Landslide Prediction using Machine Learning

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Sahil Thakur (191230)

Under the supervision of

Mr. Prateek Thakral

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

DECLARATION

I hereby declare that the work presented in this report entitled "Landslide Prediction using Machine Learning " in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from Jan 2023 to May 2023 under the supervision of Mr. Prateek Thakral, Assistant Professor, Department of Computer Science & Engineering and Information Technology.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Sahil Thakur
191230

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Prateek Thakral
Assistant Professor
Department of Computer Science & Engineering and Information Technology.
Dated:

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty God for his divine blessing makes it possible for us to complete the project work successfully. I really am grateful and wish my profound indebtedness to Supervisor Mr. Prateek Thakral, Assistant Professor, Department of CSE Jaypee University of Information Technology, Waknaghat. Deep Knowledge and keen interest of my supervisor in the field of “Machine learning” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to Mr. Prateek Thakral, Department of CSE, for his kind help to finish my project. I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Sahil Thakur

TABLE OF CONTENT

LIST OF ABBREVIATIONS

Abbreviations	Full Form
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
cGAN	Conditional Generative Adversarial Network
LAB	L*: Lightness A*: Red/Green B*: Blue/Yellow
RGB	R*:Red G*:Green B*:Blue
HTML	Hyper Text Markup Language
CSS	Cascading Style sheets

ABSTRACT

In Korea, during the rainy season, landslides are a frequent geological hazard that can cause fatalities, property damage, and economic losses. Landslides are responsible for at least 17% of all fatalities from natural disasters worldwide and nearly 25% of all fatalities from natural disasters in Korea each year. Global climate change has increased the frequency of landslides, which has led to an increase in landslide-related losses and damages. Therefore, it is essential to perform exact landslide prediction, monitoring, and early warning of ground movements in order to reduce the losses and damages caused by landslides. There has been significant recent progress in the fields of landslide prediction and landslide damage reduction as a result of the numerous studies that have been undertaken in these fields.

The use of information and geospatial technology, especially remote sensing and geographic information systems, has greatly improved recent studies on landslide hazard assessment (GIS). The most recent advancements and the state of the art in the critical fields of landslide susceptibility analysis, runout modelling, landslide monitoring, and early warning were covered in this article. This paper's major objective was to evaluate landslide susceptibility using a probabilistic approach and a physically based method. Other subjects included in the study included the evaluation of runout using a volume-based and dynamic model, on-site ground monitoring methods, remote sensing methods for landslide monitoring, and physical and rainfall thresholds for landslide early warning

Chapter-1: INTRODUCTION

1.1 Introduction

A form of natural disaster known as a landslide occurs when rocks or land tumble down a mountainside. The majority of the time, landslides occur after or immediately following a period of significant rainfall. Because the earth is moist and heavy, the adhesion force between the soil masses is reduced. Landslides can occur swiftly or slowly.

Existing people or structures that restrict soil flow could possibly endanger lives and harm property. According to statistics on the issue, the bulk of landslides in Himachal Pradesh take place in Bilaspur, Chamba, Hamirpur, Kangra, Kinnaur, Kullu, Lahaul and Spiti, Mandi, Shimla, Sirmaur, Solan, and a few other locations.

According to analysis of data from the HP Government's prior years, the significant slides in Himachal Pradesh that caused enormous damage are:

- The Maling (1968). This avalanche, which is still in motion, damaged 1 km of NH-22.
- The Kinnaur (Dec.1982) This happened at Sholding Nala, when three bridges collapsed and 1.5 miles of road disappeared.
- The Jhakri (March 1989) A 500 m stretch of road in Nathpa was damaged by this avalanche and is still in use.
- The avalanche in Luggarbhati on September 12, 1995, buried 65 people alive (39 according to the official account).
- At Marhi, Bhang, Chhyal, and Mandu in the upper catchment of the Beas River are prominent slides.

Combining a variety of topographical data for analysis has allowed for the identification of areas at risk for landslides. The occurrence of a landslide is a complicated process with many interwoven aspects. In earlier decades, established processes for assessing the susceptibility of a landslide mainly depended on statistically supported and data-driven methodologies. The methods of LR, SVM, frequency proportion, and scale of entropy models are a few that can be used to forecast potential landslides. However, a significant variety of samples are necessary for these procedures to be effective. accuracy in prediction Landslides can gravely destroy both persons and property, despite the fact that they occasionally happen in arbitrary locations.

There are several issues with conventional landslide prediction, including:

(1) a significant proportion of existing understanding and assumptions are frequently required by the models.
(2) The underlying nonlinear characteristics of landslides cannot be sufficiently characterised by the networks.
(3) The networks are too small to take into account sub-regional correlations.
(4) Forecasting accuracy is decreased by the models' over-fitting, labor-intensive computing, vulnerability of falling into edge points, and receptivity to missing data.

1.2 Problem Statement :

Landslides are the most expensive geohazard in the world, and they regularly cause, aggravate, or directly result from other hazards and disasters, such a earthquakes, tsunamis, wildfires, and volcanic eruptions. A mass of soil, rock, debris, mud, or other earth materials—including vegetation—moving down a slope is referred to as a landslide. Gravitational pull causes them to appear in areas with steep slopes and poor relief. Other names for landslides include slump and slope failure. Thus, it is basically necessary to foresee landslides so that the local government and natural disaster management authorities can take preventative action before they happen. This will be useful for civil engineers as well because more care will be paid while designing bridges, tunnels, and transportation routes in these places..

1.3 Objectives :

● Understanding the various factors (parameters) and correlation among them that led to landslide.

● Design the dataset for the selected specific area.

● Building an efficient classification model which can predict the probability of occurrence of landslide of a particular region for which the model is validated, and the result will be "high potential" or "low potential" of landslide occurrence.

1.4 Language Used :

Python 3 has been used in this project.

The main reasons for the use of this programming language for the implementation of the project are stated below:

Simple and dependable: Python stimulates readable and condensed code. Python's simplicity enables programmers to create dependable solutions because machine learning and machine intelligence are based on complex algorithms and adaptable workflows. Developers can concentrate solely on resolving an ML issue rather than on the language's finer technical details.

1.5 Python Libraries Used:

- Numpy
- Pandas
- Matplotlib
- Tensorflow
- Kaggle
- Tensorflow_io
- Pathlib
- Shutil

1.6 Technology used :

- **HTML** : The most fundamental component of the web is HTML, or Hypertext Markup Language explains the purpose and organisation of web content. Other than HTML, JavaScript and CSS are frequently employed to describe the look, feel, and functionality of online pages. Links that connect online pages on different websites or inside the same website are referred to as "hypertext." An essential component of the web are links. By posting your stuff online and linking to other people's pages, you may engage actively there in World Wide Web.

  To mark text, pictures, and also other information for output in a web browser, HTML uses "markup."

