the practicality of the proposed approach and underpin its practical Employment. The model along with the historical, topographic, climatic, and socio-environmental data provide evidence supporting the flood risk prediction while its associated predictive variables enhance decision making for disaster management. The model's generalization ability enables it to make predictions in novel areas, and its development is easily scalable and adaptable which enables rapid forecasting. This method applied to areas of Vietnam enables improved forecasting while supporting damage limitation and climate adaptation strategies. Random Forest Supervised learning in this model demonstrates extensive scientific and practical credibility to flood predictive zoning.

6. Conclusion

This study developed a flood and flash flood risk zoning model using the Random Forest algorithm and a combination of terrain and archival flood data. The model has deployed and refined user accuracy and precision while zoning high-risk areas and its precise predictive capability ca user application software designed for disaster mitigation. The model's capability of solving complex spatial data speaks to the efficacy of machine learning as well as Random Forests for flood risk analysis. Then the study will focus on expanding the spatial scope to the flood prone areas of the whole of Vietnam and enhance the data set to include real-time rainfall, land cover, population density and other parameters as well as enhance performance through state-of-the-art algorithms. The study also plans to create a web-based real-time structure with risk maps, assessing user perception and community of the system. In the end, the study aims at a balance of policy and capacity building, contributing to an ecosystem of scales to combat the impacts of climate change.

REFERENCES

1. Intergovernmental Panel on Climate Change. (2022). *Climate change 2022: Impacts, adaptation and vulnerability* (WGII, AR6). Cambridge University Press.
2. World Bank Group & Asian Development Bank. (2021). *Climate Risk Country Profile: Viet Nam*. https://climateknowledgeportal.worldbank.org/ (PDF).
3. World Bank, IFC & MIGA. (2022). *Vietnam Country Climate and Development Report (CCDR)*. (PDF).
4. Luu, C., et al. (2021). GIS-based ensemble computational models for flood susceptibility mapping in Quang Binh Province, Vietnam. *Journal of Hydrology, 599*, 126410.
5. Nguyen, H. D., et al. (2022). A novel hybrid approach to flood susceptibility mapping considering land-use change. *Hydrological Sciences Journal, 67*(7), 1230–1246.
6. Seleem, O., Ayzel, G., de Souza, A. C. T., Bronstert, A., & Heistermann, M. (2022). Towards urban flood susceptibility mapping using data-driven models in Berlin, Germany. *Geomatics, Natural Hazards and Risk, 13*(1), 1640–1662.
7. Seleem, O., Bronstert, A., & Heistermann, M. (2023). Transferability of data-driven models to predict urban pluvial flood

water depth in Berlin. *Natural Hazards and Earth System Sciences, 23*, 809–827.

8. Islam, T., Zeleke, E. B., Afroz, M., & Melesse, A. M. (2025). A systematic review of urban flood susceptibility mapping: Remote sensing, machine learning and modeling approaches. *Remote Sensing, 17*(3), 524.
9. Wahba, M., et al. (2024). Forecasting of flash-flood susceptibility mapping using Random Forest and GIS (Ibaraki, Japan). *Heliyon, 10*(5), e—. (Open access).
10. Wang, Y., Li, Z., Tang, Z., & Zeng, G. (2020). Prediction of urban flood susceptibility using Random Forest and LightGBM models. *Environmental Modelling & Software, 129*, 104759.
11. Rezvani, S. M. H. S., et al. (2024). Smart hotspot detection using GeoAI and Random Forest for flood hazard assessment. *Sustainable Cities and Society, 110*, 105.
12. Razavi-Termeh, S. V., et al. (2025). Improving flood-prone areas mapping using geospatial artificial intelligence (GeoAI): Non-parametric algorithm enhanced by metaheuristics. *Journal of Environmental Management, 360*, 120.
13. Razavi-Termeh, S. V., et al. (2025). Optimized deep learning frameworks (LSTM/RNN + GA/CSA) for flood susceptibility mapping. *Applied Water Science, 15*, Article 2548.
14. Shirzadi, A., et al. (2025). Urban flood susceptibility mapping using deep and conventional learners: DANet vs. SVM/ANN/LR. *Ecological Informatics, 80*, 102528.
15. Rahman, Z. U., et al. (2025). Flood susceptibility mapping using supervised machine-learning models: A multi-model comparison. *Geomatics, Natural Hazards and Risk, 16*(1), 1–22.
16. Li, W., et al. (2023). Assessment of a GeoAI foundation model (IBM-NASA Prithvi) for flood mapping. In *Proceedings of SIGSPATIAL '23*. ACM.
17. Mahdizadeh Gharakhanlou, N., & Hooshyaripor, F. (2022). Spatial prediction of current and future flood susceptibility under climate-change scenarios using ML. *Entropy, 24*(11), 1630.
18. Quang, N. H., et al. (2025). Boosting vs. traditional machine-learning models for flood risk zoning: A case study in Hoa Vang District, Da Nang, Vietnam. *Advances in Space Research, 76*