Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
18/1 (2025), 11-27. DOI: http://dx.doi.org/10.21609/jiki.v18i1.1304

Forest and Land Fire Vulnerability Assessment and Mapping using Machine Learning Method in East Nusa Tenggara Province, Indonesia

Hans Timothy Wijaya* and Aniati Murni Arymurthy†

Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

*E-mail: *hans.timothy@ui.ac.id, †aniati@cs.ui.ac.id*

Abstract

Forest and land fires (FLF) severely damage forest ecosystems and reduce their functionality. Predicting areas prone to fires is crucial for effective management and prevention. Machine learning (ML) has shown potential in this field. By 2022, East Nusa Tenggara (NTT) experienced the highest incidence of fires in Indonesia, with 70,637 hectares burned. This study evaluates NTT's FLF vulnerability using seven ML methods: Gaussian Naive Bayes, Support Vector Machine, Logistic Regression, Artificial Neural Network, Random Forest, Gradient Boosting Machine, and Extreme Gradient Boost (XGB). A geospatial dataset combining NTT's 2022 fire data and fourteen fire-related factors was developed with ArcGIS. Using the Information Gain Ratio for feature selection, twelve key features were identified: Elevation, Slope angle, Slope Aspect, Plan Curvature, Land Cover, NDVI, Distance to Road, Distance to Buildings, Annual Rainfall, Average Temperature, Wind Speed, and Relative Humidity. The XGB model performed best, with AUC values of 0.959 for training and 0.743 for testing. The resulting vulnerability map revealed key fire factors: low elevation, gentle slopes, curved terrain, forest cover, poor vegetation health, human activity, distant firefighting resources, low rainfall, high temperatures, high wind speeds, and low humidity. Recommendations include land management, fire-resistant vegetation, policy enforcement, community education, and infrastructure enhancement.

Keywords: *East Nusa Tenggara, forest and land fires, feature selection, machine learning, mapping*

1. Introduction

Terrestrial ecosystems such as forests play a fundamental role in ecological equilibrium, soil and water conservation, environmental enhancement, carbon sequestration, and oxygen production [1]. Nonetheless, the integrity of forests faces multifaceted threats, including urban expansion, deforestation, natural disasters such as landslides and storms, and forest fires [2]. Forest and land fires (FLF) pose significant hazards to forest ecosystems, often spiraling out of control and causing detrimental impacts on both natural resources and human welfare [3].

As per data from the Ministry of Environment and Forestry [4], Indonesia undergoes a substantial area of FLF amounting to 204.9 thousand hectares in 2022. This area predominantly comprises mineral soil, covering approximately 92.96%, followed by peat soil, accounting for 7.04%. Notably, East Nusa Tenggara (NTT) Province emerged as the region most affected by fires, with 70,637 hectares burned, primarily encompassing savanna, shrublands, and dry agricultural areas intermixed with shrubs.

The repercussions of FLF are profound, leading to the degradation of forest functions such as soil ecology, hydrology, land integrity, and erosion [4-5]. Furthermore, these fires precipitate biodiversity loss, often resulting in species extinction [3]. Given the substantial losses incurred, governmental efforts are continuously directed toward fire prevention and control through various policy frameworks and programs [4].

Accurate prediction of FLF is pivotal for effective mitigation and prevention strategies [6]. Additionally, mapping the vulnerability of regions to FLF is essential for resource allocation and land use planning [7]. Machine learning (ML) methods have been extensively explored for

Table 1. FLF-causal factors [1-2], [5-11].

Category	Specific conditioning factor
Topography	Altitude, Slope Angle, Slope Aspect, Plan Curvature
Hydrology	Distance to Rivers, Topographic Wetness Index (TWI)
Land coverage	Normalized Difference Vegetation Index (NDVI), Land Cover
Meteorology	Rainfall, Temperature, Wind Speed, Relative Humidity
Anthropogenic engineering	Distance to Roads, Distance to Buildings

predicting and mapping FLF vulnerabilities, with tree-based algorithms such as Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGB) demonstrating promising results [1-2], [5-11]. However, the superiority of these methods varies across studies, indicating that predictive performance is contingent upon the unique characteristics and circumstances of the study area. For instance, while some studies report GBM surpassing RF accuracy [8], others demonstrate RF outperforming GBM [1], [11] or XGB exceeding both RF and GBM [2], [7]. These discrepancies underscore the influence of local conditions on model efficacy.

Table 1 presents an overview of the primary factors contributing to FLF, commonly integrated into ML modeling approaches. However, these factors may not be directly applicable for modeling purposes. Thus, various feature selection techniques such as Variance Inflation Factor (VIF), Information Gain Ratio (IGR), Backward Elimination (BE), and RF are essential to identify pertinent factors contributing to FLF. Proper feature selection enhances prediction accuracy.

Based on the abovementioned issues, we have formulated three critical questions for our study: Which feature selection methodologies and ML Model combination shows optimal predictive efficacy in evaluating forest susceptibility within the NTT Province? What underlying factors contribute to the incidence of FLF within NTT Province? What strategic interventions may be posited to mitigate the occurrence of FLF within NTT Province?

To address these questions, we propose utilizing ML techniques to predict and map FLF susceptibility in NTT Province. Specifically, seven ML models—namely Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Network (ANN), RF, GBM, and XGB—are evaluated. Furthermore, in contrast to conventional research approaches, this study analyzes three distinct feature selection methodologies within each ML model. Consequently, this comprehensive analysis aims to determine the optimal combination of feature selection techniques and ML models to achieve superior prediction accuracy.

2. Related Works

Numerous investigations have been conducted across diverse scenarios to evaluate and predict the susceptibility of FLF in global regions. Hidayanto et al. [8] conducted a comparative analysis of three ML algorithms—GBM, RF, and SVM—for constructing forest fire susceptibility (FFS) models in the Pisang Island region of Central Kalimantan. Feature selection was executed utilizing Spearman's Correlation Coefficients and RF methodologies. The empirical findings reveal that GBM performs superior to other ML algorithms, demonstrating the highest predictive capability with an Area Under the Curve (AUC) accuracy of 88.2%.

Mohajane et al. [5] conducted a comparative analysis of five machine learning techniques, namely Multi-Layer Perceptron (MLP), LR, Decision Tree (DT), RF, and SVM, for the modeling and mapping of forest fires in the Hoceima region, located in the northern part of Morocco. Using the Frequency Ratio Analysis method, they performed feature selection on ten factors contributing to forest fires. The findings of this study indicate that the Random Forest method outperformed the other four methods, achieving the highest AUC accuracy of 95.2%.

Rakshit et al. [3] conducted predictive modeling of forest fires within the Montesinho Natural Park, Portugal, employing four distinct ML algorithms: K-Nearest Neighbor (KNN), DT, SVM, and NB. They curated a dataset comprising eight pertinent variables to feed into the ML above models. The assessment of model performance revealed that DT showed significant efficacy, demonstrating an AUC accuracy of 99%.

In the study by Shao et al. [1], four ML techniques, specifically RF, SVM, GBM, and MLP, were employed to assess the potential risk of forest fires across mainland China. A comprehensive dataset comprising 20 forest fire-related causal factors was assembled and refined utilizing the RF algorithm. Performance assessment through AUC analysis revealed that the RF model exhibited superior predictive capability, attaining the highest AUC accuracy of 95.1%.

Qiu et al. [9] developed an RF model employing 23 climate and land surface variables inputs. The significance of individual variables

was assessed utilizing the Shapley value method. Empirical findings indicate that the proposed approach yields notably high AUC accuracy, specifically achieving a rate of 98%.

Moayedi and Khasmakhi [10] conducted a study on wildfire susceptibility assessment employing two enhanced ML algorithms, Biogeography-based Optimization-Artificial Neural Network (BBO-ANN) and ANT Colony Optimization-Artificial Neural Network (ACO-ANN). In this research, fifteen variables was chosen utilizing the frequency ratio analysis technique. The findings revealed that the BBO-ANN approach achieved the highest performance, exhibiting an AUC accuracy of 84%.

In the study conducted by Tan and Feng [11], an analysis was undertaken to assess the effectiveness of RF, SVM, and GBM techniques for mapping forest fire risk areas within Hunan Province, China. The VIF technique was employed in this investigation to identify and select 22 pertinent factors associated with forest fire occurrence. Empirical findings revealed that the RF model exhibited superior performance to SVM and GBM, achieving a notable AUC accuracy of 97.2%.

Akinci and Akinci [2] assessed forest fire susceptibility within the Antalya district of Turkey utilizing four machine learning methodologies: ANN, GBM, RF, and XGB. Furthermore, the VIF approach was employed in this study to determine relevant factors associated with forest fires, as identified by the researchers. The observed findings highlight that the XGB model exhibits superior performance, achieving an AUC accuracy rate of 97.5% compared to other models.

Based on prior studies above, it has been observed that Decision Tree-based methodologies show superior predictive accuracy in scenarios concerning the susceptibility prediction of FLF. Notably, the RF model has demonstrated the highest accuracy in various study regions, such as Northern Morocco and China [1], [5], [11]. Additionally, the GBM model exhibited superior predictive accuracy in the research conducted by Hidayanto within Pisang Island, Central Kalimantan [8]. Conversely, the XGB model delivered the most accurate predictions in studies conducted by Seddouki [7] and Akinci [2]. These investigations highlighted that the XGB model consistently outperformed other Decision Tree-based models regarding predictive accuracy.

