

**Machine Learning & GIS Based
Landslide Susceptibility Modelling
for the Part of Himachal Pradesh, India**

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Civil Engineering

by

Pathak Yashodhar Pravinkumar

(Enrolment No. 189999912029)

Under supervision of

Dr. M.B. Dholakia

Retd. Principal

Lukhdhirji Engineering College, Morbi

Gujarat Technological University, Ahmedabad, Gujarat, India



**GUJARAT TECHNOLOGICAL UNIVERSITY,
AHMEDABAD**

[July- 2025]

© Pathak Yashodhar Pravinkumar

DECLARATION

I declare that the thesis entitled “*Machine Learning & GIS Based Landslide Susceptibility Modelling for the Part of Himachal Pradesh, India*” submitted by me (Pathak Yashodhar Pravinkumar), for the degree of Doctor of Philosophy is the record of research work carried out during the period from 2018 to 2025 under the supervision of Dr. M.B. Dholakia and this has not formed the basis for the award of any Degree/Diploma, Associate ship, Fellowship, Titles in this or any other university other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities, if noticed in the thesis.

Signature of the Research Scholar :  Date: 24 July 2025

Name of Research Scholar : **Mr. Pathak Yashodhar Pravinkumar**

Place : GTU, Chandkheda, Ahmedabad

CERTIFICATE

I certify that the work incorporated in the thesis “*Machine Learning & GIS Based Landslide Susceptibility Modelling for the Part of Himachal Pradesh, India*” submitted by **Mr. Pathak Yashodhar Pravinkumar** was carried out by the candidate under my supervision. To best of my knowledge, (i) candidate has not submitted this research work to any other institution for any degree/diploma, Associateship, Fellowship, or other similar titles (ii) The thesis submitted is record of original research work done by research scholar during the period of the study under my supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

Signature of the Supervisor : 

Date: 24/07/2025

Name of Supervisor : **Dr. M.B. Dholakia**

Place : GTU, Chandkheda, Ahmedabad

Course Work Completion Certificate

This is to certify that Mr. Pathak Yashodhar Pravinkumar enrolment No. 189999912029 is a PhD scholar enrolled for the PhD program in the Civil Engineering Branch of Gujarat Technological University, Ahmedabad.

(Please tick the relevant option(s))

- ☐ He/she has been exempted from the course work (completed during the M.Phil. course)
- ☐ He/she has been exempted from the Research Methodology Course work (successfully completed during M.Phil. course)
- ☒ He/she has successfully completed the PhD coursework for the partial requirement for the award of the PhD degree. His/her performance in the PhD course work is as follow,

Grade Obtained in Research Methodology (PH001)	Grade Obtained in Self-Study Course (PH002)
AA	AB

Supervisor's Sign: 

Name of Supervisor: **Dr. M.B. Dholakia**

Originality Report Certificate

It is certified that PhD Thesis titled "*Machine Learning & GIS Based Landslide Susceptibility Modelling for the Part of Himachal Pradesh, India*" by Mr. Pathak Yashodhar Pravinkumar enrolment No. 189999912029 has been examined by us. We undertake following:

- a. Thesis has significant new work/knowledge as compared to already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
- b. The work presented is original and own work of the author (i.e there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.
- c. There is no fabrication of data or results which have been compiled/analysed.
- d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
- e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per GTU Plagiarism Policy and instructions issued from time to time (i.e., permitted similarity index $\leq 10\%$).

Signature of the Research Scholar :



Date: 24/07/25

Name of Research Scholar :

Mr. Pathak Yashodhar Pravinkumar

Place :

GTU, Chandkheda, Ahmedabad

Signature of the Supervisor :



Date: 24/07/2025

Name of Supervisor :

Dr. M.B. Dholakia

Place :

GTU, Chandkheda, Ahmedabad

Plagiarism Report



Yashodhar Pathak

**Machine Learning & GIS Based Landslide Susceptibility
Modelling for the Part of Himachal Pradesh, India**

- Quick Submit
- Quick Submit
- Gujarat Technological University

Document Details

Submission ID
trn.oid.:1:3230046532

139 Pages

Submission Date
Apr 27, 2025, 3:07 PM GMT+5:30

27,753 Words

Download Date
Apr 27, 2025, 3:18 PM GMT+5:30

167,242 Characters

File Name
Final_YPP_25042025_Plagiarism_1.pdf

File Size
8.6 MB



5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **167 Not Cited or Quoted 4%**
Matches with neither in-text citation nor quotation marks
-  **1 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 3%  Internet sources
- 3%  Publications
- 0%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

PhD Thesis Non-Exclusive License to GUJARAT TECHNOLOGICAL UNIVERSITY

In consideration of being a PhD research scholar at Gujarat Technological University, and in interests of the facilitation of the research at the University and elsewhere, I, Pathak Yashodhar Pravinkumar, enrolment no. 189999912029 hereby grants non-exclusive, royalty free and perpetual license to the University on the following terms,

- a) The University is permitted to archive, reproduce, and distribute my thesis, in whole or in part, and/or my abstract, in whole or in part (referred to collectively as ‘work’) anywhere in the world, for non-commercial purposes in all forms of media;
- b) The University is authorized to sub-lease, sub-contract or procure any of the acts mentioned in paragraph (a);
- c) The University is authorized to submit the Work at any national or international library under the authority of their “Thesis Non-Exclusive License;”
- d) The University Copyright Notice (©) shall appear on all copies made under the authority of this license;
- e) I undertake to submit my thesis, through my university, to any library and archives. Any abstract submitted with the Thesis will be considered to form part of the thesis.
- f) I represent that my thesis is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this non-exclusive license.
- g) If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission of the copyright owners is required, I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection.
- h) I understand that the responsibility for the matter as mentioned in the paragraph (g) rests with the authors / me solely. In no case shall GTU have any liability for any acts/ omissions / errors / copyright infringement from the publication of the said thesis or otherwise.
- i) I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to my University in this non-exclusive license.
- j) GTU logo shall not be used /printed in the book (in any manner whatsoever) being published or any promotional or marketing materials or any such similar documents.

- k) The following statement shall be included appropriately and displayed prominently in the book or any material being published anywhere: "The content of the published work is part of the thesis submitted in partial fulfilment for the award of the degree of Ph.D. in Civil Engineering of the Gujarat Technological University".
- l) I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my University in this nonexclusive license. I shall keep GTU indemnified from any and all claims from the Publisher(s) or any third parties at all times resulting or arising from the publishing or use or intended use of the book / such similar document or its contents.
- m) I am aware of and agree to accept the conditions and regulations of Ph.D. including all policy matters related to authorship and plagiarism

Date: 24/07/25

Place: GTU, Chandkheda, Ahmedabad



Signature of the Research Scholar

Recommendation of the Research Supervisor: Yes, Recommended

Recommendation of the Co-Supervisor (if any): Yes, Recommended



Signature of the Research Supervisor



Signature of the Co-Supervisor

Thesis Approval Form

The Viva-voce of the PhD thesis submitted by Mr. Pathak Yashodhar Pravinkumar (Enrolment no. 189999912029) entitled "*Machine Learning & GIS Based Landslide Susceptibility Modelling for the Part of Himachal Pradesh, India*" was conducted on 24th July, 2025 at Gujarat Technological University, Ahmedabad.

(Please tick the relevant option)

- ☒ Performance of the candidate was satisfactory. We recommend that he/she be awarded the PhD degree.
- ☐ Any further modification in research work recommended by the panel after 3 months from the date of first viva-voce upon request of the supervisor or request of independent research scholar after which viva-voce can be reconducted by the same panel again.

(Briefly specify the modifications suggested by the panel)

- ☐ The performance of the candidate was unsatisfactory. We recommend that he/she should not be awarded the PhD degree.

(The panel must give justifications for rejecting the research work)


Dr. M. B. Dholakia

Name and signature of Supervisor with Seal


July, 24, 2025
Dr. Binh Thai Pham

(External Examiner 2) Name and Signature


Dr. B. J. Shah

(External Examiner 1) Name and Signature

M. Prasad Rao
Prasad Rao, July-24-2025

(External Examiner 3) Name and Signature (if any)

Abstract

Landslides pose a major threat in mountainous regions, causing substantial infrastructural damage, environmental degradation, and loss of life. Accurate Landslide Susceptibility Mapping (LSM) is vital for effective disaster preparedness and risk mitigation. This study explores and compares the performance of several machine learning (ML) models—Logistic Regression (LR), Support Vector Machine (SVM), Kernel SVM (K-SVM), K-Nearest Neighbours (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN)—along with ensemble strategies for LSM in a part of Himachal Pradesh, India. Traditional statistical methods like Frequency Ratio (FR) and Analytic Hierarchy Process (AHP) were also applied for comparison.

To accomplish this task, A landslide inventory of 1561 locations was compiled using Geological Survey of India datasets and additional sites digitized from Google Earth and Landsat 8 imagery. Various factors influencing landslides such as slope, aspect, elevation, curvature, rainfall, NDVI, Topographic Wetness Index (TWI), Stream Power Index (SPI), geology, geomorphology, soil, land cover, distance to lineaments, distance to roads and distance to streams were selected for spatial landslide susceptibility modelling. Further, an extensive feature selection approach was applied to identify the most significant factors.

The performance of each model was assessed using evaluation metrics such as Area Under the Curve (AUC), accuracy, k-fold cross-validation, precision, recall, and Cohen's Kappa. XGBoost achieved the highest AUC (0.984), while Random Forest exhibited the best overall performance and recorded the highest Landslide Density Index (LDI) value in the 'very high' susceptibility zone (3.79). The pixel-wise distribution across susceptibility classes revealed significant variability among the models, with the DT model assigning the highest proportion to the 'Very Low' class (45.16%) making it most conservative, while the KNN and NB models showed relatively greater emphasis on the 'Very High' class (22.96% and 22.16% respectively). In contrast, models like LR and RF showed a more balanced distribution across all classes, suggesting a more moderate and nuanced assessment of landslide susceptibility. Furthermore, zonal analysis of landslide distribution identified the part of Kullu as the most critical area, exhibiting the highest landslide-area ratio (1.58) and risk index (3.14), indicating intense vulnerability within a limited geographic region.

*Dedicated
To
My Family*

Acknowledgment

At the outset, I extend my sincere appreciation to my guide, Prof. (Dr.) M. B Dholakia, (Retd. Principal, L.E College, Morbi) whose invaluable guidance, expertise, and unwavering support were instrumental throughout this research journey. Your insightful feedback and encouragement pushed me to sharpen my thinking and greatly improved the quality of my research work.

