MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF
SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

HOANG ANH DUC

DATA CLASSIFICATION TECHNIQUES AND THEIR APPLICATION IN DEVELOPING FOREST FIRE AND LANDSLIDE RISK MAPPING

SUMMARY OF DISSERTATION ON INFORMATION SYSTEM
Code: 9 48 01 04

Ha noi - 2025

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy of Science and Technology

Supervisors:

Supervisor 1: Prof. Dr. Dang Van Duc, Institute of Information Technology

Supervisor 2: Prof. Dr. Le Van Hung, Thuy Loi University

Referee 1: …

Referee 2: …

Referee 3: ….

The dissertation will be examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at…….. (time, date, year…)

This dissertation can be found at:

1) Graduate University of Science and Technology Library
2) National Library of Vietnam

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. H. V. Le, N. T. Thanh, D. H. Nghị, and **A. D. Hoang**, " DEVELOPING A DEEP NEURAL NETWORK MODEL FOR PREDICTING FOREST FIRE RISK OF LAM DONG PROVINCE" in Basic Research in Earth Science and Environment, November 2019, doi: 10.15625/vap.2019.000162.

2. Hung Van Le, **Duc Anh Hoang**, Chuyen Trung Tran, Phi Quoc Nguyen, Van Hai Thi Tran, Nhat Duc Hoang, Mahdis Amiri, Thao Phuong Thi Ngo, Ha Viet Nhu, Thong Van Hoang, Dieu Tien Bui, "A new approach of deep neural computing for spatial prediction of wildfire danger at tropical climate areas," Ecological Informatics, vol. 63, pp. 101300, 2021. DOI: 10.1016/j.ecoinf.2021.101300.

3. Duc Anh Hoang, Hung Van Le, Dong Van Pham, Hoa Viet Pham, and Dieu Tien Bui, "Hybrid BBO-DE Optimized SPAARCTree Ensemble for Landslide Susceptibility Mapping," Remote Sensing, vol. 15, no. 8, p. 2187, Apr. 2023, doi: 10.3390/rs15082187.

INTRODUCTION

1. The urgency of the topic

Natural disasters pose a significant global threat, resulting in substantial economic, social, and environmental losses. Climate change exacerbates the frequency and intensity of these phenomena, presenting considerable challenges for management and prevention efforts. In Vietnam, forest fires and landslides are among the most prevalent and hazardous natural disasters. Forest fires, despite their important ecological role, can have severe consequences when occurring on a large scale, particularly in tropical regions and near populated areas. Landslides, a major geological hazard, affect millions of people worldwide, causing annual loss of life and economic damage. The increasing trend of heavy rainfall and storms, especially in mountainous regions of developing countries, is expected to escalate the incidence of landslides in the future.

Vietnam, situated in Southeast Asia with its diverse topography, is recognized as one of the world's most vulnerable countries to natural disasters. Its mountainous terrain and tropical monsoon climate create favorable conditions for both forest fires and landslides. This vulnerability is further exacerbated by global climate change, as extreme weather events become increasingly prevalent and unpredictable. These factors pose significant challenges for land-use planning, population distribution, and agricultural practices.

In this context, the prediction and mapping of natural disaster risk zones play a crucial role in risk management and mitigation strategies. Machine learning methods integrated with Geographic Information Systems (GIS) are opening new avenues of approach, offering substantial advantages over traditional methodologies. Previous studies in Vietnam on forest fire and landslide prediction have predominantly employed conventional statistical methods, not fully exploiting the potential of advanced machine learning techniques. The application of machine learning and GIS in natural disaster

forecasting enables more precise identification of high-risk areas, thereby supporting effective prevention planning, forest management, and safe land-use planning.

For these reasons, this study aims to solve a real-world problem and has both scientific and practical value. The main goal is to show how effective it is to use satellite data and mapping systems together with modern computer learning methods to create models that predict the risk of natural disasters and make risk zone maps. The results of this study will greatly help improve how Vietnam manages and prevents natural disasters. It will also provide a new method that can be used in other areas with similar conditions.

2. Research objectives

The main objective of this study is to build effective machine learning models to predict and create maps of natural disaster risk zones.

The specific objectives are: to gather and process data, to develop and train machine learning models, to assess the performance of these models, and to produce maps showing areas at risk of natural disasters.

3. Research methodology

To achieve the objectives of this dissertation, the following tasks will be undertaken:

- Examine fundamental theories of machine learning, data classification methodologies, and issues related to the construction of natural disaster risk zonation maps, as well as review the current state of research in this field.
- Develop a protocol for data collection and preprocessing for the study areas.
- Collect, process, and preprocess data for the research areas, including forest fire data from Gia Lai Province, Vietnam, and landslide data from Than Uyen District, Lai Chau Province, Vietnam.
- Generate and train models applicable to both datasets.
- Evaluate criteria, test, and assess the proposed models.

- Visualize the results through the creation of risk zonation maps for forest fires and landslides.

4. Dissertation contributions

1) Propose several techniques for synthesizing, processing, and classifying data to facilitate the zonation of forest fire and landslide risk areas.

2) Utilize the aforementioned classification results to construct risk zonation maps for forest fires in Gia Lai Province and landslides in Than Uyen District, Lai Chau Province..

5. Dissertation structure

The dissertation consists of an introduction, three chapters, and a conclusion.