In a specific case study conducted on Pisang Island, Central Kalimantan [8], the findings indicated that the GBM model exhibited the highest accuracy, with an AUC value of 0.882, followed by RF and SVM with AUC values of 0.871 and 0.869, respectively. In contrast, research conducted by Tan and Feng in China's Hunan Province [11] employed the same three ML models as utilized by Hidayanto. However, their results indicated that the RF model outperformed the others regarding predictive accuracy, followed by GBM and SVM. The AUC accuracy values for this study's RF, GBM, and SVM models were 0.972, 0.958, and 0.953, respectively. These findings highlight the influence of the specific characteristics and environmental conditions of the study area on the predictive performance of the models, thereby leading to variations in prediction outcomes across geographically diverse regions.

3. Study Area

The study area in this research is located in NTT Province, Indonesia, highlighted with red line in Figure 1. NTT spans from 80° to 120° South Latitude and 1180° to 1250° East Longitude, covering an extensive landmass of 46,452.38 km². Geographically, NTT shares borders with the Flores Sea to the north, the Indian Ocean to the south, the State of Timor Leste to the east, and West Nusa Tenggara Province to the west. Within NTT Province are 21 districts and one city distributed across seven principal islands: Flores, Sumba, Timor, Alor, Lembata, Rote, and Sabu.

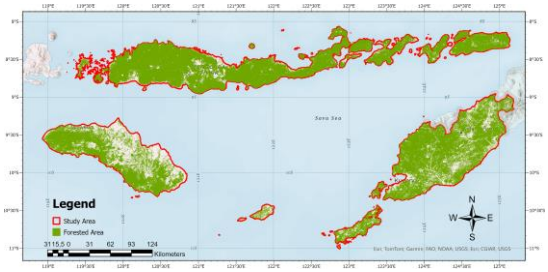

Figure 1. Study area: NTT Province is highlighted in red, with forest areas colored green. The scale bar (in kilometers) is located at the bottom left, and the north arrow is at the bottom right.

NTT Province experiences a dichotomous climate characterized by two distinct seasons: a dry period from June to September and a wet season from December to March. Due to its proximity to Australia, atmospheric currents originating from Asia undergo desiccation en route to NTT, resulting in diminished moisture content and reduced precipitation compared to regions closer to Asia. Consequently, drought conditions persist for prolonged durations within NTT Province.

Temperature profiles in NTT exhibit notable variations, with the mean highest recorded

temperature in 2022 reaching 32.8°C and the lowest at 16.2°C. However, the average temperature remains relatively elevated, between 27°C and 28°C. Regarding precipitation, NTT experienced an average of 145 rainy days in 2022, with the highest occurrence observed in Manggarai Regency and the lowest in East Flores Regency [12].

4. Methodology

This research consists of several stages. Initially, we gathered the locations where FLF happened in the NTT province along with geographical information related to that region. These FLF locations were marked with a spatial resolution of 375 meters, meaning each point on the map represents a square area on the ground measuring 375 meters by 375 meters. Typically, data points in nearby regions can show how the fire spreads. Subsequently, this data underwent processing via ArcGIS Pro software. Specifically, the FLF data points (serving as target variables) and causal factors contributing to forest fires (as predictor variables) were collated and processed, then exported in CSV format and transposed into Python data frames for application in subsequent feature selection and ML modeling phases. In order to ensure a comprehensive representation of data distribution, we randomly selected 80% of our dataset (6624 data points, with 3309 for non-FLF and 3315 for FLF classes) for training. The remaining 20% (1656 data points, with 831 for non-FLF and 825 for FLF classes) will be used for testing.

The feature selection phase encompasses multicollinearity analysis, entailing the computation of VIF values to identify multicollinearity among predictor variables. Predictor variables exhibiting VIF values exceeding five are excluded from subsequent feature selection stages. Following this, three feature selection methodologies—IGR, BE, and RF—are concurrently employed to eliminate irrelevant features during the modeling phase. Consequently, three distinct feature groups (corresponding to each feature selection method) are generated for employment in the modeling phase.

The modeling process involves the utilization of seven distinct ML algorithms—GNB, SVM, LR, ANN, RF, GBM, and XGB. Each ML model is constructed utilizing the feature above sets as predictor variables and evaluated based on accuracy metrics. The feature groups producing the most accurate model is chosen for the subsequent ML model evaluation or comparison stage. Hyperparameter tuning for each ML model is conducted using the "GridSearchCV" technique to ascertain optimal hyperparameter configurations.

Comparative analyses of prediction outcomes from the seven models are conducted via statistical methods and ROC curves. The ML model exhibiting the highest AUC value is employed for mapping FLF vulnerability, facilitating predicting vulnerability classes for individual pixels within the NTT province region. This mapping is subsequently categorized into five risk classes: very low, low, moderate, high, and very high. Finally, the FLF vulnerability map are compared with relevant causal factor raster for further analytical exploration. The insights obtained from this analysis inform strategies for addressing FLF challenges within the NTT province.

4.1 Data Preparation

4.1.1 Data collection

Spatial data were collected from diverse sources to construct the FLF dataset in this study. Initially, the Indonesian Earth Map provided by the Indonesian Geospatial Information Agency was utilized to delineate the territorial boundaries of NTT Province, serving as the designated study area. Subsequently, a Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) with a spatial resolution of 30 meters was acquired from the United States Geological Survey (USGS) website to analyze topographic characteristics within the NTT Province. This SRTM DEM dataset, released by the USGS on September 23, 2014, was employed for topographic analysis. Moreover, Landsat 8 OLI/TIRS imagery, covering September 2021 to September 2022, was sourced from the USGS website to ensure comprehensive satellite coverage of NTT Province with minimal cloud cover (below 5%).

Additionally, over 4000 FLF data points in NTT Province for the year 2022 were sourced from the NASA Fire Information for Resource Management System (FIRMS) website, derived from the SUOMI satellite's Visible Infrared Imaging Radiometer Suite (VIIRS) with a spatial resolution of 375m. Furthermore, in July 2023, data on road networks, building locations, and river networks were obtained from OpenStreetMap, facilitating analyses on the

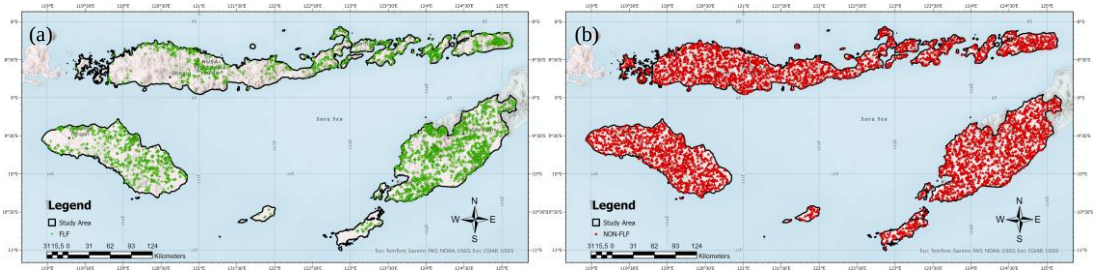

Figure 2. FLF inventory map: (a) FLF points (green) and (b) non-FLF points (red). The map shows 4140 FLF points extracted from NASA FIRMS data for 2022 and 4140 non-FLF points randomly generated across shrubland, herbaceous vegetation, agricultural land, and forested areas. The scale bar in kilometers is located at the bottom left, and the north arrow is at the bottom right.

proximity of fire occurrences to key infrastructural and natural features. Next, we sought the most recent climate data available. We acquired the 2022 rainfall data from the Climatic Research Unit (CRU) dataset of the University of East Anglia. Additionally, we sourced the 2020 data for temperature, wind speed, and relative humidity from NASA Prediction of Worldwide Energy Resource (POWER). Finally, land cover data specific to NTT Province was obtained from Copernicus Global Land Services, released in 2019, and utilized for analyses concerning land use patterns within the region.

In our FLF analysis, we didn't explicitly account for time-related factors. This choice was made because our study prioritizes examining variables like topography, hydrology, land cover, weather, and human interventions, which are commonly studied in this field [2-3], [4-7]. Nevertheless, we acknowledge that considering temporal and seasonal patterns could offer valuable insights. Future research could explore incorporating these factors for a more comprehensive analysis.

4.1.2 Data preprocessing

Spatial data processing is crucial for effective feature selection and modeling. In this study, we utilized ArcGIS Pro software to process spatial data. The FLF dataset was constructed by integrating 4140 FLF data points obtained from NASA FIRMS as feature points. These points were categorized as "1," denoting FLF presence at respective coordinates. Additionally, randomly generating non-FLF data points, especially in a 1:1 ratio, is a well-established and validated method in spatial modeling research. It ensures the balanced representation of both classes and enhances the reliability of model evaluation and susceptibility assessment [1-2], [5], [7-8], [10-11]. This process involved the creation of 4,140 random non-FLF data points, with a buffer of 500 meters between each other, which were randomly distributed across different land cover types, such as shrubland, herbaceous vegetation, agricultural land, and forest, to ensure adequate representation in areas prone to FLF occurrence [1]. The randomly generated points were labeled "0" to signify non-FLF data. Consequently, the study encompassed a total of 8280 data points. The spatial distribution of FLF and non-FLF points is illustrated in Figure 2.