I am equally indebted to my co-guide, Dr. Indra Prakash (Bhaskar Acharya Institute for Space Applications and Geo-Informatics - BISAG), for your constant advice and guidance. Your knowledge and suggestions helped to refine the objectives of my research and brought clarity. Your support has been instrumental in completing this project successfully.

My heartfelt appreciation goes to my Doctoral Progress Committee (DPC) members, Prof. (Dr.) S. D Dhiman ((Birla Vishvakarma Mahavidyalaya, Anand) and Prof. (Dr.) A. K Singh (Adani Institute of Infrastructure Engineering (AIIE), for their expert guidance and valuable inputs throughout the research process. Your perspective and recommendations greatly enhanced the quality of this work, and I am truly grateful for your contributions.

I am profoundly grateful to Government Engineering College, Bhavnagar, India for allowing me to use PARAM Shavak which is a compact, energy-efficient and High-Performance Computing (HPC) system developed by C-DAC (Centre for Development of Advanced Computing), India and provided by Gujarat Council on Science and Technology (GUJCOST). PARAM Shavak was instrumental in executing the computationally intensive aspects of my research. The access to this advanced computing facility enabled me to perform complex ANN and machine learning models that formed the core of my research work.

This academic journey would not have been possible without the emotional support of my parents. Thank you for your unconditional love, patience, and belief in my abilities. To my wife, Amisha, words cannot express my gratitude for your unwavering support at each stage of this research journey. You have been my anchor, my motivation, and my strength. And to my beloved daughter, Keya, your bright smiles, and innocent curiosity have been my greatest inspiration.

With profound reverence, I offer my deepest gratitude to my spiritual guru, Sri Kalgiribapu, Junagadh, whose divine wisdom, blessings, and spiritual guidance illuminated my path during moments of doubt and difficulty. Your teachings on perseverance, clarity of thought, and selfless dedication to knowledge have been the invisible force guiding this intellectual pursuit.

I am fortunate to have been surrounded by friends who made this journey more enjoyable. Girish Jagad, your technical insights during our numerous discussions at various stages of my research journey were invaluable. Your support for communicating with University personally as and when required for submitting various documents on behalf of me deserve special mention. Thank you all the friends from Bhavnagar for providing much-needed motivation when stress levels were high.

Pathak Yashodhar Pravinkumar

24th July, 2025

Bhavnagar, Gujarat

TABLE OF CONTENTS

Machine Learning & GIS Based Landslide Susceptibility Modelling for the Part of Himachal Pradesh, India	i
DECLARATION.....	i
CERTIFICATE.....	ii
Course Work Completion Certificate	iii
Originality Report Certificate	iv
Plagiarism Report	v
PhD Thesis Non-Exclusive License to	vii
Thesis Approval Form	ix
Abstract.....	x
Acknowledgment.....	xii
TABLE OF CONTENTS	xii
LIST OF ABBREVIATIONS	xv
LIST OF FIGURES	xix
LIST OF TABLES.....	xxi
LIST OF APPENDICES	xxii
CHAPTER 1	1
Introduction	1
1.1 Background.....	1
1.2 Himalayan Landslide Crisis.....	3
1.3 Landslide Susceptibility Mapping Techniques and Limitations	6
1.4 Research Aim	8
1.5 Scope of Work.....	8
1.6 Organization of the Thesis.....	9
Chapter 2.....	11
Literature Review	11
2.1 Landslide Causative Factors	11
2.2 Machine Learning Models in LSM	14
2.3 Evaluation Metrics for Machine Learning Models.....	32
2.4 LSM approaches and Regional Application.....	34
2.5 Key Findings and Research Gaps.....	37
2.6 Problem Statement and Objectives.....	40
2.7 Original Contribution of Thesis.....	41
Chapter 3.....	43

Modelling Approach and LSM Generation	43
3.1 Methodology.....	43
3.2 Study Area	45
3.3 Data Acquisition and Preparation.....	47
3.4 Data Statistics and Visualization	59
3.4.1 Multivariate analysis of Landslide Occurrence Based on Numerical Variables.....	73
3.4.2 Elevation and Landcover Influence on Landslide Occurrence.....	75
3.5 Feature Selection	77
3.6 GIS Based Landslide Susceptibility Modelling with Frequency Ratio Method	88
3.7 GIS Based LSM with Analytical Hierarchy Process.....	93
3.8 Analysis using Standalone Machine Learning Models	97
3.8.1 Model Performance and Statistical Evaluation	98
3.8.2 Landslide Susceptibility Maps	101
3.9 Ensemble Machine Learning Models	105
3.10 Artificial Neural Network Approach for LSM	115
Chapter 4.....	119
Discussion on Model Outputs.....	119
4.1 Performance Analysis for GIS based Model: AHP and FR	119
4.2 Performance Evaluation and Statistical Validation of Machine Learning Models	120
4.3 Comparative Analysis of LSM in Uttarakhand Region.....	125
4.4 Performance Evaluation of Artificial Neural Networks (ANN) in Landslide Susceptibility Modelling.....	126
4.5 Landslide Density Index and Percentage Pixel Distribution	128
4.6 Zonal Distribution of Landslide Susceptibility	132
Chapter 5.....	136
Conclusion.....	136
5.1 Objectives Achieved.....	136
Table 5.1 : Summary of achieved Objectives.....	136
5.2 Concluding Remarks	137
5.3 Further Research Potential.....	138
References	140
APPENDIX A : Results of GIS based Landslide susceptibility models (FR and AHP).....	154
AHP Analysis	160
APPENDIX B : Results of various Machine Learning Models	165
Landslide Density Index calculation for various ML models	166
District wise Class Distribution.....	171
Combined Pixel Percentage for all Districts for zonal statistics	175

APPENDIX C : Publications.....	182
--------------------------------	-----

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BRT	Boosted Regression Trees
CAE	Conditional Attribute Evaluation
CART	Classification and Regression Tree
CCF	Canonical Correlation Forest
CDT	Credal Decision Tree
CF	Certainty Factor
CFS	Correlation-based Feature Selection
CGD	Conjugate Gradient Descent
CHAID	Chi-squared Automatic Interaction Detection
CI	Confidence Interval
CNN	Convolutional Neural Network
CR	Classification Rate
CRU	Climatic Research Unit
CSAE	Cluster and Selection Attribute Evaluation
DA	Discriminant Analysis
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
EBF	Evidence Belief Function
ENVI	Environment for Visualizing Images
EOSDIS	Earth Observing System Data and Information System
ERDAS	Earth Resource Data Analysis System

ETA	Learning Rate (in XG-Boost Hyperparameter Tuning)
FE	Feature Engineering
FIS	Fuzzy Inference System
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FR	Frequency Ratio
FRA	Frequency Response Approach
FRT	Fractal Theory
GB	Gradient Boosting
GBDT	Gradient Boosted Decision Trees
GBM	Gradient Boosting Machine
GCM	Global Climate Model
GIS	Geographic Information System
GRASS	Geographic Resources/Research Analysis Support System
GSI	Geological Survey of India
GVIF	Generalized Variance Inflation Factor
GWR	Geographically Weighted Regression
HIRESSS	High-Resolution Slope Stability Simulation
HSU	Homogenous Susceptible Units
IDW	Inverse Distance Weightage
IMD	India Meteorological Department
IVM	Iterative Voting Mechanism
KDE	Kernel Density Estimation
Km	Kilometer
KNN	K-Nearest Neighbours
LCF	Landslide Causative Factors

LD	Linear Discriminant
LDI	Landslide Density Index
LGR	Logistic Regression
LHEF	Landslide Hazard Evaluation Factor
LHZ	Landslide Hazard Zonation
LIBSVM	Library for Support Vector Machines
LMT	Logistic Model Tree
LR	Logistic Regression
LRM	Likelihood Ratio Method
LSI	Landslide Susceptibility Index
LSM	Landslide Susceptibility Mapping
LSTM	Long Short-Term Memory
LSZ	Landslide Susceptibility Zonation
m	Meter
MATLAB	Matrix Laboratory (Software)
MCC	Matthews Correlation Coefficient
MCDA	Multi-Criteria Decision Analysis
ML	Machine Learning
MLP	Multi-Layer Perceptron
mm	Millimeter
NB	Naïve Bayes
NDVI	Normalized Difference Vegetation Index
NG	Natural Gradient
NIRS	Near-Infrared Spectroscopy
NN	Neural Network
NPV	Negative Predictive Value
OA	Overall Accuracy

OOB	Out-of-Bag Error
PPV	Positive Predictive Value
PROBSTAB	Probability of Stability
PSO	Particle Swarm Optimization
QGIS	Quantum Geographic Information System
R	Correlation Coefficient
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
RS	Remote Sensing
RSSCART	Random Subspace CART
SEDAC	Socioeconomic Data and Applications Center
SHALSTAB	Shallow Landslide Stability Model
SHAP	Shapley Additive explanations
SINMAP	Stability Index Mapping
SLIP	Slope-Informed Landslide Prediction
SMART	Spatial Machine-learning Analysis for Risk and Terrain
SMO	Sequential Minimal Optimization
SOM	Self-Organizing Map
SPI	Stream Power Index
SPSS	Statistical Package for the Social Sciences
SR	Spatial Response
SRTM	Shuttle Radar Topography Mission
STARWAR	Statistical Analysis for Rainfall-induced Warnings and Responses

STATISTICA	Statistical Analysis Software
SVM	Support Vector Machine
SW	Spatial Weights
TN	True Negatives
TNR	True Negative Rate
TP	True Positives
TRIGRIS	Topographic Relief and GIS Integration for Risk Identification System
TWI	Topographic Wetness Index
VIF	Variance Inflation Factor
VIS	Visible Spectrum (in Remote Sensing)
WMS	Web Map Service
WOE	Weight of Evidence
w.r.t	With respect to
WSN	Wireless Sensor Network
XG	Extreme Gradient
XGB	Extreme Gradient Boosting