Chapter 1: Overview of the research field

Chapter 2: A model for forest fire risk zonation mapping in Gia Lai using Deep-NC

Chapter 3: A model for landslide risk zonation mapping using the BBO-DE-StreeEns Ensemble Tree

The dissertation ends with conclusions and future research proposals.

CHAPTER 1: OVERVIEW OF THE RESEARCH FIELD

1.1 Basic theory

1.1.1 Deep learning

Deep learning, inspired by the human brain, is a subset of machine learning. It employs multi-layered artificial neural networks to analyze complex data, forming the foundation for many modern AI applications. Key components of deep learning include: artificial neural networks, weights and biases, activation functions, backpropagation, dropout, CNNs, and RNNs.

Adv.: Effective in areas like language processing and image recognition

Lim.: Requires substantial data and computational resources.

A typical deep learning model consists of an input layer, hidden layers, and an output layer. The depth of the model is determined by the number of hidden layers. This architecture can be applied to forest fire risk zonation

mapping. The hidden layers process input data, while the output layer calculates probabilities for "fire" or "no fire" classes.

1.1.2 Ensemble learning

Ensemble learning is a machine learning approach that combines multiple base models to create a more accurate and optimized predictive model. The fundamental principle is to generate a "strong model" from a group of "weak models".

Main methods in ensemble learning include:

Bagging: Training models on data subsets, combining results.

Boosting: Sequential model training, with subsequent models learning from predecessors.

Stacking: Training diverse models, then training a final model on their outputs.

Voting: Applying multiple models in parallel, with results determined by majority or average.

Blending: Combining base models using optimized weights.

Adv.: Increased accuracy through model combination. Reduced overfitting via averaging of biases and variances, and enhanced model stability.

Lim.: Computationally intensive and not suitable for all problems.

1.2 Some machine learning models

1.2.1 Support Vector Machine (SVM)

SVM is a machine learning algorithm used for both classification and regression tasks. It finds a linear or non-linear boundary to separate data points into different groups. SVM can also handle non-linear data through the use of kernel functions.

1.2.2 Relevance Vector Machine (RVM)

Similar to SVM, RVM is also a machine learning algorithm used for classification and regression. However, RVM uses a small number of important vectors to define the model, helping to reduce computational costs compared to SVM.

1.2.3 Random Forest (RF)

RF is a machine learning algorithm based on building a collection of decision trees. Each tree is built based on a subset of data and uses the "random forest" method to make predictions. RF can handle non-linear data and also has high interpretability.

1.2.4 Logistic Regression (LR)

LR is a machine learning algorithm used for classification problems. It uses a logistic function to estimate the probability of belonging to a certain class. LR is often used when probability information is needed.

1.2.5 Multilayer Perceptron Neural Network (MLP)

MLP is an artificial neural network architecture with multiple hidden layers. It is used for classification and regression problems. MLP combines neuron nodes and weights between them to learn a correlation model between inputs and outputs.

1.2.6 Split Point and Attribute Reduction Classifier (SPAARC)

SPAARC is a decision tree algorithm extended from CART. SPAARC integrates two techniques: node attribute sampling and split point sampling to speed up training without significantly affecting accuracy.

1.3 Optimization algorithm

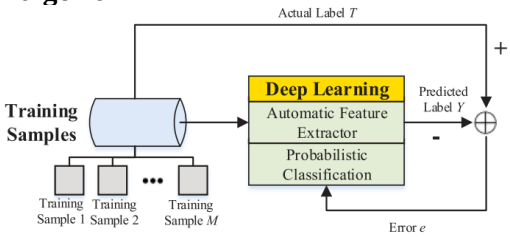

Figure 1-7 General process of network weight adaptation.

One of the significant advantages of this advanced Machine Learning method is the ability to infer high-level representations from initial features. However, the effectiveness of high-level feature extraction in deep neural networks heavily depends on the optimization algorithms used to fine-tune network weights, as illustrated in Figure 1-7. In this context, the dissertation

evaluates several popular optimization algorithms, including SGD, RMSProp, Adam, Adadelta, BBO, and Differential Evolution (DE).

1.3.1 Stochastic Gradient Descent (SGD)

SGD is an optimization algorithm commonly used in machine learning. It updates model parameters using a single training example or a small subset at each iteration.

1.3.2 RMSProp

RMSProp is an adaptive gradient optimization algorithm proposed by Geoffrey Hinton in 2012, aimed at improving the performance of Gradient Descent in training deep neural networks. It automatically adjusts the learning rate for each parameter based on the moving average of recent squared gradients, enabling faster convergence and better handling of sparse gradients in complex deep learning problems.

1.3.3 Adaptive Moment Estimation Algorithm (Adam)

Adam is an adaptive learning rate optimization algorithm, commonly used to train deep neural networks, proposed by Kingma and Ba in 2015. Adam combines the advantages of two other algorithms: AdaGrad and RMSProp.

1.3.4 Adadelta Optimization Algorithm

The Adadelta optimization algorithm, proposed by Zeiler in 2012, addresses the issue of diminishing learning rates over time. It achieves this by limiting the window size of past gradient accumulation. Notably, the Adadelta method eliminates the need to set a default learning rate.

1.3.5 Biogeography-Based Optimization (BBO)

BBO is a meta-heuristic optimization algorithm inspired by the migration process of species between biogeographical islands. BBO simulates the distribution and movement of species based on factors such as Habitat Suitability Index (HSI) and Species Suitability Index (SIV), thereby finding optimal solutions for complex optimization problems in various fields.