- **CSS :** Cascading Style Sheets (CSS) is a language for creating style sheets that govern how HTML or XML documents should look . The rendering of items on paper, a screen, in audio, or in other medium is described by CSS. According to W3C guidelines, CSS is among the fundamental languages of both the open internet and is defined for use by all web browsers. The development of multiple CSS specification components used to be coordinated, providing versioning of the most recent recommendations. Perhaps you are familiar with CSS1, CSS2.1, and perhaps CSS3. There was never a CSS3 or CSS4. Instead, CSS is now completely unversioned.
- **JAVASCRIPT :** A just-in-time compiled or lightweight programming language with first-rate capabilities is called JavaScript (JS). Although it is most famous for being a javascript library for online pages, numerous non-browser contexts, like Jquery, Apache Couchbase, and Adobe Acrobat, also use it. JavaScript is a reactive languages that covers object-oriented, declarative, and prescriptive styles and is prototype-based, multi-paradigm, and single-threaded (such as functional programming). Study up on JavaScript. The JavaScript language itself, as opposed to individual webpages and other hosting environments, is the focus of this section. See Application Programming interfaces and DOM for web-specific APIs.
- **REACTJS :** For creating reusable UI components, Javascript library is a declarative, effective, and adaptable JavaScript toolkit. This front-end component package, which is open source, is completely in charge of handling your application's view layer. It was initially created and managed by Facebook and later incorporated into services like WhatsApp and Instagram.

Chapter-2: LITERATURE SURVEY

- ## Landslide prediction from machine learning

One of the biggest challenges in studying natural disasters and their mitigation is identifying how and when disasters are likely to happen in a particular area of interest. Physical and mathematical models can explain the underlying processes of embankment failure start and runoff, but there aren't any accurate real-time observations of the soil, rock, and groundwater conditions. Predicting landslides accurately is difficult. To predict future landslide disasters based on historical distribution patterns, researchers are increasingly examining multivariate data analysis methods from the disciplines of machine learning and information mining. Despite the fact that this work clarified spatial sensitivity patterns, temporal prediction remained essentially empirical. The majority of machine learning techniques have a 75–95 percent overall success rate. The performance of the input information, the possibility for overfitting and corresponding underselection of forecasting analytics, the unintentional integration of duplicated or noisy different factors, and the unintentional inclusion of specific types and sizes of forecasting analytics all pose risks despite how promising this may appear. Technical limits for only predicting landslides continue to be a problem. Only slightly inferior forecasts are made by simpler models than by more complicated models, which should open the door for more extensive data mining applications in area landslide prediction. Planners and decision-makers should be especially informed about this strategy. Future analysis might take into account the following options: (1) Additional best practise recommendations for model choice. (2) forecast regionally the frequency and flow of significant bank failures; (3) Temporal projection of Landslides.

- **Spatial prediction models for landslide hazards: review, comparison and evaluation**

  The predictive power of logistic regression, support vector machines, and bootstrap aggregate classification trees (bagging, double bagging) are compared using the misclassification error rates of independent test datasets. Based on a resampling approach that takes spatial autocorrelation into account, error rates are estimated for predicting 'current' and 'future' landslides inside and outside the exercise area. A case study from the Ecuadorian Andes shows that logistic regression with stepwise backward selection of variables has the lowest error rate and the best generalization ability. Evaluations outside the training region show that tree-based methods tend to overfit the data.

- **Shallow Landslide Prediction Using a Novel Hybrid Functional Machine Learning Algorithm**

  To forecast the geographic variation of landslides in Iran's Salhoun Basin, we used a brand-new hybrid workable machine learning algorithm. We created the ABSGD model for landslide prediction by fusing stochastic descent (SGD), functional algorithms, and AdaBoost (AB) metaclassifiers. 20 slope stability adjustment factors are included in the model, which was ranked and use the least-squares vector machine (LSSVM) method. The modelling took 98 landslide sites into account. 30% (19) were used for the validation process, and 70% (79) were utilized for training. Sensitivity, specificity, accuracy, mean squared root error (Standard deviation), and area underneath the receiver operation characteristic (AUC) curve were used to validate the model. For validation and comparison, we also used softer computing benchmark models including Ssg, regression analysis (LR), gradient boosted tree (LMT), and component tree (FT) methods. Despite the importance of the chosen conditioning elements for landslide occurrence, the distance to the road was revealed to be the most crucial element. Its LR (0.797), SGD (0.776), Shift to the left (0.740), & FT (0.734) algorithms all performed worse than the ABSGD model (AUC=0.860). Our findings demonstrate that even a single SGD approach for geographical landslide predictions performs better when functional algorithms & meta-classifiers are used in combination to prevent overfitting, reduce noise, and enhance performance prediction. is affirmative.

Application of an enhanced BP neural network model with water cycle algorithm on landslide prediction

Mudslides caused enormous damage to the living environment and seriously threatened the The local population's safety in terms of their lives and property was greatly threatened by mudslides, which severely damaged the living environment. Direct management of landslides and the mitigation of these hazards through proper preparedness measures both benefit greatly from assessment of landslide inclination or prediction of avalanche displacement by monitoring. In order to improve the hydrologic cycle algorithm and make up for the absence of BPNN convergence speed, we created a unique BP computational model (BPNN) dynamic modeling approach (WCA-BPNN). Four years collected displace monitoring data were being used for moment analysis and prediction on the Langshuwan landslide, a typical incremental landslide that happened in the Three Gorges River part of China. The accumulating shifts are divided into two types trend shifts and cyclical shifts, and the model takes into consideration both the short-term speeding effects of climate change and the long-term insidious effects of landslides. The Gray relational analysis approach was utilised to screen key variables influencing landslide-induced periodic displacement, and the results were employed as training data. We assess the accuracy of his three models—BPNN, Support Vector Machine, and Extreme Learning Machine—under training settings using the same learning dataset in addition to comparing the simulated and experimental values produced by the models. The WAC-BPNN model is the most accurate of the four models, according to the results, and has a greater prediction performance than the traditional His BPNN model.

- **Senslide: a distributed landslide prediction system**

Discover more about the Senslide distributed sensor system's conception, development, and present status for forecasting collapses in the hillsides of western India. This region experiences landslides during the monsoon season, which results in significant property loss and fatalities. Our objective is to anticipate landslides before they happen, in contrast to current methods that detect them in this area. Our approach includes a large number of affordable sensor nodes linked by a wireless network, in contrast to earlier efforts that employed a few pricey sensor nodes to determine slope stability. Because such cheap components can introduce more faults, our system software is built to tolerate them.

We tested the concept on a modest scale with 65 sensor nodes on a lab test bench. The findings of this test bench are presented, along with simulation results for a larger system containing up to 400 sensor nodes. The system will be put to the test in the field during the Indian rainy season thanks to our positive results.

Method	Classification Accuracy	Sensitivity	Precision	AUC	F1-Score
Logistic Regression[12]	51.40%	51.40%	26.40%	50%	34.8%
RF[13]	80.70%	80.70%	80.70%	86.90%	80.70%
AdaBoost[14]	83.40%	83.40%	83.40%	83.30%	83.40%
KNN[15]	86.50%	86.50%	88.60%	93.50%	86.20%
SVM[16]	92.70%	92.70%	93%	92.20%	92.70%
Proposed Approach	96.90%	96.90%	96.90%	99.20%	96.90%

Table 2.1 Accuracy Assessment

- **Landslide identification using machine learning**

Identification of landslides is crucial for risk analysis and mitigation. Using an unified geospatial database, we suggest new deep learning and machine learning techniques in this study for predicting landslides in naturally occurring terrain. Gather information regarding the landslide, such as topographical, geology, and precipitation data, first. The creation of three integrated geodatabases follows. These three databases are the Joint Erdslide Database, the Relict Erdslide Database, and the Recently Erdslide Database (RecLD) (JLD). Then, for each database, five deep learning and deep learning algorithms were employed and assessed, namely regression models (LR), support vector regression (SVM), random forests (RF), boosted methods, and convolutional neural networks (CNN) utilised in**.** A case study is carried out at Lantau Island, Hong Kong, to show how the suggested method can be used. Due to its advantages in feature extraction and number of co two-dimensional data processing, CNN outperforms other algorithms in the case study findings, achieving 92.5percentage detection performance on his RecLD. Following RF, LR, and SVM in order of accuracy is the boosting technique. The suggested landslide identification method exhibits remarkable robustness and tremendous potential in resolving the landslide detection algorithm by utilising machine learning algorithms techniques.