LIST OF FIGURES

Figure 1.1: Representation of Landslide Phenomena [2]	1
Figure 1.2: Landslide Classification mechanism (A) Rotational Landslide, (B) Translational Landslide, (C) Block Slide, (D) Rockfall, (E) Topple, (F) Debris Flow, (G) Debris Avalanche, (H) Earthflow, (I) Creep, (J) Lateral Spread	3
Figure 1.3: Global Landslide hazard distribution	4
Figure 1.4: Organization of Thesis	10
Figure 2.1: Most used causative Factors in period (2005-2012 and 2013 -2016).....	14
Figure 2.2: Trends in LSM in period (2005-2012 and 2013-2016).....	15
Figure 2.3: Example figure for Decision Tree Model	16
Figure 2.4: K-NN workflow	21
Figure 2.5: Flowchart of the study.....	23
Figure 2.6: Landslide Studies Distribution in India	35
Figure 2.7: Study Area in Garhwal Himalaya and its Geology	36
Figure 2.8: (a) Number of Published papers in Scopus and Google Scholar on LSM, (b) Number of Papers published for LSM in Himalayan region, (c) Number of published papers with different models.....	40
Figure 3.1: Workflow for data collection, feature extraction, and dataset preparation	43
Figure 3.2: Methodological Framework for Machine Learning Modelling	44
Figure 3.3: Study Area.....	46
Figure 3.4: SOI Toposheets details of Study Area	47
Figure 3.5: Landslide Locations in the Study Area	50
Figure 3.6: Preparation of Landslide Inventory.....	50
Figure 3.7: TIN Representation for Study Area	52
Figure 3.8: Topographical Features: (a) Aspect, (b) Slope, (c) Curvature, (d) Elevation(m), (e) TWI and (f) SPI	53
Figure 3.9: Drainage Map for the Study Area	54
Figure 3.10: Hydrological Features: (a) Distance to Stream, (b) Rainfall	54
Figure 3.11: Satellite Image of the Study Area	55
Figure 3.12: Geological Features: (a) NDVI, (b) Soil, (c) Geomorphology, (d) Geology, (e) Distance to Lineaments	56
Figure 3.13: Validation of Area of Interest (AOI) for Supervised Classification	57
Figure 3.14: Landcover Map	58
Figure 3.15: Anthropogenic Feature: Distance to Road.....	58
Figure 3.16: Dependent variables Statistics	59
Figure 3.17: General Statistics for Variables.....	60
Figure 3.18: Density plot, Violin plot and Boxplot for Curvature with Landslides perspective	61
Figure 3.19: Density plot, Violin plot and Boxplot Elevation(m) with Landslides Perspective	62
Figure 3.20: Density plot, Violin plot and Boxplot for Slope(Degree) with Landslides Perspective.....	63
Figure 3.21: Density plot, Violin plot and Boxplot for Rainfall(mm/year) with Landslides Perspective.....	64
Figure 3.22: Density plot, Violin plot and Boxplot for TWI with Landslides Perspective.....	64
Figure 3.23: Density plot, Violin plot and Boxplot for SPI with Landslides Perspective.....	65

Figure 3.24: Density plot, Violin plot and Boxplot for NDVI with Landslides Perspective ..	65
Figure 3.25: Visualization of Landcover Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams.....	66
Figure 3.26: Visualization of Geological Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams.....	67
Figure 3.27: Visualization of Geomorphological Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams	68
Figure 3.28: Visualization of Soil Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams	69
Figure 3.29: Aspect Distribution w.r.t Landslide occurrence.....	70
Figure 3.30: Distance to Lineaments(m) Distribution w.r.t Landslide occurrence	71
Figure 3.31: Distance to Road(m) Distribution w.r.t Landslide occurrence	72
Figure 3.32: Distance to Streams(m) Distribution w.r.t Landslide occurrence	73
Figure 3.33: Landslide Distribution w.r.t to Elevation(m), Rainfall(mm/year) and Slope(Degree).....	74
Figure 3.34: Landslide Distribution w.r.t Elevation(m), NDVI, Rainfall(mm/year) and Slope(Degree).....	74
Figure 3.35: Landslide occurrence w.r.t Elevation(m) disaggregated by Landcover.....	76
Figure 3.36: Landslide Occurrence w.r.t to Elevation(m) and Landcover	76
Figure 3.37: GVIF of the Features.....	79
Figure 3.38: Variance of the Features	79
Figure 3.39: Spearman's Rho Correlation analysis	81
Figure 3.40: Kendall's Tau Correlation analysis.....	81
Figure 3.41: Chi-square score of the Features	83
Figure 3.42: Mutual Information Score of the Features	84
Figure 3.43: Gini Importance Plot.....	86
Figure 3.44: Mean Decrease Accuracy Plot	87
Figure 3.45: Permutation Feature Importance Plot	87
Figure 3.46: Extra Tree Classifier Importance Plot.....	87
Figure 3.47: Methodological framework for LSM using FR model	89
Figure 3.48: Prediction Rate for Variables.....	91
Figure 3.49: LSM using FR Approach	92
Figure 3.50: ROC for FR Approach	92
Figure 3.51: Methodological framework for LSM using AHP	93
Figure 3.52: Determination of Criteria Weights.....	94
Figure 3.53: LSM using AHP Approach.....	96
Figure 3.54: ROC Curve for AHP	96
Figure 3.55: Machine Learning Modelling Process	97
Figure 3.56: K-Fold Cross Validation	99
Figure 3.57: Statistical evaluation for single ML Models on Testing Datasets	100
Figure 3.58: ROC Plot for Standalone ML Models.....	101
Figure 3.59: Landslide Susceptibility Maps: (a) LR, (b) KNN, (c) SVM, (d) KSVM, (e) NB, (f) DT.....	105
Figure 3.60: Random Forest Trees Visualization.....	106
Figure 3.61: Predictor Optimization for Random Forest.....	108
Figure 3.62: ROC and AUC for Random Forest.....	109
Figure 3.63: Hyperparameter tuning (XG-Boost)	110
Figure 3.64: ROC Plot for XG Boost	111
Figure 3.65: Landslide Susceptibility Maps (a)Random forest and (b) XG-Boost.....	113
Figure 3.66: Comparison for different ML Models.....	114

Figure 3.67: Graphical Representation for the Artificial Neural Network Approach	116
Figure 3.68: Training Accuracy vs. ANN Epoch	117
Figure 3.69: Training Loss vs. ANN Epochs	117
Figure 3.70: ROC Plot for ANN.....	118
Figure 3.71: Landslide susceptibility map generated with ANN	118
Figure 4.1: Percentage of Landslide Pixels and Class Pixels for AHP and FR.....	120
Figure 4.2: Statistical evaluation of the confusion matrix parameter for ML Models	123
Figure 4.3: Chart for Landslide Density Index for Different Models	130
Figure 4.4: Percentage Pixel distribution for Landslide Susceptibility Models.....	130
Figure 4.5: Area % and Landslide % Distribution in Study Area	132
Figure 4.6: High Risk Areas with respect to Districts and Models	132
Figure 4.7: Landslide Susceptibility distribution by District along with risk index.....	133
Figure 4.8: Radar Chart showing Risk Index, Landslide/Area Ratio, %Landslide and %Area	134

LIST OF TABLES

Table 1.1 : Classification of landslide based on their movement mechanism and material.....	2
Table 1.2 : Recent Landslide events in Himalayan Region.....	5
Table 1.3 : Comparison between GIS-Based Methods and ML Approaches.....	7
Table 2.1 : Merits and Demerits of ML Models for LSM.....	32
Table 2.2 : LSM Approaches in Indian Himalayan Region	33
Table 2.3 : LSM Approaches in Indian Himalayan Region	36
Table 3.1 : Data Acquisition and Preparation.....	48
Table 3.2 : Filter based feature selection method.....	82
Table 3.3 : Prediction Rate for all Variables	90
Table 3.4 : Fundamental Scale for Pairwise Comparison	95
Table 3.5 : Confusion Matrix Results.....	98
Table 3.6 : Statistical Evaluation results for Single ML Models on Testing Datasets	100
Table 3.7 : Grid Search Results for Hyperparameter Tuning.....	107
Table 3.8 : Performance comparison for different Bagging Models.....	108
Table 3.9 : Selected hyperparameters for the XGBoost model	109
Table 3.10 : Performance comparison for Boosting Models.....	111
Table 3.11 : Performance Comparison for Voting, Stacking, Boosting and Bagging algorithms	112
Table 3.12 : Performance comparison for standalone and Ensemble ML Models.....	114
Table 3.13 : Criteria Details for ANN Modelling	115
Table 4.1 : Statistical Evaluation Parameters of ML Models.....	122
Table 4.2 : Landslide Density Index for different models.....	129
Table 4.3 : Percentage Pixel Distribution for Landslide Susceptibility Models	130
Table 4.4 : Zone wise landslide-area ratio and risk index	133
Table 5.1 : Summary of achieved Objectives.....	136

LIST OF APPENDICES

APPENDIX A : Results for GIS Based Landslide Susceptibility Models (FR and AHP).....	153
APPENDIX B : Results for Single and Ensemble Machine Learning Models.....	164
APPENDIX C : Publications.....	181

CHAPTER 1

Introduction

1.1 Background

Landslides are defined as the downward movement of rock, soil, and associated debris along a slope, primarily driven by gravitational forces. Recognized as one of the most destructive natural hazards, landslides can cause severe damage to infrastructure and result in significant casualties in affected regions. As illustrated in Figure 1.1, landslides are complex phenomena shaped by a wide range of natural and anthropogenic factors. Natural contributors include geological structure, geomorphological features, hydrological conditions, and meteorological triggers such as intense rainfall or seismic events [1-3]. At the same time, human-induced activities such as deforestation, rapid urbanization, slope cutting, and road construction on unstable terrain further aggravate slope instability [2,4,5]. Additionally, long-term climate change has been increasingly identified as a contributing factor, particularly through shifts in rainfall intensity and frequency.

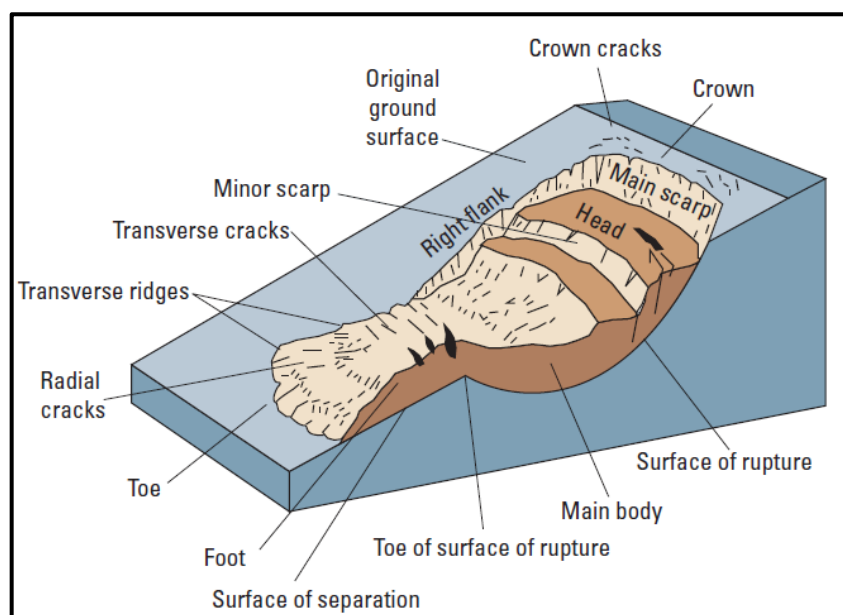


Figure 1.1: Representation of Landslide Phenomena [2]

Landslides can be systematically classified based on the type of movement involved and the nature of the slope-forming material. The major types of movement include falls, topples, slides (rotational and translational), lateral spreads, and flows, each influenced by a combination of geological, topographic, hydrological, and anthropogenic factors [2]. These movement types interact differently with materials such as bedrock, predominantly fine-grained soils, and predominantly coarse-grained soils, resulting in distinct forms of mass wasting. A detailed classification scheme is presented in **Table 1.1**, which outlines the specific movement-material combinations such as rockfalls, debris flows, earth slumps, and lateral spreads. Figure 1.2 shows the visualization for different types of movement in landslides.