1.3.6 Differential Evolution

DE is a powerful evolutionary optimization algorithm developed by Storn

and Price in 1997, using mutation, crossover, and selection operations to search for optimal solutions in multi-dimensional search spaces. DE stands out for its ability to efficiently handle non-linear and multi-objective optimization problems, while being easy to implement with few adjustable parameters, making it a popular tool in many application areas.

1.4 Research methodology

1.4.1 Research process

The process of creating a disaster risk zoning map consists of 11 steps:

Step 1: Data collection

- Collect data from local forest management and disaster prevention agencies about past forest fires and landslides.
- Gather data from satellite images, digital elevation models (DEM), geological maps, and meteorological-hydrological data.
- Conduct field surveys to collect additional information on vegetation and soil structure.

Step 2: Data preprocessing

- Use GIS software to process spatial data.
- Use ArcPy, Python source code to fill in missing values.
- Normalize all variables to the same scale.

Step 3: Analysis and selection of influencing factors

- Perform separate analyses for forest fires and landslides to select influencing factors, including 12 factors affecting forest fire probability and 10 factors affecting landslide probability.
- Identify the most important factors for each type of disaster, using the average impurity decrease (AID) algorithm to evaluate the importance of factors influencing forest fire probability and the wrapper algorithm combined with five-fold cross-validation to assess the role of 10 factors.

Step 4: GIS database construction

- Create geodatabases for each type of disaster in the two study areas.
- Create separate map layers for each influencing factor.

Step 5: Model building and training

- Design models for forest fires (Deep-NC) and landslides (BBO-DE StreeEns).
- Split data into training and testing sets for each model.
- Train the models.

Step 6: Model evaluation and refinement

- Evaluate the performance of both models.
- Refine each model separately to achieve the best performance.
- Compare the performance of the two models with reference models.

Step 7: Model application for prediction

- Use the forest fire model on the geodatabase of Gia Lai province to predict forest fire risk.
- Use the landslide model on the geodatabase of Than Uyen district to predict landslide risk.
- Save the prediction results of both models.

Step 8: Risk level classification

- Determine classification thresholds separately for forest fires and landslides.
- Classify prediction results into risk levels for each type of disaster, using the natural break classification method.

Step 9: Risk zoning map creation

- Create a forest fire risk map.
- Create a landslide risk map.

Step 10: Map verification and evaluation

- Compare prediction results with actual data for both types of disasters.
- Conduct field verification for forest fire risk in Gia Lai province and landslide risk in Than Uyen district, Lai Chau.

Step 11: Map finalization and presentation

- Create risk zoning maps for forest fires in Gia Lai province and landslides in Than Uyen district, Lai Chau.

- Add legends, scales, and other necessary information to the maps.

1.4.2 Data Preparation

- Steps for preparing research data include:

1.4.2.1 Data Collection for Gia Lai Province Forest Fire Dataset

- Forest fire data: 2530 fire locations from 2007-2016 from MARD database at http://www.kiemlam.org.vn
- Topographic data: 1:50,000 scale topographic map from MONRE

Satellite imagery: 2016 Landsat-8 OLI, acquired from EarthExplorer (http://earthexplorer.usgs.gov)

- Climate data (2007-2014): Temperature, wind speed, relative humidity, precipitation. Sourced from NCEI, Website: https://www.ncdc.noaa.gov/
- Land use data: 1:50,000 scale Gia Lai land use map with 11 categories, from 2013 National Land Inventory Project by General Department of Land Administration
- The study combined these diverse data sources to build a GIS database and analyze forest fire risk for Gia Lai province.

1.4.2.2 Data Collection for Than Uyen District Landslide

- Landslide data: 970 landslide points in Lai Chau province, including 114 in Than Uyen district. Sourced from national project by Vietnam Institute of Geosciences and Mineral Resources in 2012
- 1:50,000 scale topographic map of Lai Chau province with 20x20m resolution for Than Uyen district
- Geological and fault maps from Vietnam Institute of Geosciences and Mineral Resources
- Soil map from Vietnam Academy of Agricultural Sciences (VAAS)
- Transportation and river distribution maps from MONRE

1.4.2.3 Data Preprocessing

Preprocessing of Gia Lai forest fire data included:

- Converting unstructured data: Collection, digitization, CSV organization, GIS integration, digital map linking, fire status encoding (0/1), shapefile export
- Handling missing data: Using Python and SimpleImputer() in Scikit-learn
- Feature creation: Using ArcGIS for slope, aspect, elevation, curvature layers from DEM; Calculating NDVI, NDWI, NDMI from satellite imagery
- Model requirement fulfillment: Encoding categorical data numerically
- Data normalization: Using -1 to 1 scaler for all features

Data balancing: Random selection of 2530 non-fire points

- Data splitting: 70% (3542 samples) for training, 30% (1518 samples) for testing

Preprocessing of Than Uyen district landslide data included:

- Using ArcGIS Pro Clip tool to confine data to Than Uyen district boundaries
- DEM construction using “Topo to Raster”
- Creating influencing factor maps (elevation, slope, curvature, aspect, relief) from DEM using ArcGIS Pro Spatial Analysis tools
- Using National Soil Map to create Soil Type and Geology feature layers
- Creating distance-to-fault, road, river feature layers using Distance tool
- Identifying 114 landslide locations in Than Uyen
- Balancing data by randomly sampling 114 non-landslide points
- Splitting 228 points into training (70%) and testing (30%) sets using train_test_split()
- Classifying and normalizing data using Reclassify and Raster Calculator

1.5 Model evaluation

1.5.1 Evaluation methods

Hold-Back Method: Divides data into training and testing sets for model training and evaluation. Also known as Train/Test or Train/Validate split.