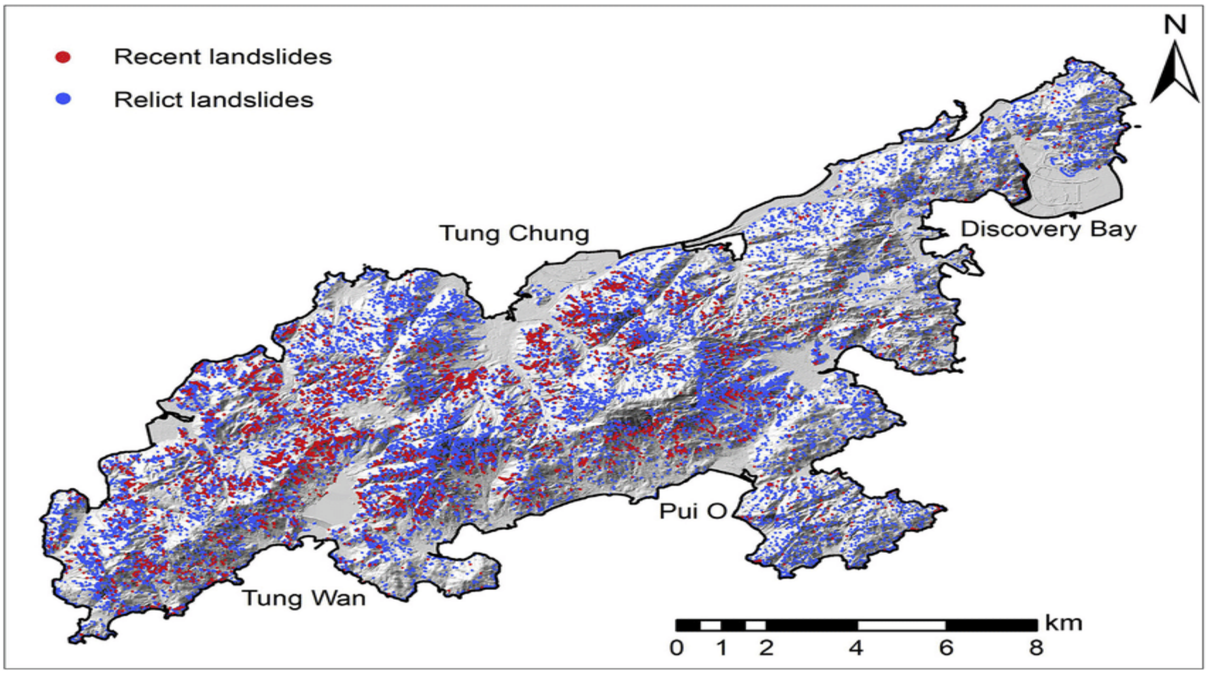


Fig 2.1 Recent landslides & Relict landslides

- **Machine Learning-Powered Rainfall-Based Landslide Predictions in Hong Kong—An Exploratory Study**

More over 7 million people live in the 1100 km2 metropolitan area of Hong Kong. The land is hilly, with over 30% of it being steeper than 30 degrees and around 63% of it being steeper than 15 degrees. As a result, infrastructure and population growth are concentrated close to naturally occurring and artificially created slopes that are vulnerable to landslides during powerful storms. As a result, the Geotechnical Office (GEO), which has been in charge of managing urban landslide risk since [1], must continue to improve this strategy. GEO has gradually enhanced a database of national landslide samples with high geographical and temporal precision over the past four decades. The region has a complex web of rain gauges (about one rain gauge every 10 km 2) enabling real-time rainfall measurements at 1-minute intervals, a thorough inventory of previous landslides, and a LiDAR-based digital topography model for the area. It exists. The creation of useful landslide prediction models, which are crucial for efficient landslide risk management, was supported by the utilisation of this high-quality data. In Hong Kong, landslide susceptibility mapping [2] and avalanche early warning [3] were conducted using the landslide prediction model. These models have largely been created up to this point utilising classic statistical methods and data-driven analysis.

Recently, GEO investigated the use of machine learning (ML) algorithms to predict landslides, making use of the adaptability of ML techniques and the abundance of landslide-related data gathered over the years. An exploratory study was started on ML-based evaluation of landslides vulnerability in natural terrain. This work presents details of the investigation, an expansion of his research released in Li et al. (2022).

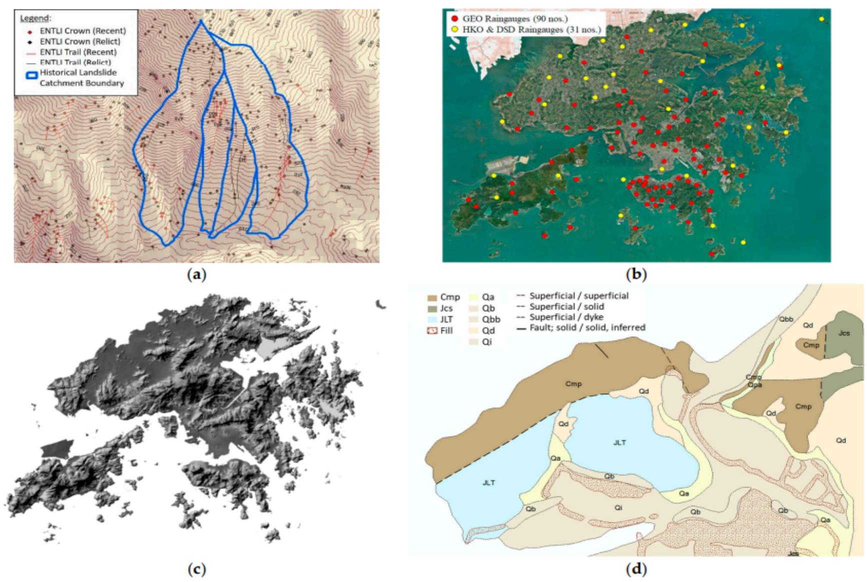


Fig 2.2 landslides vulnerability in natural terrain

- **Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia**

  Support Vector Networks (SVM), Conditional Random fields (RF), Adaptive Regression Multivariate Grooves (MARS), Artificial Neural Systems (ANN), and Nonlinear Discriminant Analysis are the seven machine learning advanced methods (MLT) that are the subject of his current study. In order to simulate landslide vulnerability, it was planned to assess the functioning of (Classifier), Linear Discriminant Assessment (LDA), and Multilayer Perceptron (NB). It is crucial to understand how to map landslide susceptibility utilising both geographical data types and machine learning algorithms. Utilizing GIS and R open-source software, the study was carried out in the Abha Delta of the Asir region of Saudi Arabia. In order to produce a landslide inventory map, a maximum of 243 landslide locations were initially found in the Abha Basin. All landslide regions were separated into two groups at random, with 70% used for training and 30% used for validation. Elevation, rock properties, distance to fault, normalised vegetation indices (NDVI), land use/land covering (LULC), proximity to road, gradient, distance to stream, and curvature of profile are among the twelve landslide factors that are generated. Slope orientation, gradient (LS), and plane curvature.To assess, validate, and contrast MLT performance, we was using the area beneath the curve (AUC-ROC) method. The findings demonstrated that the 7 MLTs' AUC scores ranging from 89.0percentage points for QDA to 95.1percentage points for RF. According to our findings, Lf (AUC=95.1%) and Ld (AUC=941.7%) outperformed other MLTs. The landslide hazard map and the study's findings will aid in environmental protection. n.