Table 1.1 : Classification of landslide based on their movement mechanism and material [6]

Type of Movement	Bedrock	Fine Grain Soil	Coarse Grain Soil
Falls	Rockfall	Earth fall	Debris fall
Topples	Rock topples	Earth topples	Debris topple
Slides (Rotational)	Rock slump	Earth slump	Debris slump
Slides (Translational)	Rock block slide	Earth block slide	Debris block slide
	Rock slide	Earth slide	Debris slide
Lateral Spreads	Rock spread	Earth spread	Debris spread
Flows	Rock flow	Earth flow	Debris flow
	Rock avalanche	-	Debris avalanche
	Deep creep	Soil creep	

The categorization shown above in Table 1.1 provides a comprehensive framework for analysing the mechanics of slope failure and is vital for planning effective hazard assessment, risk mitigation, and land-use management strategies. Given the complex and multifactorial nature of landslides, this classification highlights the importance of a multidisciplinary approach in susceptibility mapping, incorporating both natural and human-induced variables.

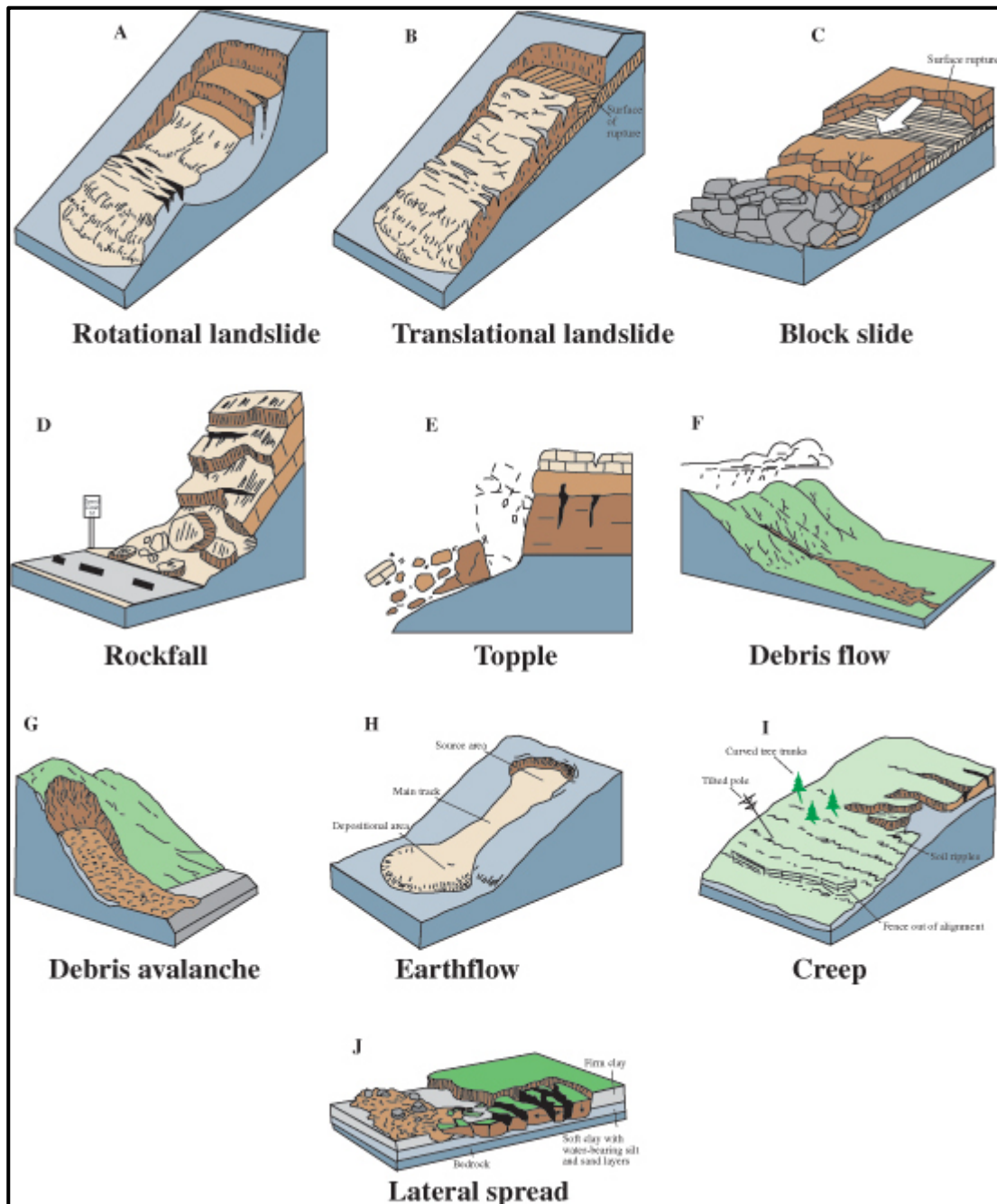


Figure 1.2: Landslide Classification mechanism (A) Rotational Landslide, (B) Translational Landslide, (C) Block Slide, (D) Rockfall, (E) Topple, (F) Debris Flow, (G) Debris Avalanche, (H) Earthflow, (I) Creep, (J) Lateral Spread [6]

1.2 Himalayan Landslide Crisis

Managing the aftermath of landslides can be challenging, but implementing effective preparedness and mitigation strategies can significantly reduce fatalities. Developing

such plans requires precise identification of landslide-prone zones, which is facilitated through Landslide Susceptibility Mapping (LSM). LSM provides policymakers with crucial insights into areas at risk of landslides. Numerous studies have explored various approaches to LSM. It involves spatial prediction of landslides and considers the interplay between landslides and related internal factors such as geotechnical properties, geological conditions, hydrological influences, vegetation cover, and topographic attributes. Formally, LSM refers to the division of land into nearly homogeneous zones ranked by their potential and actual hazard due to landslides [7].

Landslides represent one of the most significant natural hazards globally, responsible for substantial human and economic losses. Figure 1.3 provides a global perspective on landslide hazard distribution, illustrating spatial variations in susceptibility levels across different regions. Image Areas shaded in purple represent zones of low landslide risk, whereas regions marked in orange and yellow signify areas with moderate to high susceptibility. Prominent high-risk zones include the Himalayan arc, the Andes Mountains, the western coastal ranges of North America, and several other mountainous and tectonically active regions across the globe. The map shows the strong correlation between landslide hazards and steep topography, high rainfall, and seismic activity, offering valuable insight into the spatial dynamics of landslide-prone areas at a global scale.

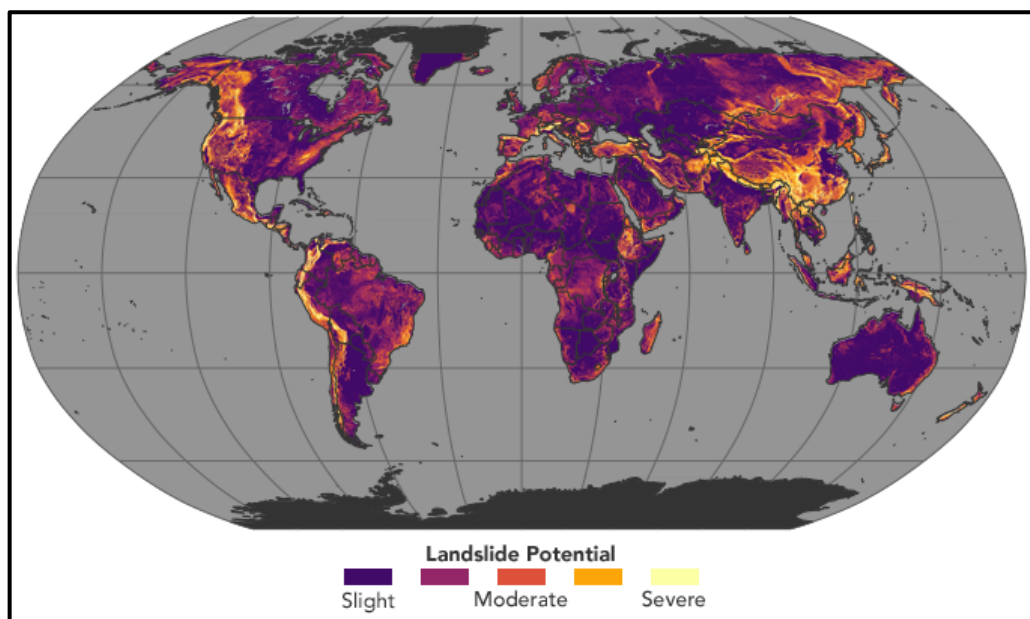


Figure 1.3: Global Landslide hazard distribution [8]

According to the Centre for Research on the Epidemiology of Disasters, landslides account for approximately 17% of fatalities associated with major natural disasters worldwide [9]. Between 2004 and 2010, there were 2,620 fatal landslide events reported globally, resulting in the deaths of over 32,000 people [10]. The Asian continent has experienced a disproportionately high impact, with an estimated 18,000 deaths and around 5.5 million people affected by landslides between 1950 and 2009, making it one of the most vulnerable regions in the world [11]. Below Table 1.2 shows some recent landslide events with its triggering mechanism in Himalayan region.