Cross-Validation: Splits data into k segments, training the model on k-1 segments and evaluating on the remaining one. This process is repeated k times to ensure each segment is used once for evaluation.

Stratified Cross-Validation: Similar to cross-validation but maintains the same class ratio in each data segment.

Bootstrapping: Generates multiple training sets from the original dataset by random sampling with replacement.

1.5.2 Evaluation Metrics

For binary classification problems, key evaluation metrics include:

Confusion matrix: Represents a matrix of predictions and actual results, with values for TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative).

Accuracy (Acc): Ratio of correct predictions to the total number of samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}};$$

Precision (PPV): Ratio of TP to total positive predictions.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}};$$

Negative Predictive Value (NPV): Ratio of TN to total negative predictions.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}};$$

Sensitivity (Recall): Ratio of TP to total actual positive samples.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}};$$

Specificity: Ratio of TN to total negative predictions.

$$\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}};$$

F1-score: Harmonic mean of Precision and Recall.

$$\text{Fscore} = \frac{2 \times PPV \times Sen}{PPV + Sen} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}};$$

ROC curve and AUC: Graphical and numerical measures of model performance across various classification thresholds.

Kappa coefficient: Statistical measure of agreement between raters.

$$Kappa = \frac{2\times(TP\times TN - FN\times FP)}{(TP+FP)\times(FP+TN)+(TP+FN)\times(FN+TN)}$$

MSE and RMSE: Common regression performance measures.

CHAPTER 2 A MODEL FOR FOREST FIRE RISK ZONATION MAPPING IN GIA LAI USING DEEP-NC

2.1 Forest fire dataset of Gia Lai province

2.1.1 Description of the study area

Gia Lai province is located in the Central Highlands, in the south central region of Vietnam, covering an area of 15,512 km². The terrain is diverse, with elevation gradually decreasing from North to South and from West to East. The climate is tropical monsoon, with two distinct seasons. 93% of the land area is used for agriculture and forestry. Forests cover 41% of the province's area. Over the past decade, the province has regularly suffered damage from forest fires. The forest fire database used in this study includes 2,530 points with fire history from 2007 to 2016, along with 2,530 randomly selected non-fire points, creating a dataset of 5,060 points in total.

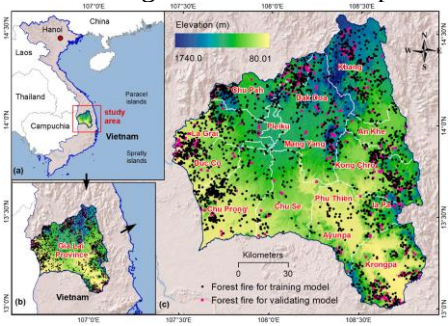

Figure 2-1 Map of Gia Lai province's location in Vietnam (a) and (b); map of Gia Lai province and forest fire points (c)

2.1.2 Forest fire data

Forest fire data was collected from the Ministry of Agriculture and Rural Development database. Data analysis shows that about 90% of fires occur in the dry season from January to May. The dataset includes 2,530 records of historical forest fire events in Gia Lai province over 10 years (2007-2016).

2.1.3 Influencing factors

Factors influencing forest fire occurrence are categorized into three groups:

- Topography: slope, aspect, elevation, curvature.
- Environment: land use, NDVI, NDWI, NDMI indices.
- Climate: temperature, wind speed, humidity, rainfall.

These factors are represented as digital map layers.

2.1.4 Development of forest fire geodatabase for Gia Lai province

The GIS database includes 12 influencing factors and a forest fire distribution map. A total of 5,060 records were preprocessed for model input. Data preprocessing are discussed in section 1.6 of the dissertation.

2.2 Deep-NC model

2.2.1 Model selection

The author evaluated SVM, ICA-RVM, Random Forest, and Deep-NC models by using Trials And Error on the same dataset. Results indicated that Deep-NC exhibited the best performance, thus it was selected for this study.

2.2.2 Evaluating factor importance

The Average Impurity Decrease (AID) method was used to assess the predictive importance of factors influencing forest fire susceptibility. Results in Table 2-3 showed that NDVI, NDWI, and NDMI were the most important factors.

2.2.3 Objective function for training the Deep-NC model

Training the Deep-NC model involves searching for and updating model weights to minimize the difference between predicted forest fire risk points and actual occurrences. An objective function is necessary to quantify this difference. In this study, the author chose Mean Squared Error (MSE) as the objective function, described as follows:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Where 'N' represents the number of data samples, while 'yi' and '$\bar{y}$i' are the actual and calculated values for sample 'i', respectively.

2.2.4 Deep-NC model architecture

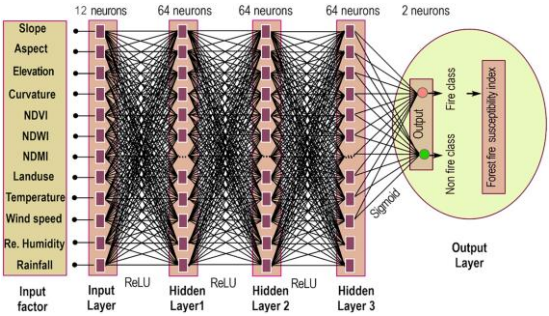

Figure 2-5: Deep-NC model structure used in the study

The Deep-NC model architecture (See Figure 2-5) consists of 5 layers:

Input layer (with 12 neurons) corresponds to 12 factors influencing forest fires, each factor represented by a separate neuron.