Four topographic factors were identified with a resolution of 30 meters. The SRTM DEM raster utilizing SRTM DEM data with a spatial was subsequently transformed into the WGS 1984 UTM Zone 51S coordinate system, specifically chosen for its compatibility with the NTT Province area. The surface toolset within ArcGIS Pro's spatial analysis tools facilitated the extraction of four causal factors: Elevation, Slope Angle, Slope Aspect, and Plan Curvature.

Furthermore, raster computations were conducted to acquire topographic wetness index (TWI) variable. Initially, the unclassified slope degree raster is converted into radians ("TanSlope") utilizing the "raster calculator" tool, as delineated by equation (1) [13].

$$TanSlope = Tan\left(Slope \times \frac{\pi}{180}\right) \quad (1)$$

Subsequently, a further conversion employing the "raster calculator" tool, as specified by equation (2) [14], is undertaken to eliminate zero values.

$$\beta = con(TanSlope == 0{,}0.001, TanSlope) \quad (2)$$

The subsequent phase entails the projection of the 30 m SRTM DEM raster into the WGS 1984 UTM Zone 51S coordinate system, with a cell resolution of 30 m. This DEM raster is then subjected to processing utilizing the "fill" tool from the "hydrology toolset," followed by additional processing using the "flow direction" and "flow accumulation" tools. The results of the Flow Accumulation raster are then rescaled to the cell size using equation (3) [13].

Table 2. FLF causal factors and their classes.

Category	Classes
Elevation (m)	(1) [<0]; (2) [0, 400]; (3) [400, 880]; (4) [800, 1200]; (5) [1200, 1600]; (6) [1600, 2000]; (7) [>2000]
Slope Angle (degree)	(1) [< 5]; (2) [5, 10]; (3) [10, 15]; (4) [15, 20]; (5) [20, 25]; (6) [25, 30]; (7) [30, 35]; (8) [35, 40]; (9) [40, 45]; (10) [>45]
Slope Aspect	(1) North; (2) Northeast; (3) East; (4) Southeast; (5) south; (6) southwest; (7) west; (8) northwest
Plan Curvature	(1) concave [< -0.05]; (2) flat [-0.05, 0.05]; (3) convex [> 0.05]
TWI	(1) [0, 5]; (2) [5, 7.5]; (3) [7.5, 10]; (4) [10, 12.5]; (5) [> 12.5]
Land Cover	(1) Shrubland; (2) Herbaceous Vegetation; (3) Cropland; (4) Built-up; (5) Bare Soil; (6) Water Bodies; (7) Herbaceous Wetland; (8) Forest
NDVI	(1) [< 0]; (2) [0, 0.33]; (3) [0.33, 0.66]; (4) [>0.66]
Distance to Roads (m)	(1) [0, 200]; (2) [200, 400]; (3) [400, 800]; (4) [800, 2000]; (5) [>2000]
Distance to Rivers (m)	(1) [0, 200]; (2) [200, 400]; (3) [400, 800]; (4) [800, 2000]; (5) [>2000]
Distance to Buildings (m)	(1) [0, 200]; (2) [200, 400]; (3) [400, 800]; (4) [800, 2000]; (5) [>2000]
Annual Rainfall (mm)	(1) [0, 51]; (2) [51, 99]; (3) [99, 141]; (4) [141, 188]; (5) [188, 235]; (6) [235, 293]; (7) [293, 344]; (8) [344, 392]; (9) [392, 434]; (10) [434, 470]
Average Temperature (Celsius)	(1) [25.388, 25.846]; (2) [25.847, 26.222]; (3) [26.223, 26.530]; (4) [26.531, 26.782]; (5) [26.783, 26.988]; (6) [26.989, 27.157]; (7) [27.158, 27.296]; (8) [27.297, 27.465]; (9) [27.466, 27.671]; (10) [27.672, 27.923]
Average Wind Speed (m/s)	(1) [1.868, 2.164]; (2) [2.165, 2.385]; (3) [2.386, 2.555]; (4) [2.556, 2.687]; (5) [2.688, 2.858]; (6) [2.859, 3.079]; (7) [3.080, 3.366]; (8) [3.367, 3.738]; (9) [2.739, 4.220]; (10) [4.221, 4.846]
Average Relative Humidity (mm/day)	(1) [72.695, 73.949]; (2) [73.950, 74.696]; (3) [74.697, 75.140]; (4) [75.141, 75.404]; (5) [75.405, 75.561]; (6) [75.562, 75.825]; (7) [75.826, 76.269]; (8) [76.270, 77.016]; (9) [77.017, 78.271]; (10) [78.272, 80.382]

$$\alpha = (FlowAccumulation + 1) \times CellSize \quad (3)$$

Ultimately, to derive the TWI raster, the flow accumulation ($\boldsymbol{\alpha}$) and slope values in radians ($\boldsymbol{\beta}$), acquired from the previous calculations, are employed. The computations are performed by equation (4) [13].

$$TWI = ln\left(\frac{\alpha}{\beta}\right) \quad (4)$$

Moreover, to generate the raster Normalized Difference Vegetation Index (NDVI), Landsat 8 OLI/TIRS satellite imagery obtained from the USGS data source was employed. NDVI was computed by measuring the normalized difference between light intensities in the red and near-infrared (NIR) bands. In Landsat 8 imagery, the red band corresponds to band 4, while the NIR band corresponds to band 5. Hence, the NDVI calculation process was executed utilizing the raster calculator tool by applying equation (5) [15].

$$NDVI = \frac{NIR-red}{NIR+red} = \frac{Band5-Band4}{Band5+Band4} \quad (5)$$

Three additional raster causal factors, including Distance to Roads, Distance to Rivers, and Distance to Buildings, were obtained by applying the Euclidean distance tool within the spatial analysis functionalities of ArcGIS Pro. This tool facilitates the computation of distances from individual raster pixels to specified geographical features. As outlined in the preceding subsection, the road, river, and building data were sourced from the OpenStreetMap shapefile.

Moreover, the procedure for handling climatic data, such as raster-based Annual Rainfall, Average Temperature, Average Wind Speed, and Average Relative Humidity, encompasses several sequential stages. Initially, the climatic data is imported into the ArcGIS Pro software environment. Subsequently, it was projected into the WGS 1984 UTM Zone 51S coordinate reference system and spatially delineated based on the designated climatic parameters. The "cell statistics" tool is employed to compute the sum or average statistics. The subsequent step entails converting raster data into point features utilizing the "raster to point" functionality. These point features are then analyzed via the "inverse distance weighting" method to generate a raster representation of the climatic dataset.

Consequently, ArcGIS Pro software incorporated Raster land cover data from Copernicus Global Land Services. This raster dataset necessitated no further preprocessing as the land cover information retrieved for this study was readily available and pre-classified into distinct classes. Furthermore, these causal factor raster were classified into distinct classes, as outlined in Table 2 and shown in Figure 3.

Subsequently, following the classification process, raster values were extracted from the fourteen aforementioned causal factor raster for each coordinate point within the FLF inventory. This data was subsequently exported as a comma-separated values (CSV) file for employment in data modeling. Furthermore, the fourteen raster-