Table 1.2 : Recent Landslide events in Himalayan Region

Region/State	Event & Year	Trigger	Impacts	Source
Uttarakhand	Tapovan Landslide, Chamoli (2021)	Rock-ice avalanche	Flash flood, destruction of hydropower stations, significant casualties	[12]
	Kedarnath Disaster (2013)	Intense rainfall	Massive landslides, flash floods, loss of life, infrastructure damage	
Himachal Pradesh	Monsoon-triggered Landslides (2023)	Intense July rainfall	Road blockages, travel disruption, community displacement	[13]
Sikkim	Landslides during Monsoon (June 2024)	Heavy rainfall	Fatalities, thousands of tourists stranded	[14]
Kashmir Himalaya	Post-Earthquake Landslides (2005)	Aftershocks from 2005 earthquake	Thousands of landslides, widespread slope failure	[15]

In recent decades, factors such as unregulated urban growth, climate-induced increases in regional precipitation, and widespread deforestation have significantly heightened the risk and frequency of landslides in many parts of the world [5,16]. Within this broader context, the Himalayan region owing to its complex geology, steep terrain, and intense monsoonal activity has emerged as a critical hotspot. One of the most striking examples comes from the Darjeeling Himalaya, a sub-region that has historically experienced high landslide activity. A notable event occurred in 1968, when an exceptionally intense rainfall event triggered widespread slope failures, resulting in the deaths of nearly 1,000 individuals.

1.3 Landslide Susceptibility Mapping Techniques and Limitations

Landslide Susceptibility Mapping (LSM) has traditionally relied on a range of GIS-based and statistical approaches that utilize either heuristic knowledge or data-driven frameworks to evaluate slope instability. One widely used GIS-based technique is the preparation of thematic layers derived from expert judgment and historical event mapping. A fundamental component of this is the landslide inventory method, which documents key attributes such as landslide location, timing, and morphology, often using point or polygon representations [17, 18]. These inventories help understand both individual and cumulative failure events across a landscape [19]. Studies comparing inventories, such as one utilizing frequency–area statistics, revealed that landslide frequency initially increases with area before declining, following a power-law distribution, thereby emphasizing the need for complete and high-resolution datasets [20]. Colombo et al. [21] demonstrated how aerial photographs and GIS databases can be employed to create detailed landslide inventories, though such efforts are often resource-intensive. Beyond inventories, other GIS-based heuristic models include the map combination method and geomorphic assessment techniques. Complementing these are statistical approaches, which are data-driven and aim to determine the likelihood of landslide occurrences by correlating causative factors with historical landslide distributions [22, 23, 24–27]. For instance, the Weight of Evidence (WOE) model, applied in regions like Rudraprayag district of the Garhwal Himalaya [28], assigns numerical weights to causative factors based on landslide frequency and integrates these layers using GIS overlays—a method also adopted in several other studies [29–31]. Similarly, the Frequency Ratio (FR) approach calculates the ratio of landslide presence in various subclasses of contributing factors to estimate their relative influence, producing a cumulative susceptibility index [32]. For more complex terrain modelling, multivariate statistical methods evaluate the combined effects of multiple parameters, often requiring external statistical software in conjunction with GIS tools. In addition, probabilistic methods offer a framework for analysing landslide occurrence by estimating spatial, temporal, and magnitude-based probabilities [19]. Among these, the Analytical Hierarchy Process (AHP) stands out for its semi-quantitative nature, involving problem definition, generation of a pairwise comparison matrix, and calculation of priority weights and consistency ratios. This method assigns values ranging from 1 to 9 based on the perceived influence of variables on landslide initiation, and has been effectively

applied in GIS-based LSM across various global regions [33–35]. Supporting all these methods is the integration of Remote Sensing and GIS, which enables efficient spatial data acquisition and terrain analysis. Advances in earth observation technologies have substantially enhanced the ability to detect, monitor, and map landslides and related hazards [36]. While these GIS-integrated approaches have formed the foundation of many LSM efforts, but they pose limitations due to their reliance on static assumptions and limited capabilities in capturing complex interactions.

In contrast, Modern techniques which utilize data driven Machine Learning (ML) approaches to uncover hidden patterns in historical landslide data. Models like Random Forest, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs) are capable of handling nonlinear interactions between multiple variables without relying on predefined assumptions. These methods learn directly from raw or pre-processed data and optimize their parameters through iterative training processes, significantly enhancing predictive accuracy and generalization. Comparison between GIS based methods and Machine learning approaches is shown in Table 1.3.

Table 1.3 : Comparison between GIS-Based Methods and ML Approaches

Aspect	GIS Overlay Methods	Machine Learning Approaches
Interpretability	High based on transparent, expert-defined weights	Lower often perceived as black-box models requiring post-hoc analysis
Data Dependency	Can function with limited datasets	Requires large, high-quality datasets for optimal training
Flexibility	Limited adaptability to new data or changing conditions	Highly adaptable – models retrain and update with new data
Computational Demand	Low simple overlay computations	High requires greater computational resources
Accuracy	Moderate may be influenced by subjective weighting	High leverages statistical optimization for enhanced predictions

From the comparison of GIS-based models and ML models it is observed that GIS methods can only be useful to provides rapid assessments in regions with limited data availability. Whereas, for planning mitigation strategy which requires higher prediction accuracy and understanding of complex non-linear relationship between different causative factors, ML models are preferred.

1.4 Research Aim

The field of landslide susceptibility mapping (LSM) has seen a significant evolution, transitioning from conventional qualitative assessments to more advanced, data-driven methodologies that incorporate machine learning (ML) with geographic information systems (GIS). These developments have notably improved the accuracy and utility of susceptibility models, yet critical gaps remain—particularly in selecting optimal input variables, determining appropriate ML algorithms, and validating outputs across diverse physical environments. To bridge these gaps, this research is centered on creating a unified framework that combines ML techniques with geospatial analysis tools to enable dynamic and responsive landslide susceptibility assessments. The proposed framework is designed to accommodate the complex nature of terrain, environmental variability, and anthropogenic influences, offering a more flexible and robust tool for identifying areas at risk. By leveraging the predictive power of ML and the spatial intelligence of GIS, the study aims to produce detailed, scalable, and decision-supportive susceptibility maps that can effectively inform land-use planning, hazard zoning, and disaster management strategies in landslide-prone regions.

1.5 Scope of Work

The scope of this study encompasses the development of a machine learning–based landslide susceptibility mapping (LSM) framework for selected regions within the Lower Sutlej River Basin, located in the northwestern Himalayas. Specifically, the analysis is focused on four districts: Kinnaur, Shimla, and selected portions of Kullu and Manali. These areas have been identified based on their history of frequent landslide activity, complex terrain, and increasing exposure to anthropogenic influences such as road construction and land-use change.

This research focuses on landslide susceptibility modelling by incorporating landslides triggered by both historical rainfall events and seismic activities, ensuring a comprehensive assessment of multiple triggering factors. The research involves the integration of geospatial techniques and machine learning algorithms to analyse landslide susceptibility patterns across varying geomorphic, geological, and climatic settings within the defined region. While a more detailed characterization of the study area, including topographic and administrative boundaries, will be presented in the methodology chapter, this section outlines the overall spatial and thematic extent of the work. The findings from this multi-district LSM exercise aim to contribute to regional-scale disaster preparedness, land-use regulation, and the formulation of targeted mitigation strategies for landslide-prone zones in the Himalayas.

1.6 Organization of the Thesis

The organization of this thesis as shown in Figure 1. is structured to provide a coherent and systematic flow of information, guiding the reader through the research process and its findings. The study begins with an introduction to the topic, highlighting the importance of landslide susceptibility mapping (LSM) and its evolution over the years. Following this, a detailed review of the existing literature sets the foundation by examining conventional methods, modern trends, and identified research gaps. The methodology chapter outlines the development of machine learning models integrated with GIS for landslide mapping, including data collection, factor selection, and validation techniques. The results and discussion chapter presents the outcomes of the study, evaluating model performance and their implications. Finally, the conclusion chapter summarizes the key findings, contributions, and suggestions for future research, emphasizing the practical applications of this study in landslide mitigation and planning.

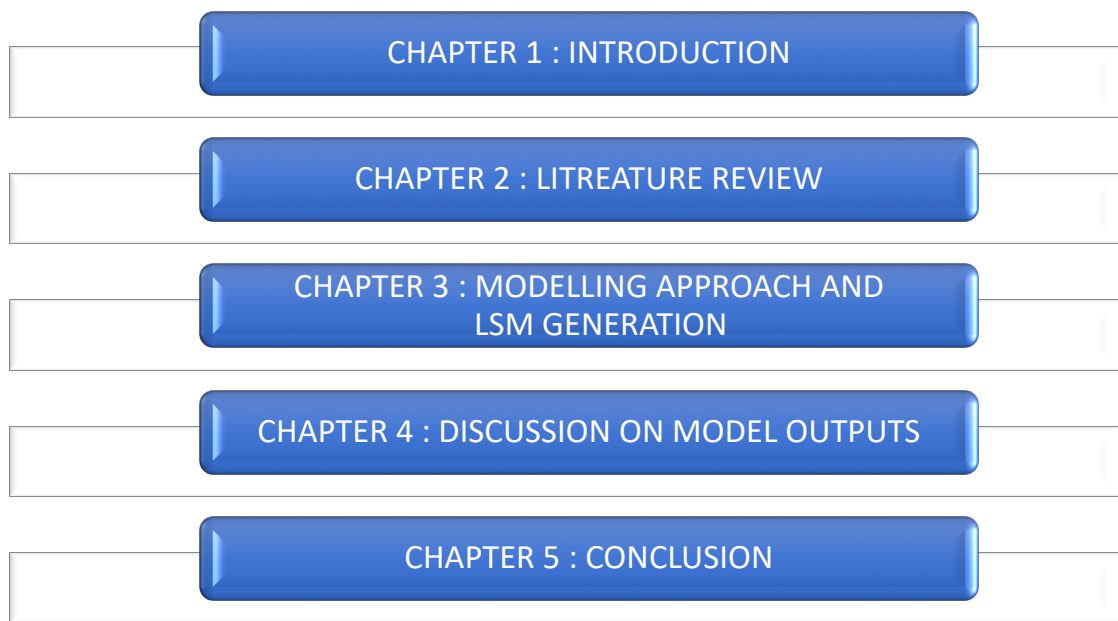


Figure 1.4: Organization of Thesis

CHAPTER 2

Literature Review

2.1 Landslide Causative Factors

Landslides are complex geomorphic phenomena resulting from the downslope movement of soil, rock, and organic matter, typically triggered by a combination of natural and anthropogenic factors. Their occurrence is influenced by a variety of geo-environmental conditions, including terrain morphology, geological structure, hydrology, land use practices, and external triggers such as rainfall and seismic activity. The extent to which each factor contributes to slope instability varies across different regions, depending on local topography, meteorological dynamics, and human interventions. There is no specific instruction available for selection of causative factors.