Hidden layers (3 layers, each with 64 neurons): These hidden layers learn complex and non-linear representations from input factors. Each neuron in the hidden layers uses the ReLU (Rectified Linear Unit) activation function. ReLU is a common non-linear activation function in deep learning models, with the formula:

$$f(x) = \max(0, x)$$

ReLU returns x if $x > 0$ and 0 if $x \leq 0$. It is often used in hidden layers as it helps reduce the vanishing gradient problem during backpropagation.

Output layer (2 neurons): This layer includes 2 neurons; each neuron represents a partitioned class: "Non-fire class" and "Fire class". The Sigmoid activation function is used in this layer to convert the model's output into probabilities.

$$f(x) = 1 / (1 + e^{(-x)})$$

To fine-tune the weights of the Deep-NC model, the author uses optimization algorithms such as SGD, RMSProp, Adam, and Adadelta. The Deep-NC model helps determine the decision boundary separating the study area map into two distinct types – "non-forest fire" and "forest fire".

Subsequently, the calculation results from the Deep-NC model can be converted into raster format for further analysis using ArcGIS software.

2.3 Model evaluation

The Deep-NC model, comprising 5 layers with 206 neurons, demonstrates high performance in forest fire prediction. The model employs the Adam algorithm to optimize 9,294 weights. When running the model on the test dataset, the AUC results of Deep-NC were compared with the SVM model (AUC = 0.786), RVM model (AUC = 0.793), and Random Forest model (AUC = 0.790). The proposed Deep-NC model, with an AUC value of 0.894, exhibits superior efficacy.

Table 2-1: Confusion matrices of RVM, SVM, Random Forest, and Deep-NC models with Training Dataset and Test Dataset

Index	**RVM**		**SVM**		**Random Forest**		**Deep-NC**	
	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset	Test dataset
TP	1760	344	1822	359	1748	340	1742	673
TN	1494	270	1372	252	1432	259	1637	564
FP	358	68	296	53	370	72	29	86
FN	624	142	746	160	686	153	134	195

Table 2-4: Performance of RVM, SVM, Random Forest, and Deep-NC models with Training Dataset and Test Dataset

Index	**RVM**		**SVM**		**Random Forest**		**Deep-NC**	
	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset	Test dataset
PPV(%)	83.10	83.50	82.53	82.52	86.02	87.14	98.36	88.7
NPV(%)	70.54	65.53	67.61	62.86	64.78	61.17	92.43	74.3
Sens(%)	73.83	70.78	71.82	68.97	70.95	69.17	92.86	77.5
Spe(%)	80.67	79.88	79.47	78.25	82.25	82.62	98.26	86.8

Index	RVM		SVM		Random Forest		Deep-NC	
Acc(%)	76.82	74.51	75.07	72.69	75.40	74.15	95.40	81.5
AUC	0.842	0.793	0.830	0.790	0.813	0.786	0.983	0.894
Kappa	0.536	0.490	0.501	0.454	0.508	0.483	0.908	0.630

2.4 Evaluation of the Deep-NC model with different optimization algorithms

Table 2-7: Confusion matrices of the Deep-NC model using 4 optimization algorithms (Adam, SGD, RMSprop, and AdaDelta) on 10 random sampling cases of the test set.

Index	Indices with Adam algorithm				Indices with SGD algorithm				Indices with RMSprop algorithm				Indices with Adadelta algorithm			
	Min	Max	Mean	STD	Min	Max	Mean	STD	Min	Max	Mean	STD	Min	Max	Mean	STD
TP	570	662	623	35.6	612	719	669	26.9	493	698	588	66.1	540	729	635	54.3
TN	537	661	605	31.9	450	546	491	32.6	419	602	528	57.5	405	565	499	47.8
FP	97	189	137	35.6	40	147	90.2	26.9	61	266	171	66.1	30	219	124	54.3
FN	98	222	154	31.9	213	309	269	31.5	157	340	231	57.5	194	354	260	47.8

Table 2-8: Classification performance of the Deep-NC model using 4 optimization algorithms (Adam, SGD, RMSprop, and AdaDelta) on ten random sampling cases of the Test set

Index	Indices with Adam algorithm				Indices with SGD algorithm				Indices with RMSprop algorithm				Indices with Adadelta algorithm			
	Min	Max	Mean	STD	Min	Max	Mean	STD	Min	Max	Mean	STD	Min	Max	Mean	STD
PPV	75.1	87.2	82	4.69	80.6	94.7	88.1	3.55	65	92	77.5	8.7	71.2	96.1	83.7	7.15
NPV	70.8	87.1	79.7	4.21	59.3	71.9	64.6	4.19	55.2	79.3	69.5	7.6	53.4	74.4	65.8	6.3
Sen	72	86.8	80.2	3.79	69.1	74.2	71.4	1.82	67.2	75.9	72.2	2.6	67.3	73.6	71.2	1.99
Spe	74	86.5	81.7	4.32	78.8	91.8	84.8	3.26	69.4	87.3	76.6	6	72.1	93.1	81.1	6.02
CA	72.9	85.8	80.8	3.74	74.1	78.3	76.4	1.14	72.1	74.5	73.5	0.9	72.8	76.8	74.7	0.96
AUC	0.82	0.93	0.89	0.03	0.81	0.86	0.84	0.01	0.81	0.84	0.82	0.01	0.82	0.85	0.83	0.01