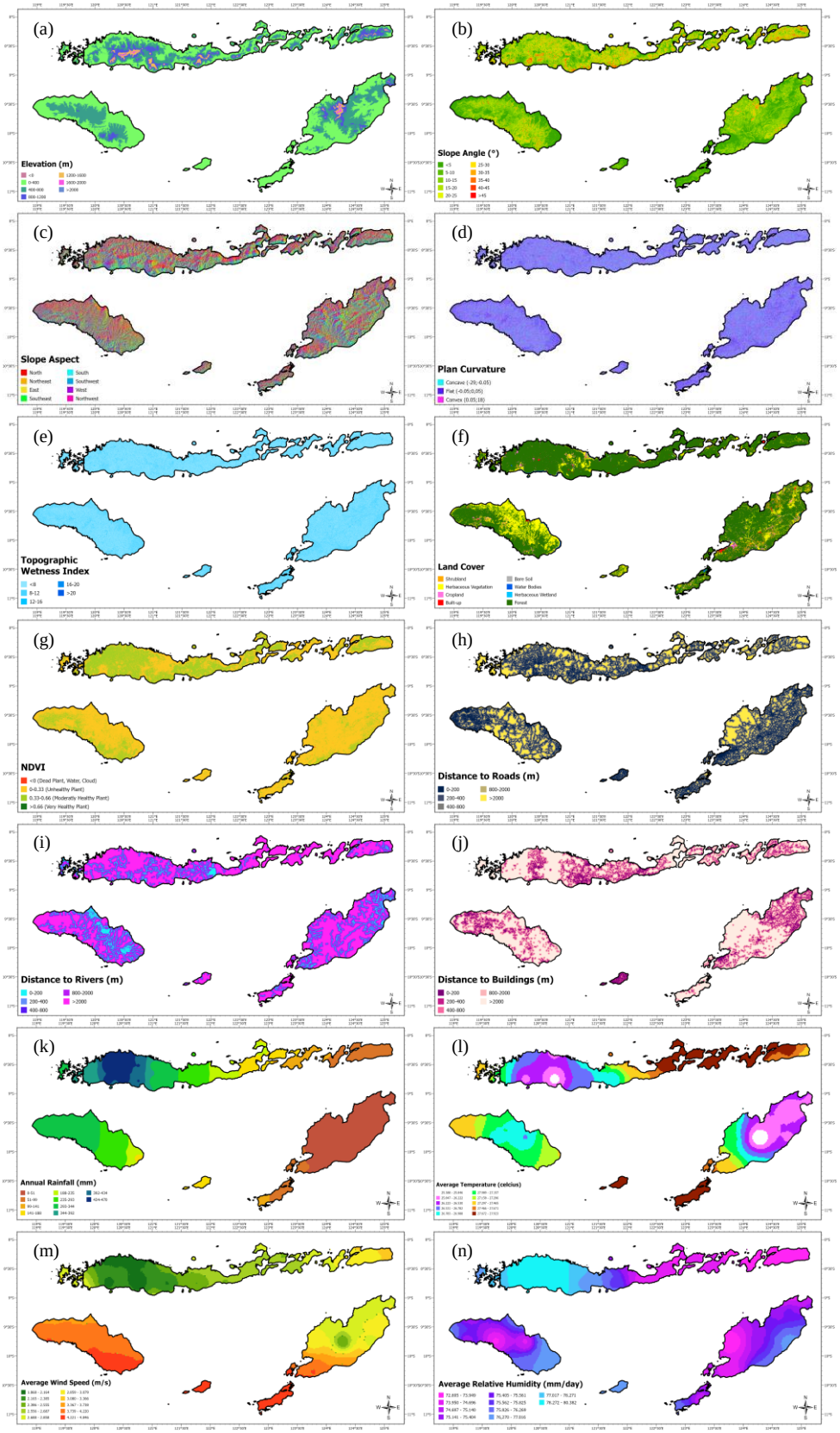

Figure 3. FLF influencing factors: (a) Elevation; (b) Slope Angle; (c) Slope Aspect; (d) Plan Curvature; (e) TWI; (f) Land Cover; (g) NDVI; (h) Distance to Roads; (I) Distance to Rivers; (j) Distance to Buildings; (k) Annual Rainfall; (l) Average Temperature; (m) Average Wind Speed; (n) Average Relative Humidity. The color indicates the raster value, and the label is at the bottom left. The north arrow is located at the bottom right.

based causal factors were exported in the Tag Image File Format (TIFF) for the following mapping tasks.

4.1.3 Data normalization

In this study, we employed the standard score method [16] for normalizing the predictor variables. This method, also known as $\boldsymbol{z}$-score normalization, is computed using equation (6).

$$x_{norm} = \frac{x-\mu}{\sigma} \quad (6)$$

where $\boldsymbol{\mu}$ represents the mean of the data values, and $\boldsymbol{\sigma}$ represents the standard deviation of the data values. Additionally, the standard deviation, where $\boldsymbol{n}$ represents the number of data points and $\boldsymbol{x_j}$ $(\boldsymbol{j = 1, 2, \dots, n})$ represents the individual data values, can be calculated using the equation (7). By applying $\boldsymbol{z}$-score normalization, the resulting normalized values ($\boldsymbol{x_{norm}}$) will have a mean of 0 and a standard deviation of 1. This process enables more effective comparison and interpretation of the data across different variables.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \mu\right)^2} \quad (7)$$

4.2 Feature selection

4.2.1 Multicollinearity analysis

It is essential to detect and mitigate collinearity among explanatory factors or predictor variables to ensure the precision of predictive outcomes. Collinearity typically arises when two or more predictor variables exhibit a substantial association. In certain instances, collinearity may extend to involve three or more predictor variables, termed multicollinearity. This study employs the VIF method to identify multicollinearity. VIF is defined as the ratio of the variance of an influencing factor ($\boldsymbol{F_i}$) when incorporated into a comprehensive regression model to the variance of the same influencing factor ($\boldsymbol{F_i}$) when included individually in a regression model without that same factor ($\boldsymbol{F_i}$) while considering the presence of other variables [17].

$$VIF(F_i) = \frac{1}{1-R^2_{(X_i|X_{-i})}} \quad (8)$$

The VIF computation for each causal factor is presented in equation (8). Where $\boldsymbol{R^2_{X_i|X_{-i}}}$ denotes the coefficient of determination resulting from regressing the $\boldsymbol{i}$-th variable ($\boldsymbol{X_i}$) against all predictor variables ($\boldsymbol{X_{-i}}$). The computation of $\boldsymbol{R^2_{X_i|X_{-i}}}$ is elucidated in equation (9).

$$R^2_{(X_i|X_{-i})} = 1 - \frac{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (9)$$

Where $\boldsymbol{y_i}$ and $\boldsymbol{\widehat{y_i}}$ signify the observed and predicted values of the target variable for the ith data point, respectively. Additionally, $\boldsymbol{\bar{y}}$ denotes the mean value of the observed target variable. $\boldsymbol{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ represents the total sum of squared errors (TSS), measuring the total variance in the target variable prior to regression. Meanwhile, $\boldsymbol{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$ denotes the residual sum of squared errors (RSS), measuring the residual variability after regression.

A coefficient of determination closing to 0 indicates the absence of collinearity, resulting in a VIF value of 1. Conversely, a coefficient of determination close to 1 suggests the presence of collinearity, leading to higher VIF values. Generally, a VIF exceeding 5 indicates a substantial collinearity among predictor variables, denoting multicollinearity. Hence, in this investigation, any causal factor exhibiting a VIF value surpassing five was eliminated to mitigate the repercussions of multicollinearity.

4.2.2 Information Gain Ratio

The IGR serves as a pivotal metric for evaluating a feature's significance and predictive efficacy within a dataset acquired through the computation of the information gained relative to the entropy of the target variable. Entropy, denoting the degree of randomness inherent in the dataset, spans from 0 to 1, with higher values signifying augmented randomness and concurrently diminished predictive utility [18].

$$Entropy\ (S) = -\sum_{i=1}^{c} P_i \log_2(P_i) \quad (10)$$

Equation (10) initially computes the entropy, indicating data uncertainty, on the training dataset $\boldsymbol{S}$. Here, the variable $\boldsymbol{c}$ denotes the count of potential classes. In predicting FLF vulnerability, the count of potential classes is fixed at two: FLF versus non-FLF. Meanwhile, $\boldsymbol{P_i}$ signifies the proportion of samples affiliated with class i.

$$Entropy\left(S_j\right) = -\sum_{j=1}^{c} P_{ij} \log_2\left(P_{ij}\right) \quad (11)$$

Subsequently, equation (11) facilitates the computation of data entropy for each data

partition ($\boldsymbol{S_1, S_2,\ldots, S_n}$) upon causal factors. Here, the variable $\boldsymbol{F}$ represents the causal factor under scrutiny, with data partitioning predicated on the values of the causal factor $\boldsymbol{F}$. Concurrently, $\boldsymbol{P_{ij}}$ represents the proportion of samples within partition $\boldsymbol{S_j}$ (resultant partition from causal factor $\boldsymbol{F}$) attributed to class i.

$$Gain(F) = Entropy\,(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot Entropy\,(S_j) \quad (12)$$

Upon acquiring the values of Entropy ($\boldsymbol{S}$) and Entropy ($\boldsymbol{S_j}$), the information gained from the causal factor $\boldsymbol{F}$ is derivable. Information gain embodies the reduction in entropy within dataset $\boldsymbol{S}$, consequent to introducing causal factor $\boldsymbol{F}$ or Entropy ($\boldsymbol{S_j}$). Equation (12) encapsulates the computation of information gain, with $\boldsymbol{v}$ denoting the count of partitions produced by causal factor $\boldsymbol{F}$. Meanwhile, $\boldsymbol{|S_j|}$ signifies the count of samples within partitions $\boldsymbol{S_j}$, and $\boldsymbol{|S|}$ denotes the total sample count in dataset $\boldsymbol{S}$.

$$SplitInfo(S,F) = -\sum_{j=1}^{v} \frac{|S_j|}{|S|} \log_2 \left(\frac{|S_j|}{|S|}\right) \quad (13)$$

After this, the $\boldsymbol{SplitInfo}$ calculation measures the potential information extracted from segmenting training data $\boldsymbol{S}$ into $\boldsymbol{v}$ subsets, effectively quantifying how adept causal factors $\boldsymbol{F}$ are at data partitioning. SplitInfo is computed via equation (13).

$$IGR\,(S,F) = \frac{Gain(F)}{SplitInfo(S,F)} \quad (14)$$

Finally, IGR can be calculated using equation (14). IGR itself is the ratio between information $\boldsymbol{Gain}$ and $\boldsymbol{SplitInfo}$. A higher IGR signifies enhanced predictive efficacy, while an IGR of 0 indicates the incapability of causal factor $\boldsymbol{F}$ to discriminate among various data classes, rendering it ineffective for predictive decision-making.

4.2.3 Backward Elimination

BE is a feature selection methodology commonly employed to eliminate features that substantially enhance or decrease classification accuracy [19]. This procedure commences by formulating a regression model encompassing all predictor variables, followed by an assessment of the significance of each predictor through statistical tests, such as the computation of the p-value. The p-value measures the statistical significance of the association between the predictor variable and the target variable, with predictors exhibiting higher p-values deemed least significant and thus removable from the modeling phase.