Geomorphological Factors

Key geomorphological parameters influencing landslide occurrence include slope angle, aspect, elevation, and curvature. These topographic features, often derived from Digital Elevation Models (DEMs), serve as foundational inputs for landslide susceptibility mapping [37,38]. Among these, slope angle is widely recognized as a primary determinant of slope failure. As slope steepness increases, the gravitational force acting on soil and rock materials tends to overcome shear strength, reducing slope stability [39]. While landslides commonly occur in moderate to steep slopes (typically between 15° and 45°) [40,41], failures have also been observed on gentle slopes, particularly during debris flows or rapid mass movement events [2]. Nevertheless, not all steep slopes are prone to failure—those composed of competent rock can remain stable, while comparatively gentle slopes with weaker lithology may be more susceptible [4].

Aspect, defined as the orientation of the slope, indirectly influences landslide risk by altering surface moisture regimes, weathering processes, and vegetation growth through differential exposure to solar radiation and precipitation [3, 43]. It is typically classified into eight directional categories and a flat class [44, 45]. Elevation, another critical factor, affects weathering intensity and moisture content. Rocks at higher elevations generally

undergo slower weathering, often retaining greater mechanical strength than those at lower altitudes [46, 47].

Curvature both plan and profile is also relevant as it influences surface water flow and consequently, terrain moisture. Plan curvature, the curvature perpendicular to the slope direction, affects lateral flow convergence or divergence, while profile curvature, measured along the slope direction, influences flow acceleration or deceleration [48]. Convex surfaces tend to accumulate less moisture and are often more failure-prone compared to concave areas [49].

Geological Factors

Geological characteristics, including lithology, structural discontinuities, and soil types, play a vital role in slope stability [50]. Lithology determines the physical and mechanical properties of slope materials. Changes in lithology and rock structure can significantly influence strength, permeability, and overall slope behavior [51]. Rocks with similar composition are usually grouped into distinct lithological formations to analyse susceptibility patterns [52,53].

Lineaments representing faults, fractures, and other tectonic features serve as indicators of structural weaknesses in the terrain. The proximity and density of lineaments are frequently used in susceptibility analysis, given their role in slope destabilization [46]. Similarly, soil type governs landslide risk by determining shear strength, permeability, and overall material behaviour. Classification is typically based on mechanical and chemical properties, and includes assessments of mineralogical and structural features [3, 54].

Land Cover and Land Use Factors

Vegetation cover is known to enhance slope stability by reinforcing soil structure through root systems and reducing surface moisture via evapotranspiration [4, 55]. Forested regions generally experience fewer landslides compared to barren or sparsely vegetated areas. Additionally, vegetation serves to mitigate the erosive impact of rainfall and other climatic forces on the slope surface.

Hydrological and Drainage Parameters

Drainage patterns influence subsurface moisture and pore water pressure, making them important considerations in landslide prediction. Areas close to rivers or drainage lines

are particularly prone to saturation, often resulting in slope failures [46]. Metrics such as distance to rivers and stream density are commonly incorporated into hazard assessments [56, 57].

Triggering and External Factors

Among the external triggers, rainfall is the most frequent and critical factor in many regions. Prolonged or intense rainfall events raise groundwater levels and pore pressure, reducing effective stress and triggering slope failures [4]. Four types of rainfall metrics are typically used in such analyses: short-term intensity, storm duration, total rainfall, and antecedent precipitation [58]. While spatial analyses often use total annual rainfall [59], temporal prediction models benefit from short-term or antecedent rainfall data [60].

Earthquakes also play a significant role in initiating landslides by inducing ground shaking, which disturbs slope equilibrium and may lead to liquefaction or mass movement [61]. These earthquake-induced failures are often large in scale and rapid in onset, typically manifesting as slides or rock falls [3].

Anthropogenic activities such as slope cutting for roads, overloading of slopes, irrigation, or poor drainage practices can disrupt the natural balance of slopes, significantly increasing landslide risks [2]. Road construction alters slope geometry, reduces stability, and introduces artificial drainage, all of which can accelerate erosion and slope failure [62].

Although, there are no specific set of instructions to select the causative factors for landslide susceptibility mapping, in their study [63] has reviewed most used causative for landslide susceptibility mapping within the period of 2005 to 2016. As can be seen from Figure 2.1, Slope, lithology, aspect, and land use appear as the most frequently utilized factors in both timeframes, showing a consistent trend. There is a noticeable increase in the number of studies incorporating factors like precipitation, topographic wetness index, and stream power index in the later period, indicating a growing interest in hydrological and vegetation-related parameters. Factors such as drainage density and curvature were the least applied across both periods, reflecting their relatively limited use in landslide susceptibility modelling.

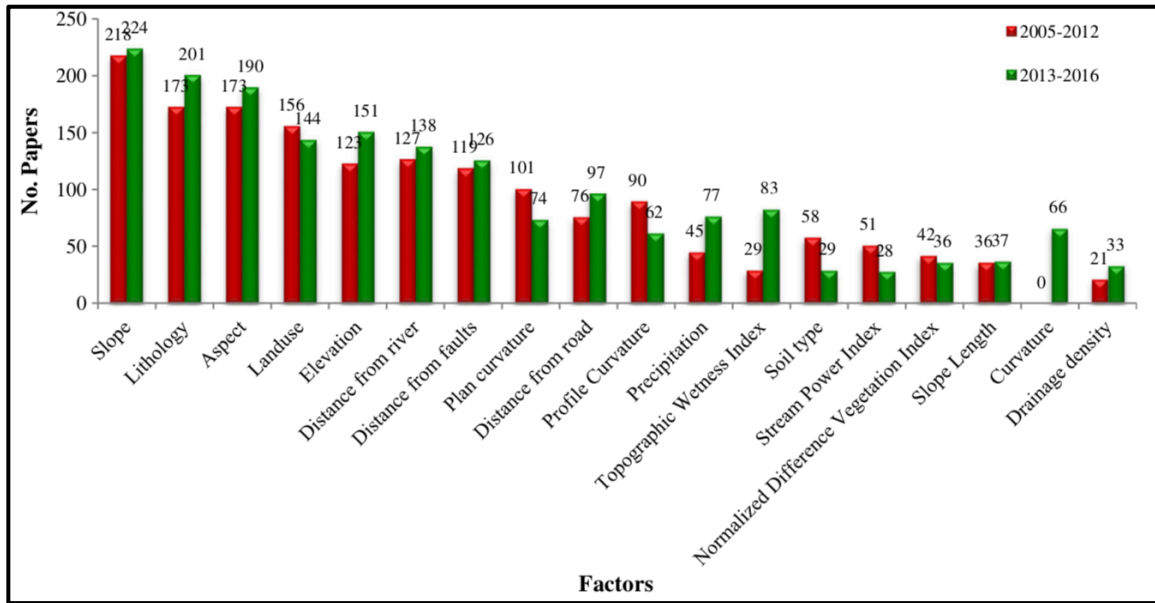


Figure 2.1: Most used causative Factors in period (2005-2012 and 2013 -2016) [63]

2.2 Machine Learning Models in LSM

The trend in landslide susceptibility modelling has shown a gradual shift from traditional methods such as Frequency Ratio (FR), and Analytical Hierarchy Process (AHP) towards more advanced machine learning (ML) models and ensemble techniques. As depicted in Figure 2.2, LR and FR remained dominant between 2005–2012, but a notable rise in the use of ML models like Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF) is evident in the period from 2013–2016. Particularly significant is the increased adoption of combined models (ensemble approaches), which have seen a marked growth in recent years. This reflects a growing preference for data-driven, adaptive methodologies that offer improved predictive accuracy and generalization over conventional techniques. The figure also reveals that Random Forest and ensemble models, though emerging more recently, are gaining traction, underlining their importance in contemporary LSM research. This shift supports the increasing emphasis on ensemble ML models for enhanced landslide hazard prediction and mapping. Popular ML models used in LSM are discussed below in detail.

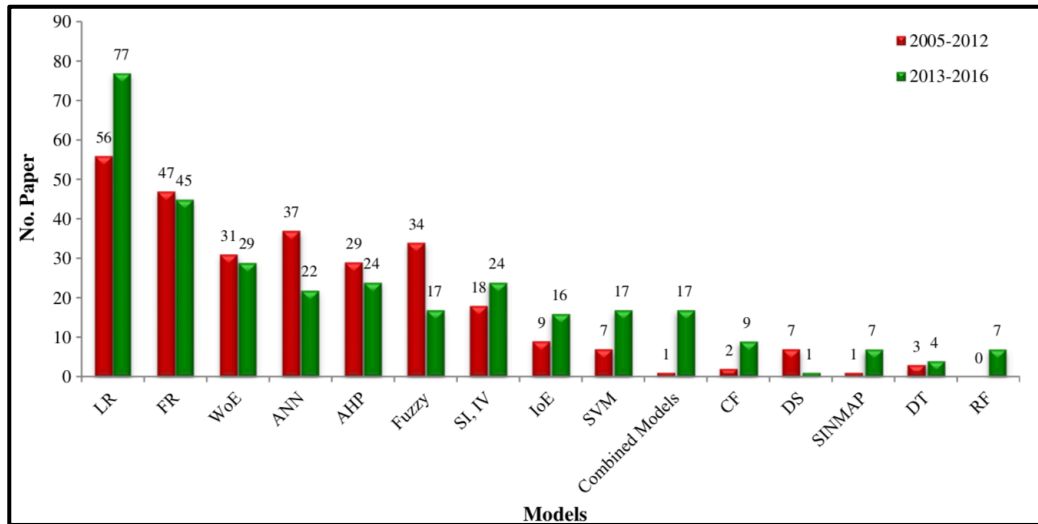


Figure 2.2: Trends in LSM in period (2005-2012 and 2013-2016) [63]

Decision Tree and Random Forest

Decision Tree (DT) models represent a type of supervised learning algorithm that is both non-parametric and non-linear in nature. These models function by partitioning the input data into subsets based on the values of independent variables. They are versatile in handling both categorical and continuous types of input and output data. Decision Trees operate using a rule-based structure that generates a binary tree through a process known as binary recursive partitioning (BRP). This method systematically divides each decision node into branches guided by simple logical rules (True/False or 1/0). As a result, the internal nodes derive decision outcomes by evaluating predictor variables from the input dataset. The foundation of many Decision Tree algorithms lies in the ID3 algorithm, introduced by [64], which follows a top-down approach for building the tree structure as shown in Figure 2.3.