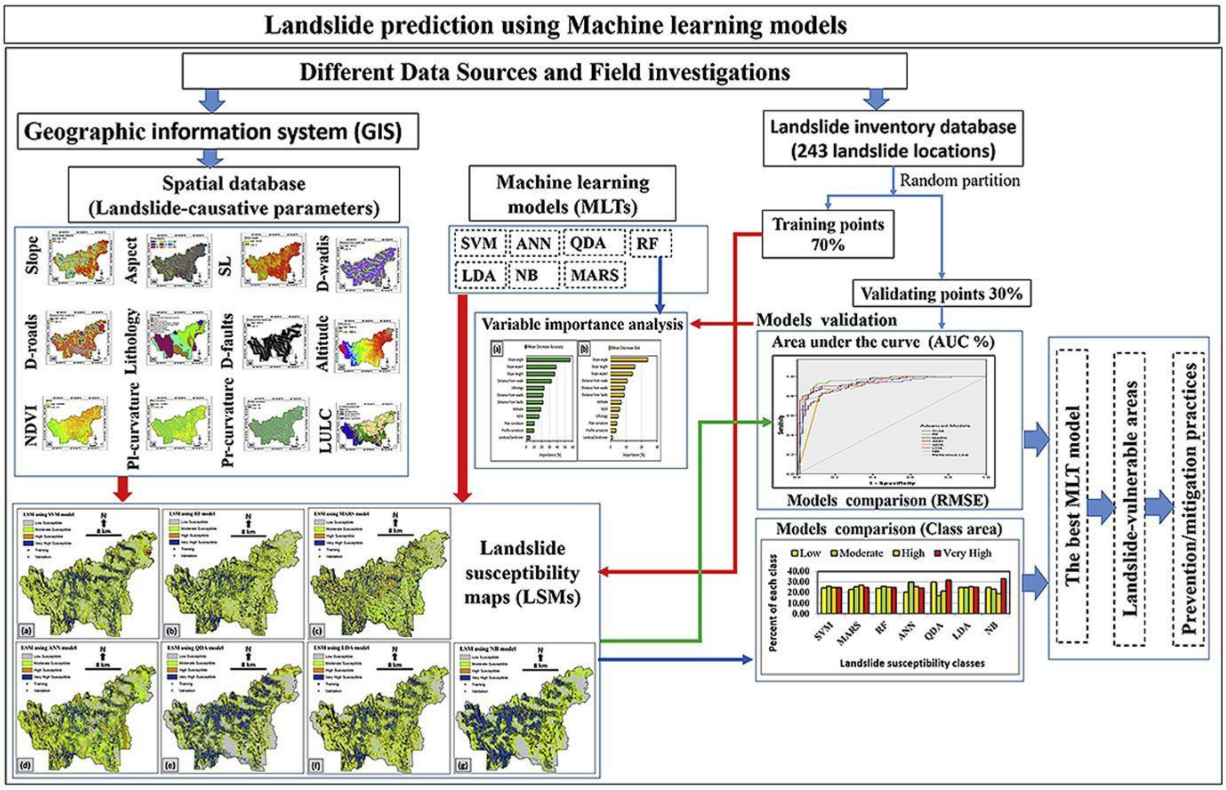


Fig 2.3 Landslide Prediction using ML model

- **Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia**

Support Vector Models (SVM), Conditional Random fields (RF), Multivariate Adaptive Grooves (MARS), Artificial Neural Systems (ANN), and Nonlinear Discriminant Analysis are the seven sophisticated machine learning methods (MLT) that are the subject of his current study. In order to simulate landslide vulnerability, it was planned to assess the functioning of (QDA), Lda Analysis (LDA), and Bayesian Network (NB). It is crucial to understand how to map landslide susceptibility utilising both geographical data types and machine learning algorithms. Employing GIS and Matlab open source, the study was carried out from the Abha Basin of the Asir region of Saudi Arabia.. In order to produce a landslide inventory map, a maximum of 243 landslide locations were initially found in the Abha Basin. All landslide regions were separated into two groups at random, with 70% used for training and 30% used for validation. Altitude, lithology, proximity to fault, normalised crop yields (NDVI), lands use/land cover (LULC), proximity to road, slope, proximity to stream, and curvature of profile are among the twelve landslide factors that are generated. Slope orientation, slope length (Lb), and plane curvature. To assess, validate, and contrast MLT performance, we used the area under a curve (AUC-ROC) method. The findings demonstrated that the 7 MLTs' AUC scores ranging from 89.0percentage points for QDA to 95.1percentage points for RF. According to our findings, RF (AUC=95.1%percantage) and LDA (AUC=941.7%) outperformed other MLTs. The landslide hazard map and the study's findings will aid in environmental protection.

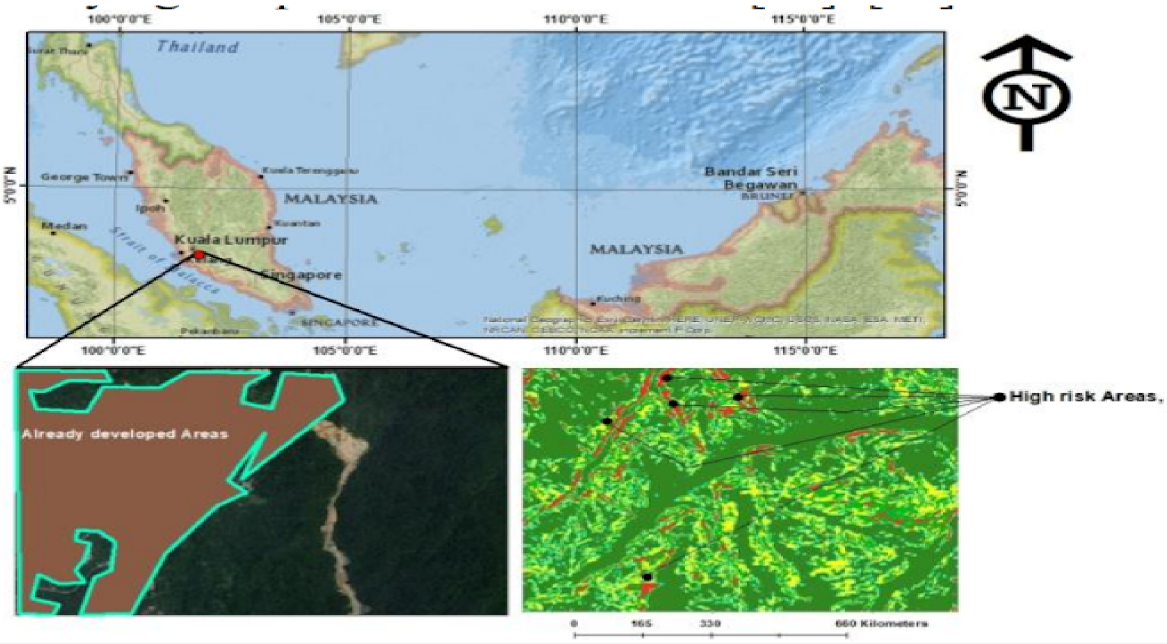


Fig 2.4 Landslide susceptibility mapping using machine learning algorithms

- **A Research on Deep Learning Advance for Landslide Classification using Convolutional Neural Networks**