Determining the deletion criterion hinges upon the value of ρ or retention threshold. A predictor variable is slated for deletion if its p-values surpass the retention threshold. The rationale behind eliminating predictors failing to meet this threshold is potentially enhancing classification accuracy in subsequent model iterations.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (15)$$

To calculate the p-value ($\boldsymbol{z}$), equation (15) can be utilized, where $\boldsymbol{n}$ represents the sample size, and $\boldsymbol{\hat{p}}$ and $\boldsymbol{p_0}$ denote the sample proportion and hypothesized proportion, respectively. While a retention threshold of 0.05 is conventionally employed to assess the statistical significance of predictor variables, in this study, a more lenient retention threshold of 0.2 is adopted to enable a more permissive approach to feature selection.

4.2.4 Random Forest

The assessment of feature importance generated by the RF model constitutes a widely utilized metric for feature prioritization across diverse domains [20]. The RF model undergoes training utilizing the designated training dataset within the feature selection procedure. Subsequently, the Mean Decrease Impurity (MDI) of the causal factor $\boldsymbol{F}$ is computed utilizing equation (16). Specifically, this feature importance measure relies on impurity, quantified through the Gini impurity as in equation (17).

$$MDI\,(F) = \sum(Imp_{Parent} - Imp_{child}) \quad (16)$$

$$GiniImpurity = 1 - \sum_{i=1}^{c} {p_i}^2 \quad (17)$$

Here, the variable $\boldsymbol{c}$ denotes the count of target classes, denoting two classes for FLF and Non-FLF. Meanwhile, $\boldsymbol{p_i}$ represents the proportion of samples allocated to class-i. $\boldsymbol{MDI}$ ($\boldsymbol{F}$) serves as a metric delineating the significance of factors contributing to $\boldsymbol{F}$. The term $\boldsymbol{Imp_{Parent}}$ denotes the Gini impurity value before the tree is split, whereas $\boldsymbol{Imp_{child}}$ signifies the average gini impurity post-split across all child nodes.

4.3 Machine learning models

4.3.1. Gaussian Naive Bayes

The NB technique is a statistical methodology designed to determine the prior probability of an event based on the proportion observed in a specified output class [17]. This approach is based on Bayes' Theorem, as represented by equation (18).

Where $\boldsymbol{P(Y|X)}$ denotes the posterior distribution of the target variable $\boldsymbol{Y}$ given the predictor variable $\boldsymbol{X(X_1, X_2, \dots, X_n)}$, and $\boldsymbol{P(X|Y)}$ represents the likelihood of the predictor variable $\boldsymbol{X(X_1, X_2, \dots, X_n)}$, given a specific value of variable $\boldsymbol{Y}$. The research employs the Gaussian Naive Bayes (GNB) techniques, assuming that the predictor variables follow a gaussian distribution [17]. The gaussian distribution assumes that each feature in the data has an independent influence in predicting the target variable. Consequently, the likelihood probability $\boldsymbol{P(X_i|y_i)}$ is estimated by considering the mean $(\boldsymbol{\mu_i})$ and standard deviation $(\boldsymbol{\sigma_i})$ of each predictor variable conditioned on its respective class. The formulation of Gaussian NB is presented in equation (19). Through this equation, GNB effectively captures the distributional characteristics of the data, enabling probabilistic predictions predicated on the assumption of a gaussian distribution.

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \quad (18)$$

$$P(X_i|y_i) = \left(\frac{1}{\sqrt{2\pi\sigma_i}}\right) \cdot \exp\left(-\frac{(X_i - \mu_i)^2}{2\sigma_i^2}\right) \quad (19)$$

4.3.2. Support Vector Machine

The SVM is a statistical machine learning algorithm designed to determine the optimal hyperplane for data classification. SVM works to separate distinct classes within the target variable while maximizing classification margins [21]. Its framework relies on two fundamental principles: employing kernel functions and conceiving an optimal classification hyperplane. The search for the optimal hyperplane entails solving a linearly constrained quadratic programming problem, formulated as equation (20).

$$\begin{cases} min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right) \\ y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, 2, \dots, n \end{cases} \quad (20)$$

Here, $\boldsymbol{w}$ and $\boldsymbol{b}$ denote weight vectors determining the hyperplane orientation and bias, respectively, while $\boldsymbol{\xi_i}$ serves as a positive slack variable for data points. The regularization parameter $\boldsymbol{C}$ balances between training and generalization errors [22]. A smaller $\boldsymbol{C}$ widens the margin, potentially leading to more misclassifications, whereas a larger $\boldsymbol{C}$ narrows the margin, risking overfitting.

The Lagrange method is employed to solve equation (20) [23], yielding the decision function:

$$y = sign(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b) \quad (21)$$

Here, $\boldsymbol{y}$ denotes the predicted class label, $\boldsymbol{y_i}$ represents the class label of the $\boldsymbol{i}$-th support vector, $\boldsymbol{\alpha_i}$ signifies the weight or coefficient assigned to each support vector, and $\boldsymbol{K(x_i, x)}$ represents a kernel function.

$$K(x_i, x) = exp\left(-\frac{||x_i - x||^2}{2\sigma^2}\right) \quad (22)$$

In SVM, the Radial Basis Function (RBF) kernel is often employed to address nonlinear classification challenges. The RBF kernel, utilized within SVM, calculates the distance between support vector $\boldsymbol{x_i}$ and input $\boldsymbol{x}$ through the following equation (22) [24]. Where $\boldsymbol{\sigma}$ denotes the variance parameter controlling the width of the Gaussian distribution. The choice of $\boldsymbol{\sigma}$ significantly influences the smoothness of the decision boundary and the model's flexibility [1].

4.3.3. Logistic Regression

LR is a statistical method for addressing binary classification tasks. The LR model generates a likelihood value as its output, indicating the probability of a specific class occurrence based on the observed values of predictor variables [5]. This model effectively captures the probabilistic relationship between the target variables and predictors.

$$P(y|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \quad (23)$$

The logistic function, also known as the sigmoid function, is utilized within the LR models to transform a linear combination of predictors into values ranging between 0 and 1. The LR function, as depicted in equation (23), mathematically expresses this transformation. In this equation, $\boldsymbol{P(y|X)}$ represents the probability of the target variable $\boldsymbol{Y}$ given the predictors $\boldsymbol{X(X_1, X_2, \dots, X_n)}$. $\boldsymbol{\beta_0}$ denotes the intercept constant, while $\boldsymbol{\beta_i}$ represents the regression coefficient associated with each predictor $(\boldsymbol{i = 1, 2, \dots, n})$.

4.3.4. Artificial Neural Network

ANN is a computational model inspired by the structure and function of neural networks in the human brain. This model aims to process and analyze complex information using interconnected neurons. The training process in ANN aims to find the best weight for each neuron unit [23]. ANN consists of three main layers, namely the input layer, hidden layer, and output layer. The input layer is an input layer whose neurons are adjusted according to the number of predictor variables. The hidden layer functions to process data received from the input layer. The output layer produces output from the ANN, so the output neurons in the output layer are used as indicators of FLF vulnerability [2].

ANN uses a backpropagation algorithm, which consists of several important stages. First, the forward pass stage is carried out to make predictions based on the existing or initial weight. The second stage of error measurement is carried out to calculate the error from the forward pass stage. Third, the reverse pass stage is carried out to calculate the error contribution for each neuron connection. Finally, the gradient descent stage is used to update the ANN parameters [25].

4.3.5. Random Forest

The RF algorithm, an ensemble learning technique, is frequently applied in classification and regression tasks. RF comprises numerous decision trees whose collective predictions enhance the accuracy of prediction. By aggregating the outputs of individual trees, RF yields a final prediction [2]. The convergence of the generalization error in RF is contingent upon augmenting the number of trees. Thus, determining the optimal number of trees becomes imperative for achieving convergence [26].

The learning process of RF entails iterative procedures involving data resampling with replacement and random alterations in predictor sets. Leveraging this stochasticity ensures that each tree within the ensemble concentrates on distinct facets of the dataset. Employing replacement during sampling introduces variability among decision trees, mitigating overfitting issues and enhancing the algorithm's generalization capability [17].

4.3.6. Gradient Boosting Machine

The GBM represents an ensemble learning technique rooted in tree-based algorithms aimed at enhancing the predictive capabilities of a singular model by amalgamating multiple models. In contrast to the RF algorithm, GBM iteratively constructs trees, with each subsequent tree endeavoring to rectify the errors stemming from its predecessor. Notably, the subsequent tree is derived by minimizing residual errors resulting from the preceding tree's predictions. This iterative process persists until either the predictive outcomes stabilize, or the predefined maximum threshold of trees is attained [2].

4.3.7. Extreme Gradient Boosting

XGB represents a scalable and efficient tree-based ensemble learning algorithm renowned for its efficacy in classification and regression tasks. Unlike GBM, XGB differentiates itself through its adeptness at mitigating overfitting during training. This is achieved by implementing two supplementary methodologies: shrinkage and column subsampling. The incorporation of these techniques not only enhances model robustness but also elevates overall predictive accuracy [2].