Table 2-7 presents the Confusion Matrices of Deep-NC with 4 optimization algorithms on 10 random samples. False Negative (FN) - the number of cases where the model incorrectly predicts no forest fire - is a crucial index, representing the model's tendency to miss forest fire occurrences. The Area

Under the Curve (AUC) is used as the primary metric to compare the predictive performance of the Deep-NC model with different optimization algorithms. Consequently, the Deep-NC model optimized with Adam outperforms other models in forest fire prediction, achieving the highest mean AUC value of 0.893.

2.5 Thành lập bản đồ phân vùng nguy cơ cháy rừng

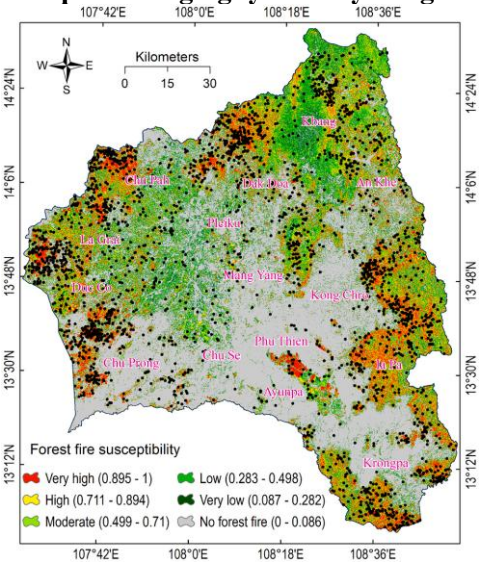

Figure 2-8: Forest fire risk zonation map of the study area using the Deep-NC model with Adam optimization

Utilizing the capabilities of the Deep-NC model, the forest fire zonation index was calculated for the entire province. The zonation results were converted to raster format and ArcGIS Pro was used to create a forest fire risk zonation map (See Figure 2-8). The map was divided into 6 levels ranging from no fire risk to very high risk using the Natural Break method in ArcGIS. Forest areas in districts such as Ia Grai and Chư Păh demonstrate very high fire risk. Conversely, newly forested areas in Mang Yang and Chư Sê exhibit low risk.

CHAPTER 3 A MODEL FOR LANDSLIDE RISK ZONATION MAPPING USING THE BBO-DE-STREEENS ENSEMBLE TREE

3.1 Landslide dataset of Than Uyen district, Lai Chau province

3.1.1 Description of the study area

Than Uyen is a mountainous district in northern Vietnam, characterized by complex terrain and frequently affected by flash floods and landslides.

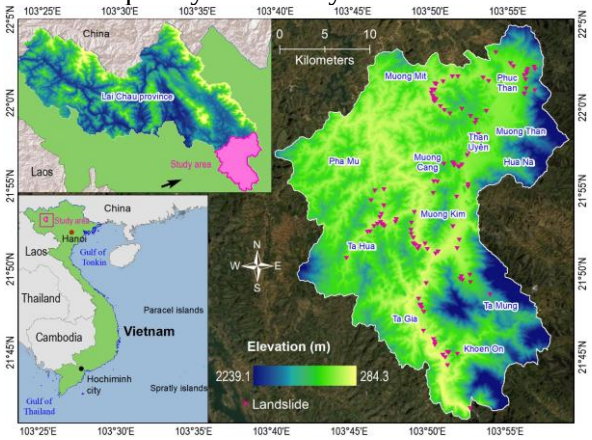

Figure 3-1: Location of Than Uyen district, Lai Chau province

3.1.2 Historical landslide data

The study utilizes a comprehensive map of landslide locations in Than Uyen district, developed within the framework of the project "Investigation, assessment, and warning of areas at risk of landslides in mountainous regions of Northern Vietnam" conducted by the Institute of Geology and Minerals from 2012. This map includes 114 landslide locations from the past 10 years in Than Uyen, primarily occurring along road networks.

3.1.3 Influencing factors

Ten factors influencing landslides were considered:

- Topographic factors: elevation, slope, curvature, aspect, relief.
- Anthropogenic and environmental factors: soil type, geology.
- Distance factors: to roads, rivers, faults.

These factors were represented as digital map layers in ArcGIS Pro.

3.1.4 Developing a Geodatabase for landslides in Than Uyen district

The GIS database was constructed comprising 10 influencing factors and a landslide distribution map. A total of 228 records were preprocessed for model input. The data processing and preprocessing procedures are discussed in section 1.6 of the dissertation.

3.2 BBO-DE-StreeEns hybrid model optimized by BBO-DE

3.2.1 Model proposal

The researcher proposes the BBO-DE-StreeEns model for landslide susceptibility mapping in Than Uyen district, Vietnam, replacing the ineffective Deep-NC model for small datasets (228 records). The new model combines:

- Subbagging and Random Subspace to create subsets of data.
- SPAARC algorithm to construct decision trees, reducing building time by 70% compared to Random Forest.
- Hyperparameter optimization using a combination of BBO and DE

This method aims to improve performance, processing speed, and accuracy compared to the traditional Random Forest algorithm, while addressing the limitations of Deep-NC with small datasets.