Simply said, a landslide is when big boulders, rubble, or terrain slip down a slope due to gravity. In a wide range of habitats, including mountains, coastal reefs, and undersea, slides can occur on both gentle and severe slopes. Slope stability is impacted by landslide-causing factors such earthquakes, torrential rain, floods, and slope building. Landslides are a growing concern that is anticipated to get worse in the future due to a number of factors, including uncontrolled urbanisation, deforestation, and unexpected increases in rainfall. In Marin village, Karnataka, a landslide happened around July 30, 2014, killing 151 people and left hundreds unknown. Over 5,700 people perished in landslides brought on by the Kedarnath flooding in Uttarakhand in July 2013. Future landslide disasters are anticipated to rise as a result of the consequences of severe rainfall and weather changes. As a result, one of the most popular research areas is landslides. Landslide estimation is a challenging practical issue. It is important to understand the time and space distribution of avalanches and how they might be controlled to lessen damage and dangers to people & their property. Landslides are a threat to lives and businesses in her country. There have previously been significant initiatives Lessen them.

A novel method of landslide detection is called landslide classification. Surveys for landslides can assist find them and give early warning indications, so precautions should be implemented right away. Recent developments in deep learning have improved the processing of images, videos, audio, and speech. enables the learning of computational models with multi-layered processing using data abstractions at several levels. Deep learning models use backpropagation to demonstrate how machines adjust weights in order to detect patterns in datasets. It is used to analyze how each layer's data representation from earlier layers looks.

Fig 2.5 landslide in hilly areas

Chapter 03: SYSTEM DEVELOPMENT

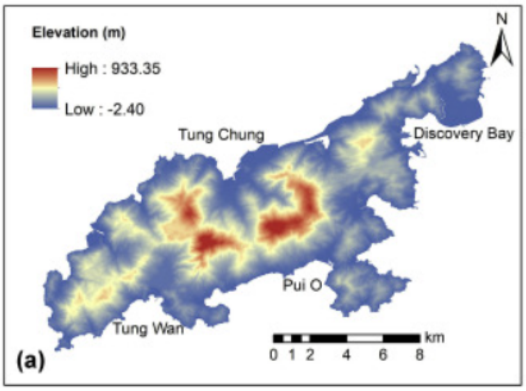
3.1 Date Set Used in this Project:

Enhanced Natural Terrain Landslide Inventory.

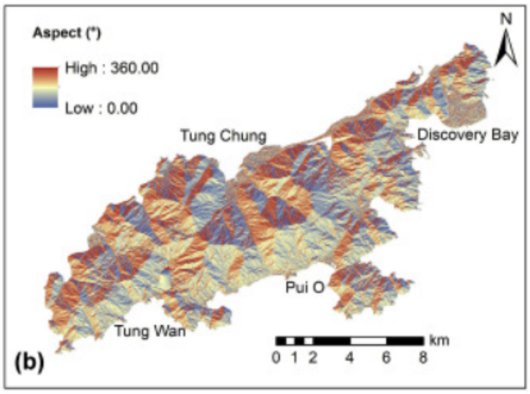
3.2 Date Set Features

Geotechnical Engineering Office has created a landslides inventory called "Enhanced Natural Terrain Landslide Inventory (ENTLI)" (GEO 1996, Maunsell Fugro Joint Venture & GEO, 2007). In the research area, the ENTLI records both recent and relict landslides from 1924 to 2009, and Fig. 6 shows how they are distributed spatially. The inventory of landslides in research region shown in Table . There are 20,884 relict landslides and 5810 current ones. The three types of recent landslides are channelized landslides, open hillslope landslides & costal landslides. These are determined through visual interpretation of aerial pictures. Based on API and topographical characteristics, the relict landslides that happened before the time period seen in the aerial pictures are identified and divided into classes A and B. or C, with varying levels of interpretation assurance.

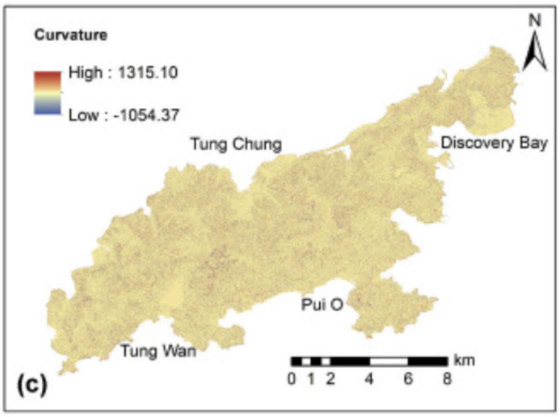
1. There is a separate class for coastal landslides, similar to the recent disasters. The Relict Landslide Inventory (RelLI), the Recent Landslide Inventory (RecLI), and the Joint Landslide Inventory are three landslide inventories that are created using the ENTLI data, as stated in the methodology (JLI).