4.4 Evaluation metrics

4.4.1 Statistical methods

Statistical indices are increasingly prevalent in evaluating the efficacy of individual ML models. Various metrics, such as accuracy, precision, recall, and F1-score, are frequently employed. The mathematical formulations are delineated in Table 3 [19], where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative predictions, respectively.

Table 3. Statistical evaluation metric [19].

Evaluation Metrics	Equation
Accuracy (Acc)	$\boldsymbol{Acc = \frac{TP + TN}{TP + TN + FP + FN}}$
Precision (Prec)	$\boldsymbol{Precision = \frac{TP}{TP + FP}}$
Recall (Rec)	$\boldsymbol{Recall = \frac{TP}{TP + FN}}$
F1-Score (F1)	$\boldsymbol{F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}}$

4.4.2 Receiver operating characteristic curve

In assessing the performance of predictive models concerning FLF vulnerability, the ROC curve—a commonly employed evaluation technique within relevant academic investigations—is employed in this research. This graphical representation illustrates various pairs of statistical measures, such as True Positive Rate (TPR) and False Positive Rate (FPR), across different thresholds [19]. A higher AUC value signifies enhanced predictive capacity within the model. Thus, an increased AUC value indicates the model's effectiveness and quality in predicting susceptibility to FLF.

$$TPR = \frac{TP}{TP+FN} \quad (24)$$

$$FPR = \frac{FP}{FP+FN} \quad (25)$$

Within the ROC curve, adjusting the classification threshold downwards results in the classification of more data as positive, thereby refining the predictions of FP and TP. The computations for TPR and FPR for each of these assessment metrics are presented in equation (24) and equation (25) [27].

5. Result and Analysis

5.1 Multicollinearity analysis

The multicollinearity among predictor variables can affect the accuracy of predictive was employed to assess multicollinearity across models for FLF vulnerability. The VIF technique the fourteen causal factors to address this issue. A VIF value exceeding 5 denotes significant multicollinearity among predictor variables.

The outcomes of the multicollinearity examination are presented in Table 4. According to the findings, no evidence of multicollinearity was detected among the predictor variables. This is shown by the high VIF value, which stands at 2.916 for the "Average Relative Humidity" factor below the threshold of 5. Consequently, no elimination of predictor variables was done at this phase.

Table 4. Multicollinearity analysis.

Causal Factors	VIF
Elevation	1.193
Slope Angle	1.376
Slope Aspect	1.008
Plan Curvature	1.292
TWI	1.440
Land Cover	1.183
NDVI	1.204
Distance To Roads	1.392
Distance To Rivers	1.118
Distance To Buildings	1.357
Annual Rainfall	2.078
Average Temperature	2.028
Average Wind Speed	1.365
Average Relative Humidity	2.916

5.2 Elimination of the less important causal factors

Following the multicollinearity analysis, the fourteen existent causal factors underwent further selection through three different selection methods: IGR, BE, and RF. The IGR method assesses the predictive effectiveness of each causal factor by measuring the informational gain it confers relative to the entropy reduction of the target variable, where higher IGR values denote better predictive ability.

As shown in Figure 4, the IGR computations delineate the performance of each predictor variable. Upon feature selection analysis, variables such as Distance to Rivers and TWI exhibited an IGR value of 0. Consequently, these two factors were eliminated, leaving twelve predictor variables as the initial feature group for subsequent modeling stages.

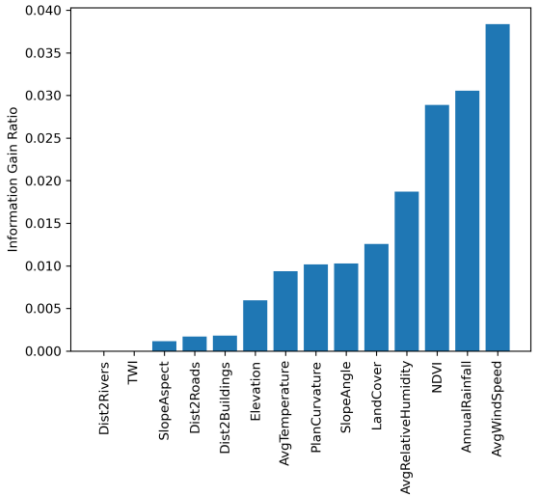

Figure 4. IGR of Predictor Variables. Higher IGR values indicate increased predictive efficacy. Distance to Rivers and TWI display an IGR of 0, indicating ineffectiveness for predictive decision-making.

Subsequently, the BE method identifies causal factors that either markedly reduce or minimally increase classification accuracy, emphasizing factors bearing high p-values. The removal criterion was established within this selection methodology at a p-value threshold of 0.2.

Table 5 shows the results of p-value calculations for each predictor variable using the BE method. As explained in the previous paragraph, a higher p-value (exceeding the threshold) indicates the inability of this factor to increase classification accuracy. Therefore, the Slope Aspect, Plan Curvature, Distance to Roads, and Distance to Rivers exceeded the retention threshold, thereby causing these features to be removed from the secondary features group.

Table 5. p-value computations for each predictor variable.

Causal factors	P-value
Elevation	0.001
Slope Angle	0.004
Slope Aspect	0.672
Plan Curvature	0.572
TWI	0.111
Land Cover	0.001
NDVI	0.0
Distance To Roads	0.503
Distance To Rivers	0.511
Distance To Buildings	0.0
Annual Rainfall	0.0
Average Temperature	0.161
Average Wind Speed	0.008
Average Relative Humidity	0.01

Lastly, the RF method in Figure 5 quantifies feature importance through gini impurity assessment, facilitating the elimination of causal factors or predictor variables registering importance values below the mean value.

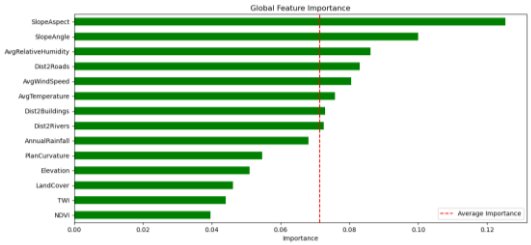

Figure 5. Feature importance for each predictor variable. The red line represents the average feature importance across all predictors. Plan Curvature, TWI, NDVI, Land Cover, and Annual Rainfall exhibit importance values below the average.

As illustrated in Figure 5, the feature importance analysis via the RF method outlines variables such as Elevation, Plan Curvature, TWI, NDVI, Land Cover, and Annual Rainfall, each exhibiting importance values below the average. Consequently, these six factors were excluded from the third feature group. Employing three different feature selection methodologies yields three groups of features selected for utilization in the subsequent ML modeling phase.

5.3 Model evaluation

The modeling procedure involves the implementation of seven distinct ML algorithms. Each ML model undergoes training in 10 iterations, utilizing three groups of features derived from feature selection processes. Subsequently, these models are compared based on their average accuracy and best F1-score to identify the optimal feature selection method. Hyperparameter optimization for each ML model is conducted through the "GridSearchCV" approach to find hyperparameter. in Table 6, configurations that yield optimal outcomes. The bolded values represent the specific hyperparameters selected for this study.

The best model obtained by the best feature selection method will then be used for the testing stage to be further identified using statistical methods and the ROC curve. This evaluation aims to assess each model's relative performance and predictive ability.

Based on the findings presented in Table 7, it is evident that the choice of feature selection method significantly impacts the performance of different machine learning models. For GNB modeling, both the IGR and BE methods yield the same highest F1-score and average accuracy of 0.682 and 0.643, respectively, indicating that either method can be effectively used. In SVM modeling, the IGR method demonstrates superior

Table 6. hyperparameter pool for each method.

ML model	Hyperparameter	Value
GNB	var_smoothing	**1e-9**, 1e-8, 1e-7, 1e-6, 1e-5
SVM	Regularization	0.1, **1**, 10
	Kernel type	linear, **rbf**
	Kernel coefficient	**0.1**, 1
LR	Penalty type	l1, **l2**
	Inverse of regularization strength	**0.001**, 0.01, 0.1, 1, 10, 100
	Solver	**liblinear**
	Maximum iteration	**100**, 200, 300, 400, 500
ANN	Hidden layer	**2**
	Neuron for each hidden layer	**16**
	Hidden layer activation function	**ReLU**
	Output layer activation function	**Sigmoid**
	Loss function	**Binary cross entropy**
	Optimizer	**Adam optimizer**
	Learning rate	**0.001**
RF	Maximum tree depth	**10**, 20, 50, 100
	Minimum sample leaf	1, **2**, 4
	Minimum sample split	2, **5**, 10
	Number of estimators	50, **100**, 200
GBM	Maximum tree depth	**10**, 20, 50, 100
	Minimum sample leaf	**1**, 2, 4
	Minimum sample split	2, **5**, **10**
	Number of estimators	50, **100**, 200
	Learning rate	**0.01**, 0.1, 0.2
XGB	Maximum tree depth	10, 20, **50**, 100
	Minimum child weight	**1**, 5, 10
	Subsample ratio	**0.6**, 0.8, 1.0
	Number of estimators	50, 100, **200**
	Learning rate	**0.01**, 0.1, 0.2
	Gamma	**0**, 0.1, 0.2
	Scale pos weight	1, **2**, 5

Table 7. feature group comparison for each ML modeling.