3.2.2 Process of creating landslide susceptibility maps using the BBO-DE-StreeEns model

To develop the landslide risk assessment model, we follow these steps as shown in Figure 3-6:

- Collect data on 10 influencing factors: elevation, slope, curvature, aspect, relief, soil type, geology, distance to faults, roads, and rivers (See section 1.6.2.2)
- Build landslide database (See section 1.6.4):

+ Extract values of 10 factors from raster maps
+ Convert discrete data to numerical values for corresponding factors

After collecting the dataset and preprocessing the data, a landslide database is formed with 10 influencing factors and locations of landslide events, creating a set of 228 points (including 114 landslide points and 114 non-landslide points). This data is divided into two sets: training (70%) and validation (30%). At this stage, steps are taken to build the proposed BBO-DE-StreeEns model (Figure 3-7). The steps are as follows:

Step 1: Divide the dataset into training (70%) and testing (30%) sets.

Step 2: Initialize the initial hyperparameters of the BBO and DE algorithms. These hyperparameters include mutation rate, crossover rate, population size, number of iterations, and constraint values for the optimization parameters.

Step 3: Optimize 3 hyperparameters of the model including TotalTrees, SizePercentage, and SubSpaceSize using the hybrid BBO-DE algorithm. These hyperparameters are used to calculate the objective function MAE on the training set.

Step 4: Evaluate the model on the test set. If the results are satisfactory, stop the optimization process. Otherwise, adjust the hyperparameters.

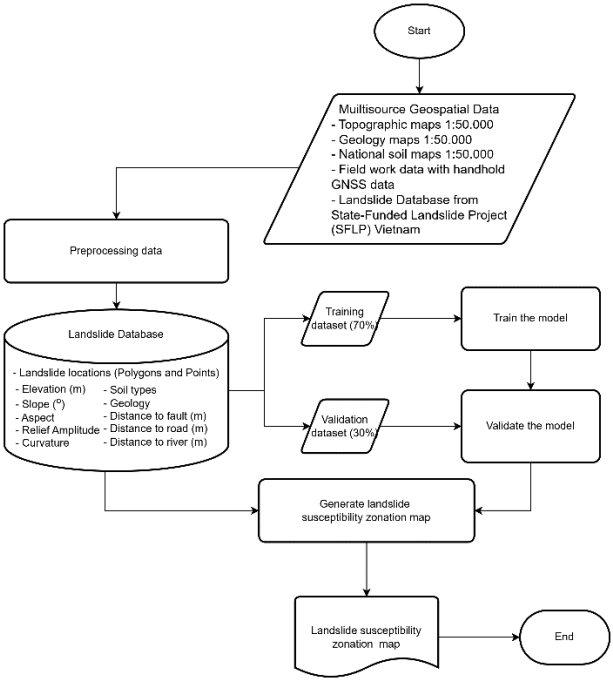

Figure 3-6: Process of developing a landslide risk assessment model

Step 5: Repeat Steps 3 and 4 until the optimal set of hyperparameters is found.

Step 6: The model at this stage is optimal, and the hyperparameters can be used in the model to generate a landslide susceptibility map in the next step.

These parameters are extracted into the ensemble model, after which BBO-DE will optimize three hyperparameters: SizePercentage, TotalTrees, and

subSpaceSize. In this study, they are: SizePercentage = 0.9 (90%), TotalTrees = 30, and subSpaceSize = 0.5 (50%).

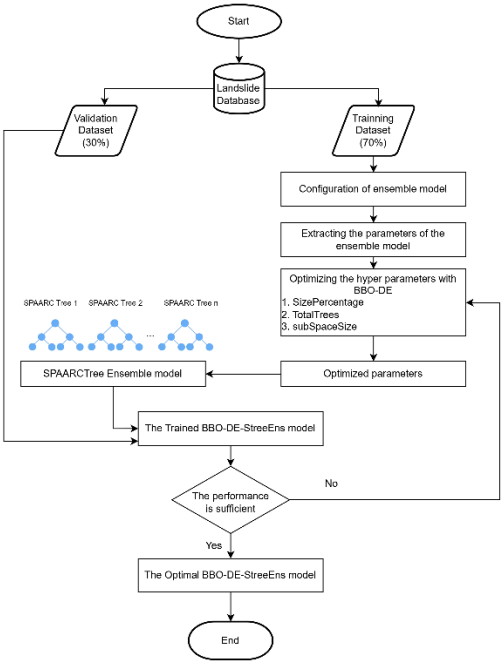

Figure 3-7: Architecture of the proposed BBO-DE StreeEns model

3.2.3 Impact of factors influencing landslides

The contribution of 10 factors influencing landslides in the BBO-DE-StreeEns model was evaluated using the wrapper method and 5-fold cross-validation. Results: slope has the greatest influence (0.299), followed by distance to roads (0.224) and elevation (0.142). The remaining factors have less influence (0.026-0.084).

3.2.4 Cost function and hyperparameters

To achieve optimal performance, the BBO-DE-StreeEns model depends on the selection of 3 hyperparameters: TotalTrees, SizePercentage, and SubSpaceSize.