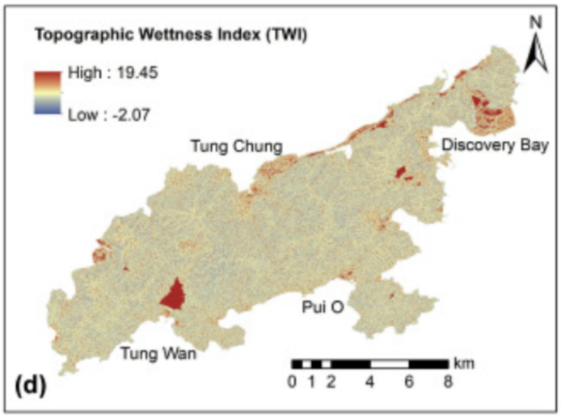
(a) elevation



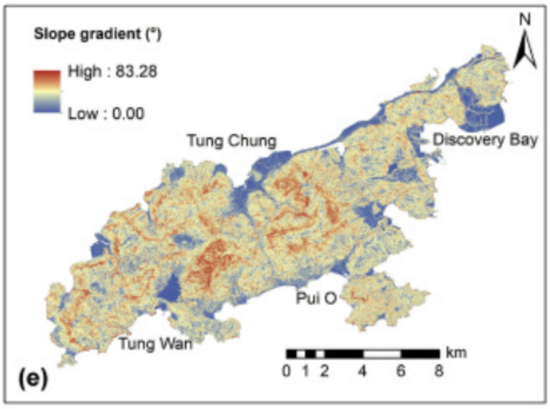
(b) aspect



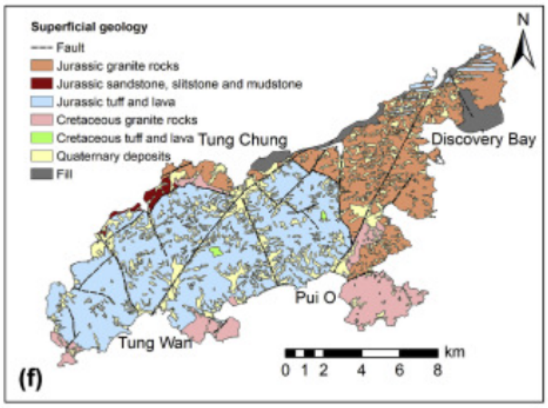
(c) curvature



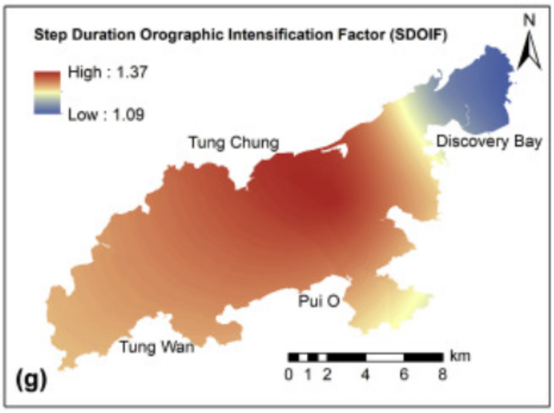
(d) TWI

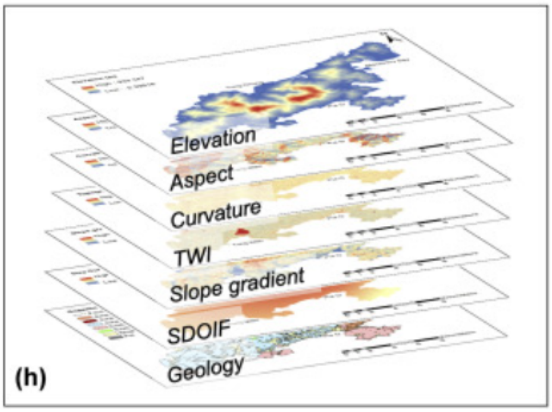


(e) slope gradient



(f) superficial geology





(g) SPDIF (h) all layers stacked

Recent landslides		Relict landslides	
Class	**No. of landslides**	**Class**	**No. of landslides**
Channelized	2593	A	6480
Open hillslope	3144	B	9007
Costal	73	C	4529
		Costal	868
Subtotal	5810	Subtotal	20,884

Table 3.1 provides a summary of the 26,694 landslides included in the study area's Enhanced Natural Terrain Landslide Inventory.

3.3 PREPARATION OF DATABASES

To combine the data into a multi-layer dataset, a layer stacking procedure is used after the production of the seven predictor data layers (Fig. 7h). The data from the entire study region are arranged in this phase into a tensor form with dimensions of 12,637 by 9737 by 13, which is a tensor. The initial superficial geology data layer is expanded to 7 dimensions as the dummy variable approach to the superficial geology layer is used. There are a total of 13 layers when the additional six data layers are included.

The three landslide inventories (RelLI, RecLI, and JLI) are individually mapped on the research region after the creation of the fundamental dataset, and data extraction is completed in each case. The entire research area is divided into cells of 2 m by 2 m, as was previously described. Landslide cells can then be verified in accordance with the mapped landslide sites after the landslide mapping. The size of the data extraction area, or the sample size, must then be decided. The sample size is set to be 22 m 22 m in order to reduce the computational burden of model training in later stages while capturing as many terrain features as possible because the widths of 93.03% of the landslide records in this case study are smaller than 22 m. Therefore, by using Fig. 2b and c as a guide, a 22 m 22 m area is used to conduct data extraction for each positive or negative sample. Positive samples are taken from areas with landslide cells, indicating that there are landslides that have been recorded in the inventory, and negative samples of the same number are taken at random from areas without landslide records.

Fig. shows examples of the positive and the negative sample. It should be noted that the superficial geology layers are not processed using the training variable approach for the convenience of the demonstration. Every positive sample typically contains a distinct landslide scar (shown in Fig. 8 with red lines), but no landslide characteristics can be seen in a negative sample. Following data extraction, three landslide databases—the Recent Landslide Database (RecLD), the Relict Landslide Database (RelLD), and the Joint Landslide Database—are created by combining the samples

(JLD). Few coastal landslides occur closer to the coast than 22 meters, and those records are not included in the databases.