ML model	Feature group (FS method)	Best F1-score	Average Accuracy
GNB	**IGR**	**0.682**	**0.643**
	BE	**0.682**	**0.643**
	RF	0.655	0.607
SVM	**IGR**	**0.696**	**0.667**
	BE	0.687	0.657
	RF	0.668	0.642
LR	IGR	0.666	0.630
	BE	**0.669**	**0.632**
	RF	0.606	0.606
ANN	IGR	0.670	0.647
	BE	**0.691**	**0.667**
	RF	0.633	0.621
RF	**IGR**	**0.713**	**0.681**
	BE	0.711	0.678
	RF	0.690	0.655
GBM	IGR	0.696	**0.665**
	BE	**0.701**	**0.665**
	RF	0.688	0.660
XGB	**IGR**	**0.733**	0.663
	BE	0.726	**0.667**
	RF	0.717	0.63

performance, achieving an F1-score of 0.696 and an average accuracy of 0.667, making it the preferable choice for this model. Conversely, LR modeling shows that the BE method produces slightly better results, with an F1-score of 0.669 and an average accuracy of 0.632. Similarly, ANN modeling benefits more from the BE method, which results in an F1-score of 0.691 and an average accuracy of 0.667. In RF modeling, the IGR method achieves the highest F1-score and average accuracy of 0.713 and 0.681, respectively, highlighting its effectiveness. GBM modeling reveals that the BE method provides a marginally better F1-score of 0.701, while both BE and IGR methods yield the same average accuracy of 0.665. Lastly, in XGB modeling, the IGR method stands out with the highest F1-score of 0.733, though it has a lower average accuracy of 0.663 compared to the BE method, which shows an average accuracy of 0.667.

Overall, the IGR method generally enhances model accuracy across various modeling techniques. This occurs because the IGR method effectively identifies and selects the most relevant features by measuring the information gain ratio, thus improving model performance by reducing noise and focusing on significant attributes. However, in models like XGB, the higher F1-score with the IGR method, despite a lower average accuracy, suggests that the IGR method is better at identifying relevant features that improve the balance between precision and recall.

Table 8. statistical evaluation of ML models in percentage (%).

Dataset	Model	Accuracy	Precision	Recall	F1
Training	GNB	0.626	0.605	0.730	0.662
	SVM	0.702	0.667	0.810	0.732
	LR	0.626	0.610	0.705	0.654
	ANN	0.708	0.682	0.779	0.727
	RF	0.775	0.729	0.876	0.796
	GBM	**0.810**	**0.760**	0.908	**0.827**
	XGB	0.788	0.705	**0.989**	0.823
Testing	GNB	0.643	0.613	0.770	0.682
	SVM	0.661	0.627	0.792	0.700
	LR	0.630	0.606	0.741	0.666
	ANN	0.647	0.627	0.719	0.670
	RF	**0.681**	**0.645**	0.798	0.713
	GBM	0.664	0.636	0.762	0.693
	XGB	0.663	0.605	**0.930**	**0.733**

An examination of ML models was carried out using a dataset selected through the IGR method, as shown in Table 8. During the training phase, the GBM model showed the highest accuracy, precision, and F1-score, with values of 0.810, 0.760, and 0.827, respectively, surpassing other models. However, the XGB model excelled in the recall, achieving an impressive value of 0.989, highlighting its ability to capture forest fires during training accurately.

Moving to the testing phase, the RF model had the highest accuracy at 0.681 and a precision of 0.645, indicating its effectiveness in predicting forest fire occurrences. Nonetheless, the XGB model once again led in recall with a value of 0.930 and an F1-score of 0.733. Additionally, in Figure 6, XGB demonstrated the highest AUC value of 0.959 and 0.743, during training and testing, confirming its ability to identify FLF vulnerability within the NTT Province.

Upon thorough evaluation, the superior performance displayed by the XGB model in both

training and testing phases makes it the best choice for forest fire risk analysis. Its exceptional recall rates demonstrate its ability to accurately identify forest fire occurrences, which is essential for reducing risks. Moreover, its consistently high AUC values underscore its reliability in identifying FLF vulnerability, enhancing its usefulness as a tool for forest fire management strategies. Consequently, the XGB model represents the most effective approach for predicting FLF vulnerability within the NTT Province.

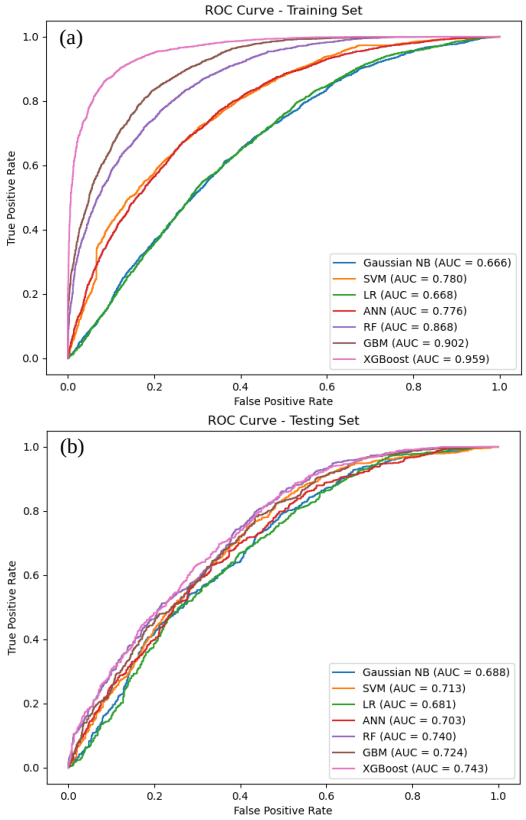

Figure 6. The ROC curve of GNB, SVM, LR, ANN, RF, GM, XGB, and ANN models in FLF Vulnerability assessment: (a) training and (b) testing. XGB demonstrated the highest AUC value during training and testing.

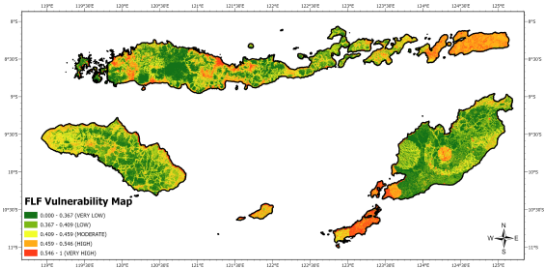

Figure 7. FLF Vulnerability Map. Prediction Probabilities Classified into Five Risk Classes: Very Low (0%-30%), Low (30%-50%), Moderate (50%-70%), High (70%-90%), and Very High (90%-100%).

5.4 FLF vulnerability mapping

The trained XGB model was subsequently applied to generate predictions regarding the vulnerability of FLF across all pixels within the raster dataset of NTT Province. Using the Python library, Rasterio facilitated the extraction of values from individual pixels, enabling the deployment of ML models designed for predictive analysis on raster datasets. The resulting predictions were then exported as TIFF files for further classification utilizing ArcGIS Pro.

Previously, the model was initially trained for binary classification, utilizing a binary system (0-1) based on the training dataset. However, in line with common practice [2], [5], [8], [10], models are now required to produce vulnerability maps with five different levels or classes during their application phase. This modification process involves using the trained initial model to predict each pixel's probability. Subsequently, through post-processing techniques, we manually categorize these probabilities into five distinct classes: very low, low, moderate, high, and very high, thus enabling the model to produce vulnerability assessments across multiple classes despite its binary classification training. The FLF vulnerability map within NTT Province is shown in Figure 7.

5.5 Causal analysis

The FLF vulnerability map generated in this study was subsequently analyzed to find the underlying factors contributing to FLF in the NTT Province. Comparative analysis was conducted between the FLF raster map (Figure 7) and various raster representing pertinent causal factors (Figure 3). The analysis aimed to identify the raster classes associated with the highest incidence of FLF, particularly those falling within the high and very high vulnerability categories.

The findings reveal that areas with elevations between 0-400 meters (Class 2) have the highest FLF risk, with a 67.77% class distribution. Slopes with an angle of less than 5 degrees (Class 1) also show a high FLF risk, comprising 30.8% of the distribution. Northern-facing slopes (Class 1) have a higher, but less significant, FLF risk at 13.32%. Plan curvature in concave and convex shapes (Classes 1 and 3) is associated with higher FLF risks (40.45% and 42.38%, respectively) than in flat areas. Forest land cover (Class 8) is particularly vulnerable, with a risk distribution of 75.03%. Areas with unhealthy or less fertile vegetation, indicated by an NDVI range of 0 to 0.33 (Class 2), show a substantial risk of 69.96%.

Moreover, distance to roads within 200 meters (Class 1) is a significant risk factor at 32.33%. Buildings located more than 2000 meters from forested areas (Class 5) have a risk distribution of 30.78%. Areas with annual rainfall between 51-99 mm (Class 2) have a risk distribution of 29.75%. The highest FLF risk, 31.5%, is found in areas with an average temperature of 27.672-27.923 degrees Celsius (Class 10). Similarly, average wind speeds of 2.859 to 3.079 m/s (Class 6) contribute 16.79% to the risk distribution. Finally, regions with an average relative humidity of 72.695 to 73.949 mm/day (Class 1) show a

significant risk, representing 22.25% of the class distribution.