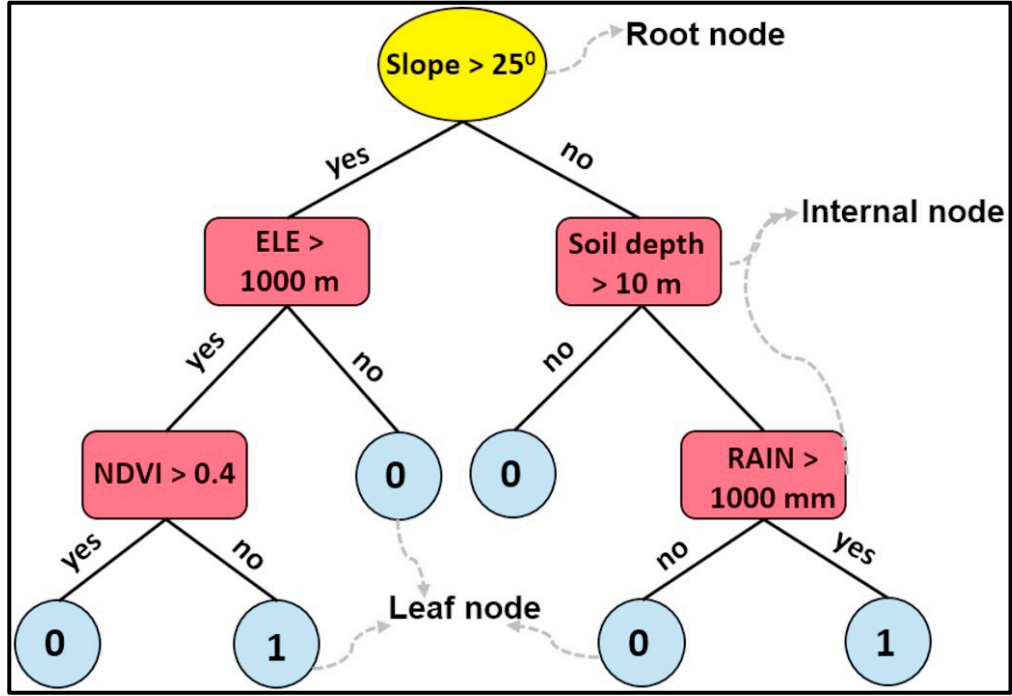


Figure 2.3: Example figure for Decision Tree Model [64]

The Random Forest (RF) algorithm operates by constructing multiple decision trees using input data and predicting outcomes through voting (for classification tasks) or averaging (for regression tasks) [65,66]. Essentially, it is an ensemble of decision trees (DT). Increasing the number of trees reduces the risk of overfitting but adds to computational complexity. In classification tasks, each decision tree predicts a class, and the class receiving most votes is selected as the final prediction. For regression tasks, the final output is determined by averaging the predictions of all trees.

The performance of the RF model heavily depends on the input variables (landslide causative factors), making the selection of these variables a critical step [67]. According to the guideline provided in [69], the ideal number of observations required for input data can be approximated by the expression $\log_2(M + 1)$, where M denotes the number of input variables. Additionally, the mean squared error (e) associated with the Random Forest (RF) model and is represented using Equation (2.1) where $\vartheta_{Observed}$ and $\vartheta_{response}$ is the representation of observed data and response data.

$$\varepsilon = (\vartheta_{Observed} - \vartheta_{response})^2 \quad (2.1)$$

Average of all trees prediction can be given by Equation (2.2) where S is the prediction of forest and K is applicable to all trees.

$$S = \frac{1}{K} \sum k^{th} \vartheta_{response} \quad (2.2)$$

Several studies have demonstrated the effectiveness of this model. For instance, a study by [69] utilized the RF model for classification, employing 50 "ntrees" and 8 "mtry," which resulted in a 98.3% success rate and an AUC prediction value of 97.7%.

As highlighted by [70], there is no fixed rule for determining the optimal values of "ntrees" and "mtry" in the RF model. Another study by [71] which trained an RF model on 700 landslide pixels using the classification approach with a specific equation (2.3). These studies underscore the importance of tuning RF model parameters to achieve accurate and reliable predictions.

$$o = (i) = \max[\sum_k f(d)] \quad (2.3)$$

Where; input data is indicated by I, output computed is indicated with o and f(d) is Indicator function which is defined as;

$$f(d) = \{1, \text{if } d \text{ is 'YES'} 0, \text{ Otherwise} \quad (2.4)$$

In the equation (2.4 “YES” represents a landslide-prone zone, while “OTHERWISE” denotes a non-landslide zone. After parameter tuning, the value of k was set to 1000, and three features were evaluated at each node. The Out of Bag (OOB) method, commonly used to measure prediction error, was employed to assess the accuracy of the Random Forest (RF) model. The results indicated an accuracy of 98%.

A comparative study conducted by [72] examined the performance of traditional statistical models (Certainty Factor), conventional machine learning models like Support Vector Machine (SVM) and RF, and hybrid models such as CF-SVM and CF-RF. In this study, the RF model utilized 500 decision trees with three randomly selected features, and 464 samples of landslide and non-landslide data were used for training. The model was tested on 200 samples, of which 156 were accurately predicted, resulting in a prediction accuracy of 78%.

Another study by [73] explored the use of RF for feature selection to improve classification accuracy by eliminating irrelevant features and carefully selecting landslide causative factors. The RF model was implemented with 500 trees and four key features, achieving an AUC (Area Under the Curve) value of 0.93. This study used a dataset of 2100 landslide-related data points, dividing them equally into two groups: one for

training (1050 points) and the other for testing. The analysis was conducted using STATISTICA, and the RF method achieved an AUC score of 0.7394.

The reviewed literature highlights the growing popularity of RF as a robust method for landslide susceptibility mapping. RF's advantages, such as ease of implementation, superior ranking of causative factors, elimination of irrelevant variables, and higher prediction accuracy, make it a preferred choice for researchers in this domain.

Support Vector Machine

The Support Vector Machine (SVM) model is a supervised machine learning approach that can be utilized for both classification and regression tasks. The fundamental principles of SVM, as described by [74], are outlined as follows:

1. **Hyperplane Determination:** SVM identifies a line or a hyperplane that divides input data into distinct classes effectively.
2. **Maximum Margin Principle:** For classification tasks, the hyperplane with the largest margin (i.e., the greatest distance between the hyperplane and the nearest data points of each class) is considered optimal. The SVM model is trained to find this hyperplane.
3. **Handling Nonlinear Data:** In practical applications, datasets are often not linearly separable. For such cases, the soft margin technique is employed, allowing some misclassifications to ensure better generalization.
4. **Kernel Function:** The kernel function is a mathematical tool in SVM that transforms input data into a higher-dimensional space, enabling the separation of nonlinear datasets. This transformation is key to SVM's ability to handle complex, nonlinear relationships.

The basic classification function of the SVM model is expressed by the following equation (2.5):

$$y = \text{sgn}(f(x)) = \text{sgn}\left(\sum_{i=1}^n w_i x_i + b\right) = \text{sgn}(w \cdot x + b) \quad (2.5)$$

In the equation, the coordinates are denoted by n , where x_i represents the parameters in the vector x in the original space. The class label is given by y , and the hyperplane parameters are defined by the weight w and the bias b . The function sgn represents the sign function. For datasets that are non-linear or noisy, a soft margin approach is

adopted to account for empirical errors. The revised classification equation is provided in equation (2.6). In this context, the weight w is expressed as $w = \sum_{i=1}^m \alpha_i y_i x_i = 1 \dots m$, α_i signifies the weight of the i -th training example, which acts as a support vector.

$$f(x) = \text{sgn} \sum_{i=1}^m \alpha_i y_i (x_i \cdot x) + b \quad (2.6)$$

To address non-linear decision boundaries, SVM employs a mapping of datasets into a higher-dimensional space where data points become linearly separable. This involves replacing x_i with $\phi(x_i)$, which maps the data into higher dimensions. A kernel function is used to compute values analogous to the dot product in this higher-dimensional space. Equation illustrates the decision function using the kernel trick.

$$f(x) = \text{sgn} \sum_{i=1}^m \alpha_i y_i k(x_i, x) + b \quad (2.7)$$

The final output of the SVM is a probabilistic binary classification, indicating landslide membership as either 0 or 1. In a study by [75], four kernel functions—linear, polynomial (degree 2), sigmoid, and RBF—were evaluated for generating landslide susceptibility maps (LSM). The training was conducted using ENVI 5.1 software, with the results exported to ArcGIS for visualization. The findings showed that the RBF-SVM yielded the highest success rate and prediction accuracy, followed by polynomial-SVM, linear-SVM, and sigmoid-SVM in success rates. For prediction rates, the order was RBF-SVM, linear-SVM, polynomial-SVM, and sigmoid-SVM.

In another study [76], the Gaussian radial basis function was used with parameters $\gamma=2$ and $C=100$ for model fitting. The datasets were divided into two equal parts: 50% for training and 50% for testing. The datasets were then swapped, and the arithmetic mean of both runs determined the model's accuracy. The study was iterated multiple times, yielding accuracy levels of 71.0% for a 1% sample size, 71.0% for a 5% sample size, and 88% for a 50% sample size. The open-source LIBSVM package was employed, and the results demonstrated that small datasets could yield good accuracy for susceptibility prediction.

Another comparative study [77] evaluated the performance of three ML models—Evidence Belief Function, Logistic Regression (LGR), and SVM—for landslide susceptibility mapping. The Exelis Visual Information Solutions software, ENVI 5.0, was used for the SVM model. The default kernel function in ENVI 5.0 (RBF) was applied

as shown in (2.8) where K represents the kernel function and γ is the kernel parameter. The results showed that SVM had the lowest AUC among the three models. Despite this, SVM's predictive performance for landslide mapping was deemed acceptable.

$$K(X_i, X_j) = \exp(\gamma X_i - X_j^2) + b \quad (2.8)$$

Additional studies have explored improvements in LSM accuracy by integrating various techniques. For instance, [78] combined geographically weighted regression (GWR), SVM, and particle swarm optimization (PSO) to enhance prediction accuracy. The study area was divided into prediction zones using GWR, with each region subsequently classified for landslide susceptibility using SVM. PSO was then utilized to optimize the SVM classifier parameters. This approach achieved an AUC of 0.978 for landslide susceptibility prediction.

K-Nearest Neighbours

The k-Nearest Neighbours (KNN) algorithm is one of the most intuitive and straightforward machine learning classifiers and has seen increasing application in domains such as landslide susceptibility mapping [76, 79]. This method classifies data points based on the proximity of their features to those in the training set. Specifically, the algorithm determines the class of a new observation by analysing the majority class among its 'k' closest samples in the feature space, typically using a voting mechanism [80,81]. When k is set to 1, classification relies solely on the closest data point, while higher values of k offer more generalized results.

KNN is classified as a non-parametric, instance-based, and "lazy" supervised learning method. The term "lazy learning" denotes that the algorithm delays model generalization until a query is made. Unlike "eager" learners, which build a model during training, KNN stores the entire training dataset and performs computation only during prediction [82]. Its non-parametric nature makes it especially suitable for landslide modelling since the spatial distribution of landslide events often does not conform to predefined statistical distributions. Furthermore, this flexibility allows KNN to adapt naturally to complex data structures.