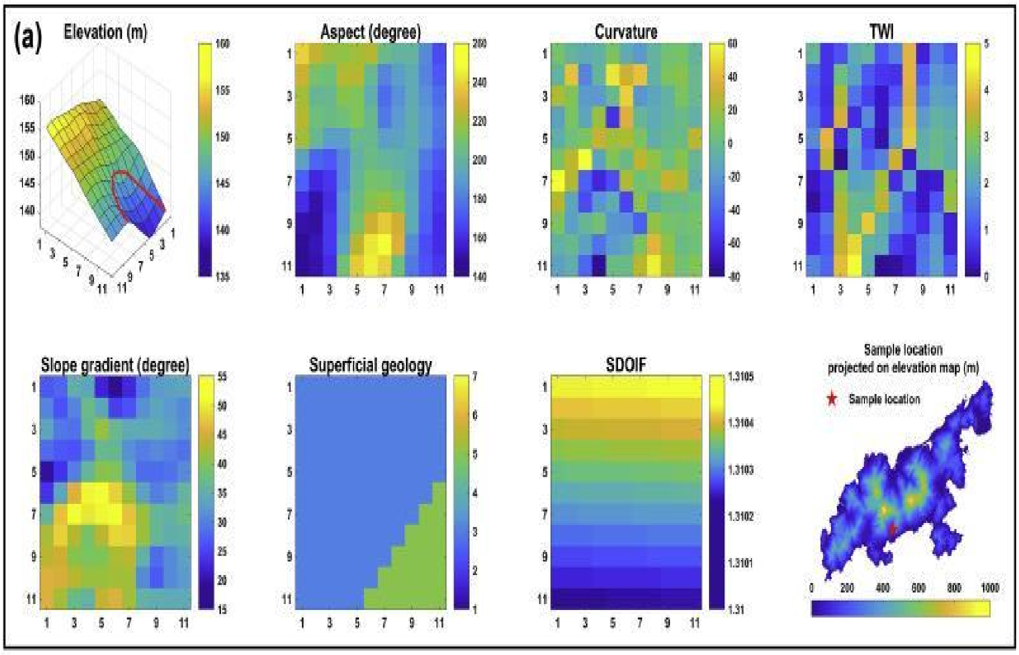


Fig 3.2 positive sample

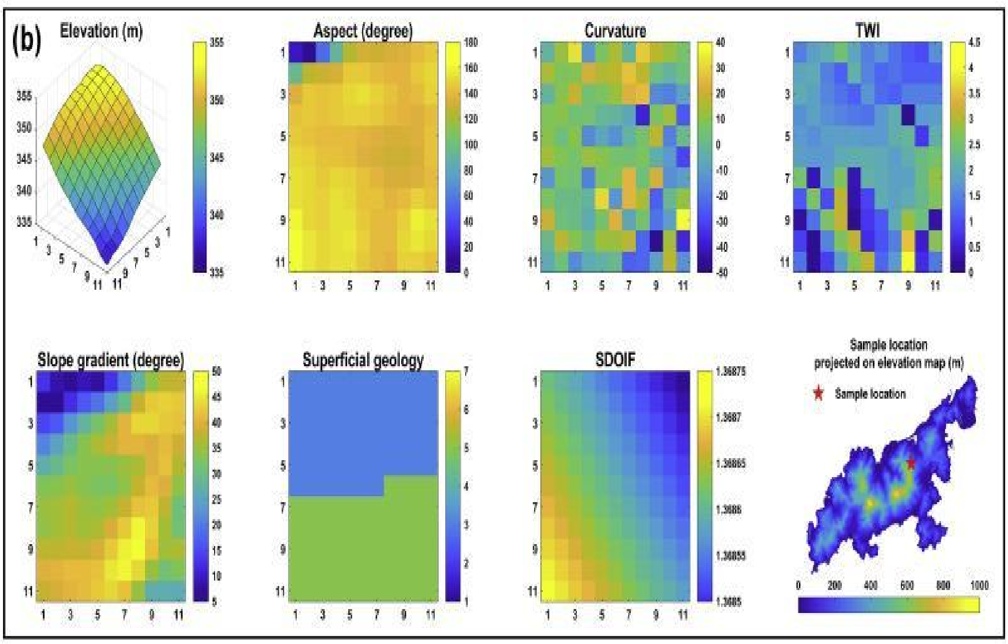


Fig 3.3 negative sample

3.4 Algorithm / Pseudo code of the Project Problem

1) **Logistic Regression:**

Logistic regression, one of the most used classification methods, functions effectively with only a modest amount of processing (James et al., 2013). For multivariate logistic regression, the logistic function is expressed as:

$$P(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_m}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_m}} \qquad \ldots\ldots\ldots(3.2)$$

where: X = (X1, …, Xm) = m predictors

set of β = training parameters ,

P(X) = probability.

Given:

predictor X = positive.

Maximum likelihood methods are often used to estimate training parameters.

2) **Support Vector Machines (SVM)**

Support Vector Machines are another popular classification approach developed in the 1990s. It is considered one of the most adaptable algorithms due to its excellent performance in various environments. Support vector machines use kernels to extend the feature space and quantify the similarity of two observations to obtain nonlinear decision boundaries for classification (James et al., 2013). For example, formula. (3) describes a general kernel called the radial kernel:

$$K\left(x_i, x_{i'}\right) = \exp\left(-\lambda \sum_{j=1}^{m} \left(x_{ij} - x_{i'j}\right)^2\right)$$

………… 3.3

where xij and xi'j are the i-th observation pair of the j-th predictor and m is the number, λ is the tuning parameter used for smoothness. is the decision boundary and K represents the core functionality.

3) **Random Forest:**

Random forest is basically a tree-based method that combines a large number of decision trees to arrive at a single consensus prediction, thereby showing good and reliable prediction performance. A key feature of random forests is their inability to consider the majority of available features in each split of the tree (Friedman et al., 2001). For example, a common way to choose the number of divisions is

$$m' = \sqrt{m}$$

. where m' is the number of predictors in each fold and m is the number of all predictors.

4) **Boosting**

Boosting, a subset of the ensemble approach, combines the output of a handful of weak classifiers into the a powerful "committee" and is hence regarded as a strong classifier. The Discrete AdaBoost prediction, for instance, is stated as follows:

$$F(x) = sign\left(\sum_1^M c_m f_m(x)\right)$$

..........(3.5)

where $fm(x)$ = weak classifier which creates positive predictions or negative predictions,

cm = coefficient calculated by understanding bout weights,

M = declaration of weak classifier.

The function returns a positive or negative prediction and F(x) is the corresponding prediction. By combining multiple models, boosting techniques can achieve better predictive performance compared to single models. The main step in boosting is to fit a set of weak learners (discriminant analysis, k-nearest neighbors, decision trees, etc.) to a weighted version of the training data (Friedman et al., 2000). In this study, three general boosting algorithms were chosen to implement landslide identification. Discrete AdaBoost, LogitBoost and Gentle AdaBoost.

5) Convolutional Neural Networks (CNN)

One of the most popular deep learning algorithms, convolutional neural networks have received much attention due to their notable contributions to computer vision (Goodfellow et al., 2016). A typical CNN consists of four main components: convolutional layers, activation layers, pooling layers, and fully connected layers. Based on these layers, many well-designed CNN structures have been proposed in many research areas. In this study, each sample with different data planes is like a special " image " with multiple channels. Classification of high-dimensional data is also a specialty of CNN.

3.5 Model Evaluation

This study evaluates the performance of each model using seven performance metrics. To compute these seven metrics, we need to clarify the concept of four types of prediction samples for classification learning.

1) False Positive (FP): The predicted class is positive, but the prediction does not match the actual class.
2) True Negative (TN): The predicted class is negative, but the prediction matches the actual class.
3) False Negative (FN): The predicted class is negative and the prediction does not match the actual class.

Seven performance measures, i. H. precision, precision, recall, specificity, false positive rate (FPR), F1 score, and Matthews correlation coefficient (MCC), are defined as:

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

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{F_1\ score} = \frac{2\cdot\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

3.6 Flow chart

fig 3.4 flow chart for landslide prediction using ML

3.7 Screenshots of the various stages of the Project

1. Preprocessing functions for images:

1. **Load image:** Reads and decodes a jpeg file from a given path as tensor variable.

```
@tf.function()
def load_img(target_img_path):
  target_img = tf.io.read_file(target_img_path)
  target_img = tf.io.decode_jpeg(target_img, channels=3)
  target_img = tf.image.resize(target_img, (IMG_WIDTH, IMG_HEIGHT))
  target_img = tf.cast(target_img, tf.float32)

  return target_img
```

2. **Resize:** Resizes the given image tensor according to the provided height and width.

```
@tf.function()
def resize(target_img, height, width):
  target_img = tf.image.resize(
      target_img, (width, height),
      method=tf.image.ResizeMethod.NEAREST_NEIGHBOR
  )

  return target_img
```

3. **RGB to LAB:** Converts an RGB image (tensor) into LAB colorspace image for both visualisation and neural network.

```
@tf.function()
def lab2rgb(lab_img, isnormalized = False, comp = ''):
  if isnormalized:
    tf.Assert(tf.reduce_any(tf.equal(comp, ['vis', 'net'])), data=[comp], name='Lab_Normalization_Error')
    if comp == 'vis':
      lab_img = lab_img * [100.,255., 255.] + [0, -128, -128]; # from 0-1 range
    else:
      lab_img = (lab_img + [1.,0., 0.]) * [50., 127.5, 127.5]; # from -1 to 1 range

  # Take LAB Image in Unnormalized Form.
  rgb = tfio.experimental.color.lab_to_rgb(lab_img)
  return rgb
```

1. Model:

1. Base Layer:

- **Down Sampling Layer:**

Data input is mapped into the a lower dimension latent representation using this layer.

```
def downsample(filters, kernel_size, batch_normalization = True, activation = True):
  pipeline = tf.keras.Sequential()

  pipeline.add(tfa.layers.SpectralNormalization(
      tf.keras.layers.Conv2D(
        filters=filters, kernel_size=kernel_size,
        strides=2, padding='same',
        use_bias = False
      ), 10
  ))

  if batch_normalization:
    pipeline.add(tf.keras.layers.BatchNormalization())

  if activation:
    pipeline.add(tf.keras.layers.LeakyReLU())

  return pipeline
```