In this study, these hyperparameters were optimized using the BBO-DE hybrid technique, with Mean Absolute Error (MAE) as the objective function.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| LS_i - \widehat{LS_i} \right|$$

Where: LS_i: Predicted landslide susceptibility value for the i-th sample

$\widehat{LS_i}$: Corresponding actual value.

n: Total number of samples in the training set

3.2.5 Evaluation results

Five models (BBO-DE-StreeEns, LRegr, MLPNeuNet, SVM, SPAARC) were trained using 10-fold cross-validation on 228 points (114 landslides, 114 non-landslides). On the training set, the proposed model (with optimal hyperparameters: TotalTrees=30, SizePercentage=0.9, subSpaceSize=0.5) and SPAARC yielded the best results in terms of metrics (AUC, Kappa, F-score, Accuracy).

Table 3-5: Confusion matrix of the proposed BBO-DE-StreeEns model along with reference models on the test dataset.

Model	**TP**	**TN**	**FN**	**FP**
BBO-DE-STreeEns	28	31	6	3
Lregr	24	28	10	6
MLPNeuNet	26	18	8	16
SPAARC	28	29	6	5
SVM	24	28	10	6

Table 3-6: Performance evaluation metrics of the proposed BBO-DE-STreeEns model and reference models on the test dataset

Model	**Performance indices**							
	PPV (%)	NPV (%)	Sen (%)	Spe (%)	Acc (%)	Fscore	Kappa	AUC
BBO-DE-STreeEns	90.3	83.8	82.4	91.2	86.8	0.862	0.735	0.940
Lregr	80.0	73.7	70.6	82.4	76.5	0.750	0.529	0.853
MLPNeuNet	61.9	69.2	76.5	52.9	64.7	0.684	0.294	0.748
SPAARC	84.8	82.9	82.4	85.3	83.8	0.836	0.676	0.915
SVM	80.0	73.7	70.6	82.4	76.5	0.750	0.529	0.767

On the test set, BBO-DE-StreeEns and SPAARC maintained the best predictive ability. The LRegr, SVM, and MLPNeuNet models performed worse on both training and test sets. Thus, the proposed BBO-DE-StreeEns model is the best among the surveyed models.

3.3.6 Generation of landslide susceptibility map for Than Uyen district

Table 3-8: Indices of 4 landslide susceptibility levels of the proposed model

No.	Susceptibility level	Range of susceptibility index values	Area ratio(%)	Area (km^2)	Number of landslide points	Percentage of landslide points (%)
1	Low	0,062–0,508	50,00	394,5	12	10,53
2	Moderate	0,508–0,606	20,00	157,8	7	6,14
3	High	0,606–0,737	20,00	157,8	23	20,17
4	Very high	0,737–0,910	10,00	78,9	72	63,16
Total			**100**	**789**	**114**	**100**

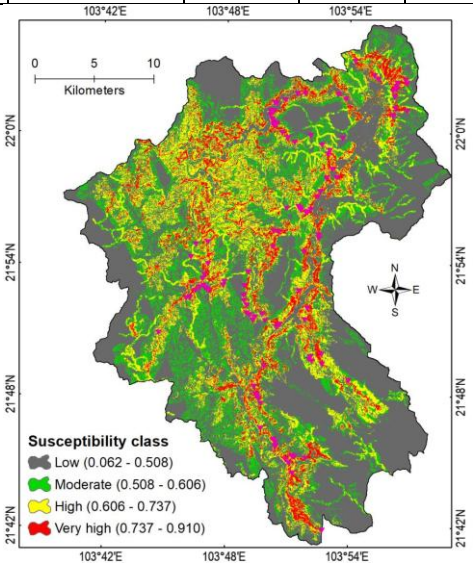

Figure 3-9: Landslide susceptibility map of Than Uyen district using the BBO-DE-StreeEns model

The BBO-DE-StreeEns model was selected to calculate the landslide susceptibility index (0.062-0.910) for each pixel in Than Uyen. The landslide susceptibility map was created by classifying the index into 4 levels using the Natural Break method: very high, high, moderate, and low. The thresholds for these levels were determined through analysis of the graph showing the percentage ratio of landslide area to susceptibility map. Distribution: very high 50% (0.737-0.910), high 20% (0.674-0.737),

moderate 20% (0.502-0.674), low 10% (0.062-0.502). The very high and high categories account for 83.33% of landslide points. The final map is based on these thresholds and classifications.

CONCLUSION

The dissertation presented the process of data collection and processing from Gia Lai province and Than Uyen district, Lai Chau province, while proposing and demonstrating the effectiveness of two important models. The Deep-NC model was applied in forest fire risk prediction, outperforming RVMs, SVMs, and RF models, with the Adam optimization algorithm yielding the best performance. For landslide susceptibility mapping, the BBO-DE-StreeEns model, combining SPAARC Trees and hybrid BBO-DE optimization, demonstrated superior performance, identifying slope and distance to roads as the most significant factors. Both models proved effective in classifying forest fire and landslide risks in Vietnam, with BBO-DE-StreeEns also notable for its fast training speed and high accuracy.

FUTURE RESEARCH PROPOSAL

Future research directions include optimizing the Deep-NC model structure and training algorithm, incorporating additional data such as rainfall into the BBO-DE-StreeEns model, and validating the models' wide applicability across various regions and scenarios. Furthermore, the research aims to integrate Deep-NC and BBO-DE-StreeEns through ensemble learning or multi-task learning, develop Online Learning to update results in real-time, and expand the research to other provinces. Throughout this process, it is necessary to consider the unique characteristics of terrain, climate, vegetation, and the relationships between influential factors in each locality, while comparing and adjusting the models based on similarities or differences between regions.