Based on the analysis of findings, it was determined that various factors contribute to FLF in NTT Province. Therefore, several strategies are proposed to mitigate these fires. Initially, land management practices should target high-risk areas by introducing fire-resistant vegetation species, such as Laban, Dadap Duri, Mediterranean cypress, Agarwood, banana, areca palm, and pawpaw trees, known for their high moisture content and resilience to fires [28]. Secondly, sustainable forest management policies should be enforced, with strict monitoring of fire-prone activities [29]. Satellite surveillance can identify areas with low NDVI values, prompting proactive measures to enhance plant health [30]. Public outreach and education campaigns are crucial for raising awareness about fire prevention and environmental conservation [31]. Thirdly, infrastructural improvements are necessary to combat FLF effectively, including enhancing firefighting facilities and road accessibility to vulnerable areas. Moreover, monitoring water sources and ensuring adequate water availability are vital measures.

Addressing FLF in NTT Province demands collaborative efforts across sectors. Implementation of these solutions requires cooperation among government agencies, communities, and relevant stakeholders to achieve optimal response and prevention outcomes.

6. Conclusion

The analysis identifies the best combination for predicting FLF vulnerability in NTT Province as the IGR method with the XGB model. The AUC scores show 0.959 for training and 0.743 for testing. Factors causing FLF in NTT Province include Elevation, Slope Angle, Slope Aspect, Land Cover, NDVI, Distance to Roads and Buildings, Annual Rainfall, Average Temperature, Average Wind Speed, and Average Relative Humidity. The FLF vulnerability map generated in this study was subsequently analyzed to find the underlying factors contributing to FLF in the NTT Province. Based on these findings, some approaches were proposed to reduce FLF risk in NTT Province, including land management strategies, planting fire-resistant vegetation, enforcing sustainable forest management policies, monitoring fire-prone activities, running public awareness campaigns, and improving infrastructure.

References

[1] Y. Shao, Z. Feng, L. Sun, X.Yang, Y. Li, B. Xu, and Y. Chen, “Mapping China’s Forest Fire Risks with Machine Learning,” *Forests*, vol. 13, p. 856, 2022, doi:10.3390/f13060856.

[2] H. A. Akinci and H. Akinci, “Machine learning based forest fire susceptibility assessment of Manavgat district (Antalya), Turkey,” *Earth Science Informatics*, vol. 16, pp. 397-414, 2023, doi:10.1007/s12145-023-00953-5.

[3] M. A. Enoh, U. C. Okeke, and N. Y. Narinua, “Identification and modelling of forest fire severity and risk zones in the Cross – Niger transition forest with remotely sensed satellite data,” *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 24, pp. 879-887, 2021, doi:10.1016/j.ejrs.2021.09.002.

[4] Kementerian Lingkungan Hidup dan Kehutanan, *Laporan Kinerja KLHK 2022*, Kementerian Lingkungan Hidup dan Kehutanan, 2022.

[5] M. Mohajane, R. Costache, F. Karimi, Q. B. Pham, Essahlaoui, H. Nguyen, G. Laneve, and F Oudija, “Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area,” *Ecological Indicators*, vol. 129, p. 107869, 2021, doi:10.1016/j.ecolind.2021.107869.

[6] P. Rakshit, S. Sarkar, S. Khan, P. Saha, S. Bhattacharyya, N. Dey, S. M. N. Islam, and S. Pal, “Prediction of Forest Fire Using Machine Learning Algorithms: The Search for the Better Algorithm,” *2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA)*, pp. 1-6, 2021, doi: 10.1109/CITISIA53721.2021.9719887.

[7] M. Seddouki, M. Benayad, Z. Aamir, M. Tahiri, M. Maanan, and H. Rhinane, “Using Machine Learning Coupled with Remote Sensing for Forest Fire Susceptibility Mapping. Case Study Tetouan Province, Northern Morocco,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)*, vol. 47, pp. 333-342, 2023, doi:10.5194/isprs-archives-XLVII4-W6-2022-333-2023.

[8] N. Hidayanto, A. H. Saputro, and D. E. Nuryanto. “Peatland Data Fusion for Forest Fire Susceptibility Prediction Using Machine Learning” *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, vol. 544. 2021. doi:10.1109/ISRITI54043.2021.9702762.

[9] L. Qiu, J. Chen, J., L. Fan, L. Sun, and C. Zheng, “High-resolution Mapping of Wildfire Drivers in California Based on Machine Learning,” *Science of the Total Environment*, vol. 833, p. 155155, 2022. doi:10.1016/j.scitotenv.2022.155155.

[10] H. Moayedi, and M. A. S. A. Khasmakhi, “Wildfire Susceptibility Mapping using Two Empowered Machine Learning Algorithms,” *Stochastic Environmental Research and Risk Assessment*, vol. 37, pp. 49-72. 2022. doi:10.1007/s00477022-02273-4.

[11] C. Tan and Z. Feng, “Mapping Forest Fire Risk Zones Using Machine Learning Algorithms in Hunan Province, China,,” *Sustainability*, vol. 15, p. 6292, 2023. doi:10.3390/su15076292.

[12] Badan Pusat Statistik Provinsi Nusa Tenggara Timur, *Provinsi Nusa Tenggara Timur dalam Angka*, Badan Pusat Statistik Provinsi Nusa Tenggara Timur, 2023.

[13] D. Odonohue, “*Topographic Wetness Index in ArcGIS Pro*,” Mapscaping, https://mapscaping.com/topographic-wetness-index-in-arcgis-pro/, (accessed Feb. 21, 2024).

[14] ArcMap “*Con*,” ArcGIS Desktop, https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/con-.htm, (accessed Feb. 21, 2024).

[15] A. K. Taloor, Drinder Singh Manhas, and G. Chandra Kothyari, "Retrieval of land surface temperature, normalized difference moisture index, normalized difference water index of the Ravi basin using Landsat data," Appl. Comput. Geosci., vol. 9, no. December 2020, p. 100051, 2021, doi: 10.1016/j.acags.2020.100051.

[16] E. Kreyszig, *Advanced Engineering Mathematics*, 10th ed. Wiley Global Education, 2010.

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed. in Springer Texts in Statistics. New York, NY: Springer US, 2021. doi: 10.1007/978-1-0716-1418-1.

[18] Kurniabudi, A. Harris, and A. E. Mintaria, "Komparasi Information Gain, Gain Ratio, CFs-Bestfirst dan CFs-PSO Search Terhadap Performa Deteksi Anomali," *Jurnal Media Informatika Budidarma*, vol. 5, pp. 332-343, 2021, doi:10.30865/mib.v5i1.2258.

[19] B. T. Pham, B. Pradhan, D. Tien Bui, I. Prakash, and M. B. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environmental Modelling & Software*, vol. 84, pp. 240–250, 2016, doi: 10.1016/j.envsoft.2016.07.005.

[20] A. Hjerpe, "Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data," *Master theses at CSC*. KTH Computer Science and Communication, 2016.

[21] Y. Huang and L. Zhao, "Review on landslide susceptibility mapping using support vector machine," *Catena*, vol. 165, pp. 520–529, 2018, doi:10.1016/j.catena.2018.03.003.

[22] D. T. Bui, P. Tsangaratos, V. T. Nguyen, N. Van Liem, and P. T. Trinh, "Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment," *Catena*, vol. 188, p. 104426, 2020, doi: 10.1016/j.catena.2019.104426.

[23] C. Zhou et al., "Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China," *Computers and Geosciences*, vol. 112, no. September 2017, pp. 23–37, 10.1016/j.cageo.2017.11.019.

[24] B. Altınel, M. C. Ganiz, and B. Diri, "A corpus-based semantic kernel for text classification by using meaning values of terms," *Engineering Applications of Artificial Intelligence*, vol. 43, pp. 54–66, 2015, doi:10.1016/j.engappai.2015.03.015.

[25] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Inc, 2019.

[26] A. M. Youssef, H. R. Pourghasemi, Z. S. Pourtaghi, and M. M. Al-Katheeri, "Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia," *Landslides*, vol. 13, no. 5, pp. 839–856, 2016, doi:10.1007/S10346-015-0614-1/TABLES/4.

[27] Google for Developers, "*Classification: Roc curve and AUC*", Google, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, (accessed Jan. 5, 2024).

[28] Food and Agriculture Organization of the United Nations, *Guidelines on sustainable forest management in drylands of sub-Saharan Africa*, Food and Agriculture Organization of the United Nations, 2010.

[29] UN-REDD Programme, *National Forest Monitoring Systems: Monitoring and Measurement, Reporting and Verification (M & MRV) in the Context of REDD+ Activities*, UN-REDD Programme, 2013.

[30] N. Pettorelli, W. F. Laurance, T. G. O'Brien, M. Wegmann, H. Nagendra, and W. Turner, "Satellite remote sensing for applied ecologists: opportunities and challenges," *Journal of Applied Ecology*, vol. 51, no 4., pp. 839-848, 2014, doi: http://www.jstor.org/stable/24032484.

[31] L. Nurhidayah, R. Astuti, H. Hidayat, and R. Siburian, "Community-Based Fire Management and Peatland Restoration in Indonesia*," Environmental Governance in Indonesia*, vol. 61, pp. 135-150, 2023. doi:10.1007/978-3-031-15904-6_8.