Fig Down Sampling

- **Up Sampling Layer:**

This layer serves as a mapping between input data and the desired output in a low-dimensional latent form.

```
def upsample(filters, kernel_size, dropout=False, activation = True):
  pipeline = tf.keras.Sequential()

  pipeline.add(tfa.layers.SpectralNormalization(
      tf.keras.layers.Conv2DTranspose(
          filters, kernel_size, strides=2,
          padding = 'same',
          use_bias = False
      ), 10
  ))

  if dropout:
    pipeline.add(tf.keras.layers.Dropout(0.5))

  if activation:
    pipeline.add(tf.keras.layers.ReLU())

  return pipeline
```

2. Discriminator:

```
def Discriminator():
  inp = tf.keras.layers.Input(shape=(128,128,1))
  tar = tf.keras.layers.Input(shape=(128,128,2))

  x = tf.keras.layers.concatenate([inp, tar])

  down1 = downsample(32, 4, batch_normalization = False)(x)        # (64, 64, 32)

  down2 = downsample(64, 4, )(down1)                               # (32, 32, 64)

  #down3 = downsample(256, 4, )(down2)

  conv1 = tf.keras.layers.Conv2D(128, 4, strides = 1, use_bias = False, padding = 'same') (down2)
  batchnorm1 = tf.keras.layers.BatchNormalization()(conv1)
  leaky_relu1 = tf.keras.layers.LeakyReLU()(batchnorm1)            # (32, 32, 128)

  #down4 = downsample(256, 4, )(down3)

  #zpad1 = tf.keras.layers.ZeroPadding2D()(down3)

  conv2 = tf.keras.layers.Conv2D(256, 4, strides = 1, use_bias = False, padding='same')(leaky_relu1)
  batchnorm2 = tf.keras.layers.BatchNormalization()(conv2)
  leaky_relu2 = tf.keras.layers.LeakyReLU()(batchnorm2)            # (32, 32, 256)

  last = tf.keras.layers.Conv2D(1, 4, strides=1, padding='same')(leaky_relu2)
                                                                  # (32, 32, 1)
  return tf.keras.Model(inputs=[inp, tar], outputs=last)
```

Chapter 04: PERFORMANCE ANALYSIS

4.1 Discussion on the Results Achieved

5 Performance evaluation of eight deep learning and machine learning algorithms on RecLD.

Machine learning models	Accuracy on training set	Accuracy on test set	Precision on test set	Recall on test set	Specificity on test set	FPR on rest set	F_1 score on test set	MCC on test set	OA on test set
LR	0.8492	0.8315	0.8283	0.8283	0.8346	0.1654	0.8283	0.6629	2.3227
SVM	0.7940	0.7726	0.8349	0.6690	0.8725	0.1275	0.7428	0.5542	2.0696
RF	0.9861	0.8396	0.8226	0.8729	0.8050	0.1950	0.8470	0.6801	2.3667
Discrete AdaBoost	0.9175	0.8529	0.8554	0.8554	0.8503	0.1497	0.8554	0.7056	2.4139
LogitBoost	0.9014	0.8592	0.8470	0.8826	0.8350	0.1650	0.8644	0.7188	2.4424
Gentle AdaBoost	0.9217	0.8589	0.8579	0.8661	0.8514	0.1486	0.8620	0.7177	2.4386
CNN-6	0.9168	0.8875	0.8837	0.8968	0.8779	0.1221	0.8902	0.7749	2.5526
DCNN-11	0.9309	0.8932	0.9258	0.8507	0.9343	0.0657	0.8866	0.7886	2.5685

Table 4.1 Performance evaluation of eight deep learning and machine learning models using RelLD.

Machine learning models	Accuracy on training set	Accuracy on test set	Precision on test set	Recall on test set	Specificity on test set	FPR on rest set	F_1 score on test set	MCC on test set	OA on test set
LR	0.8229	0.8181	0.8197	0.8107	0.8254	0.1746	0.8152	0.6362	2.2695
SVM	0.7295	0.7267	0.7620	0.6509	0.8010	0.1990	0.7021	0.4573	1.8861
RF	0.9899	0.8140	0.8079	0.8307	0.7968	0.2032	0.8191	0.6279	2.2610
Discrete AdaBoost	0.8704	0.8383	0.8323	0.8430	0.8337	0.1663	0.8376	0.6766	2.3526
LogitBoost	0.9092	0.8412	0.8396	0.8416	0.8425	0.1575	0.8406	0.6841	2.3659
Gentle AdaBoost	0.9212	0.8360	0.8326	0.8366	0.8353	0.1647	0.8346	0.6719	2.3425
CNN-6	0.8995	0.8725	0.8759	0.8649	0.8800	0.1200	0.8704	0.7450	2.4879
DCNN-11	0.9012	0.8750	0.8809	0.8642	0.8856	0.1144	0.8725	0.7500	2.4975

Fig 4.2 Comparison of the performance of eight machine learning and deep learning models on JLD.

4.2 Limitation of the Project

Even if the suggested strategy performs well in the case study, there are certain drawbacks that need to be mentioned. First, the DTM used was recorded in 2011, whereas the ENTLI covers landslides from 1924 to 2009. This suggests potential discrepancies between the terrain data and the landslide data. This can cause the performance of the suggested method to be underrated. Second, as already mentioned, the ENTLI records for relict landslides are not entirely accurate. Third, when more powerful computing is available in the future, CNNs with more layers should be researched.

4.3 Future Work

The accuracy of this can be increased and enhanced by using advanced deep learning algorithms and this project can be improved into a real time prediction system.

References

- Keefer DK Landslides caused by earthquakes. Geological Society of America Bulletin. 1984 Apr 1 95(4)406-21

- Malamud BD, Turcotte DL, Guzzetti F and Reichenbach P 2004 Landslides, earthquakes, and erosion. Earth and Planetary Science Letters Dec 30229(1-2)45-59

- Collins BD and Znidarcic D2004Stability analyses of rainfall induced landslides Journal of geotechnical and geoenvironmental engineering Apr130(4)362-72

- Tsaparas I, Rahardjo H, Toll DG and Leong EC2002Controlling parameters for rainfall-induced landslides Computers and geotechnics Jan 1 29(1)1-27

- Wang G and Sassa K 2003Pore-pressure generation and movement of rainfall-induced landslides: effects of grain size and fine-particle content Engineering geology Apr 1 69(1-2)109-25

- Chandrasekaran SS, Owaise RS, Ashwin S, Jain RM, Prasanth S and Venugopalan RB2013 Investigation on infrastructural damages by rainfall-induced landslides during November 2009 in Nilgiris India Natural hazards Feb 1 65(3)1535-57

- Senthilkumar V, Chandrasekaran SS and Maji VB 2018Rainfall-induced landslides: case study of the Marappalam landslide Nilgiris District Tamil Nadu India International Journal of Geomechanics18(9)05018006

- Dikshit A, Sarkar R, Pradhan B, Segoni S and Alamri AM 2020 Rainfall induced landslide studies in Indian Himalayan region: A critical review Applied Sciences Jan10(7)2466

- Sengupta A, Gupta S and Anbarasu K 2010Rainfall thresholds for the initiation of landslide at Lanta Khola in north Sikkim India Natural hazards Jan 152(1)31-42

- Chen W, Peng J, Hong H, Shahabi H, Pradhan B, Liu J, Zhu AX, Pei X and Duan Z2018 Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County Jiangxi Province China Science of the total environment Jun 1 6261121-35

- Zhu X, Xu Q, Tang M, Nie W, Ma S and Xu Z 2017 Comparison of two optimized machine learning models for predicting displacement of rainfall-induced landslide: A case study in Sichuan Province China Engineering geology 23(218) 213-22