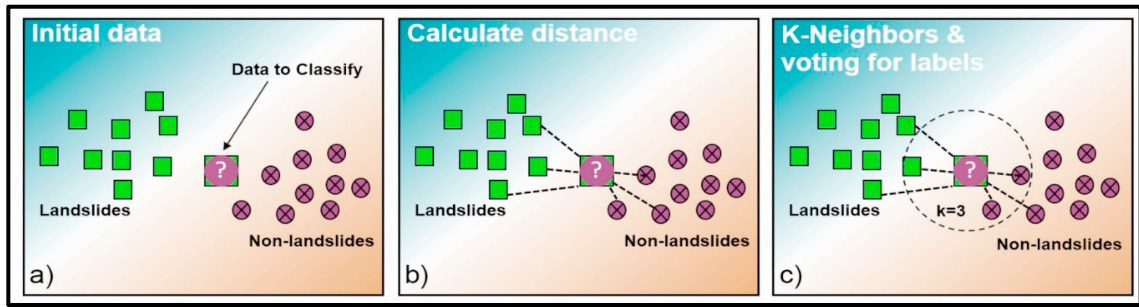


Figure 2.4: K-NN workflow

The typical workflow for implementing KNN Figure 2.4 includes the following steps: (i) input of unlabelled data, (ii) measurement of distance between the new data point and training samples using metrics such as Euclidean or Manhattan distances, (iii) identification of the k-nearest neighbours, and (iv) decision-making based on a voting scheme or averaging process. The Euclidean distance is calculated as the square root of the sum of squared differences between feature values, while the Manhattan distance sums the absolute differences between corresponding features. These distance-based techniques make KNN a powerful, yet computationally intensive approach, particularly effective in cases where patterns in the data are locally defined.

Logistic Regression

Logistic Regression is a supervised machine learning technique primarily employed to address binary classification problems. It serves as a precursor to Artificial Neural Networks (ANN), as ANN is essentially a generalization of logistic regression [83]. Despite its name suggesting it is a regression model, logistic regression is, in fact, a classification method [84]. Unlike traditional regression models, it does not necessitate a linear relationship between input and output variables [85].

This method calculates the probability of an event by evaluating the ratio of occurrences to non-occurrences. Input variables in logistic regression are assigned weights, and there exists an exponential relationship between these weights and the model's output [86]. The mathematical representation of logistic regression is provided in the equation below.

$$\text{Logistic Function} = \frac{1}{1 + e^{-x}} \quad (2.9)$$

In the above equation, the input variable is represented as x , and the logistic function corresponds to the natural logarithm of the odds. The probability value of 0.5 is mapped to a logistic function value of 0. The logistic function operates within the range of 0 to 1,

defined as $P \in [0,1]$ $P \in [0,1]$ [84]. Logistic regression predicts binary outcomes such as Yes/No, Success/Failure, or Occur/Not occur. By using causative factors as input variables, it is possible to predict landslide occurrences in given study area.

As observed by [87], logistic regression is one of the most widely used multivariate methods, effectively establishing relationships between causative factors and the occurrence or non-occurrence of landslides. Logistic function is given by equation (2.10, where dependent variable is represented by p and $p/1-p$ gives the odds ratio.

$$\text{Logistic}(p) = \frac{p}{1-p} \quad (2.10)$$

In studies conducted by and [88], several common machines learning models, including support vector machines (SVM), logistic regression (LGR), and random forests (RF), were combined with convolutional neural networks (CNN). CNN, with its superior learning capabilities, extracts feature that enhance the performance of machine learning models. The findings of the study revealed that the combination of LGR and CNN exhibited a better predictive ability compared to other hybrid models analysed, achieving an overall accuracy of 79.82%. These results demonstrate that hybrid models can outperform conventional machine learning techniques.

Additionally, another study by [89] conducted a comparative analysis between SVM, LGR, artificial neural networks (ANN), and conditional probability methods. Similarly, the work of [90] employed logistic regression and classification and regression trees (CART) as benchmark models, integrating them into a Random Subspace-based Classification and Regression Tree (RSSCART) approach. This study utilized 14 causative factors and 203 landslide locations as inputs. The results highlighted that RSSCART achieved a training AUC of 0.852 and a validation AUC of 0.827, surpassing the performance of both LGR and CART.

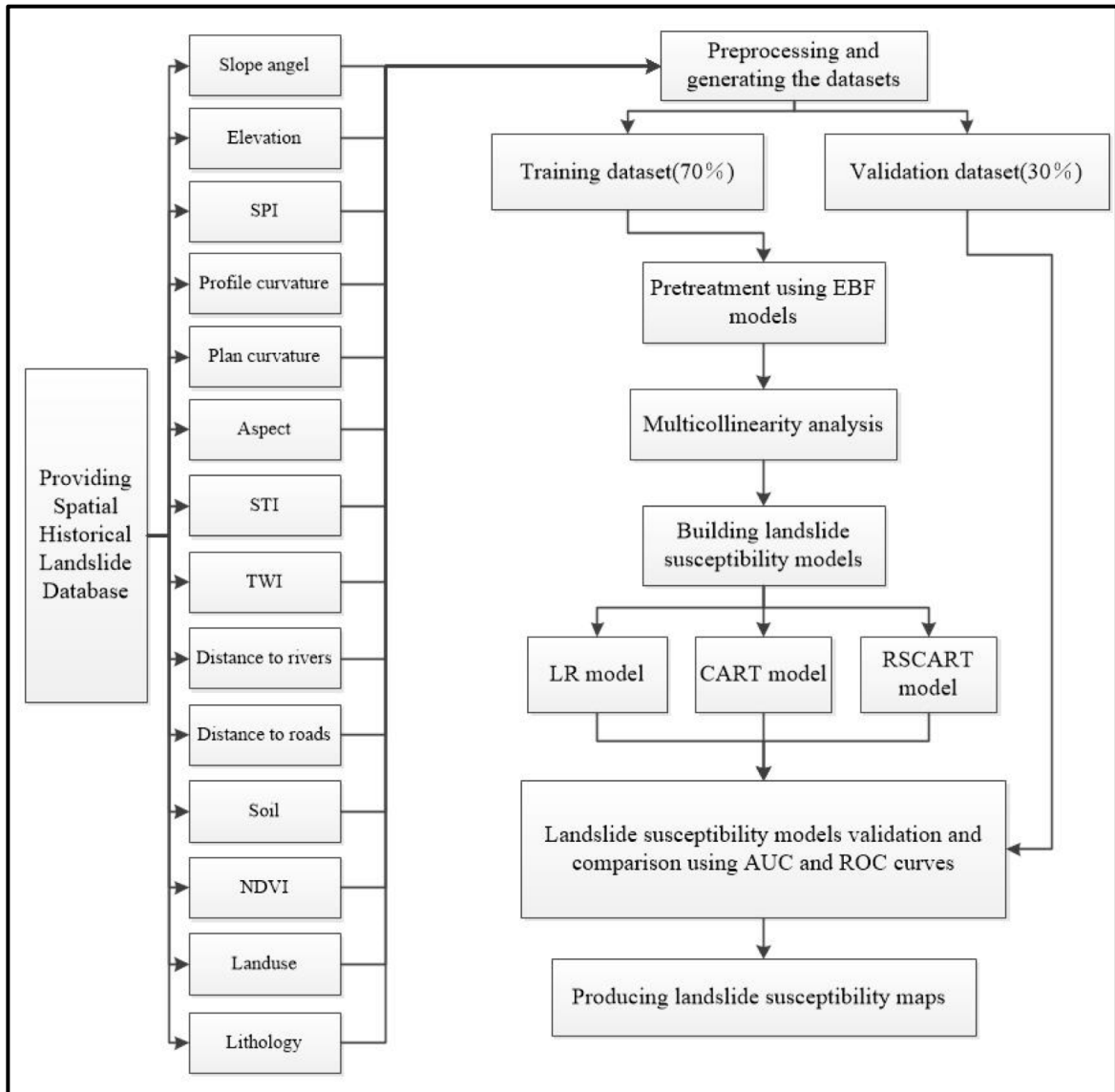


Figure 2.5: Flowchart of the study [90]

Naïve Bayes

The learning algorithm derived from Bayes' theorem is commonly referred to as the Naïve Bayes (NB) model. This algorithm operates under the assumption that classification features are independent of one another within a given class [91]. Although this assumption is often violated in practical applications, the NB model still delivers competitive accuracy.

The key features of the Naïve Bayes model are summarized below:

1. **Efficient Computation:** The computational time for training and classification is proportional to the number of features, making the model computationally efficient.

2. Low Variance, High Bias: The absence of a search mechanism in the model results in minimal variance but may lead to higher bias.
3. Cumulative Learning: The model learns from lower-order probabilities derived from available training data and updates its predictions when new training data is introduced.
4. Direct Posterior Probabilities: The NB model predicts posterior probabilities directly, simplifying the classification process.
5. Noise Resistance: The model utilizes all available features for prediction, minimizing the impact of noise.
6. Robustness to Missing Features: As all features contribute to prediction, missing features do not significantly affect the model's performance, making their absence negligible.

Bayesian network classifier is given in equation (2.11);

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)} \quad (2.11)$$

Considering conditionally independent feature, value of feature can be given by equation (2.12);

$$P(y|x) = \prod_{i=0}^n P(x_i|y) \quad (2.12)$$

In Equation 2.12, n represents number of feature and x_i gives the x value of i^{th} feature.

Furthermore, normalising the numerators in Equation 2.11 of right-hand side,

$$P(x) = \prod_{i=1}^n P(c_i)P(x|c_i) \quad (2.13)$$

Several studies have employed the NB model in landslide susceptibility mapping (LSM). For instance, a study by [92] utilized the NB classifier with 17 landslide causative factor vectors alongside occurrence and non-occurrence vectors. The classification was performed using the NB classifier, as represented by the equation (2.14 below.

$$Y_{NB}f(x) = \{argmax P(y_i), y_i = [event, non - event], \prod_{i=1}^{17} P\left(\frac{x_i}{y_i}\right)\}$$

$$P\left(\frac{x_i}{y_i}\right) = \frac{1}{\sqrt{2\pi}\alpha} e^{\frac{-(x_i - \eta)^2}{2\alpha^2}} \quad (2.14)$$

In above equation, $P(x_i/y_i)$ represents conditional probability, $P(y_i)$ represents prior probability of y_i , standard deviation α for x_i , η indicates the mean.

In study [92], achieved a prediction accuracy of 78.3% and a success rate of 79.2%. The primary goal was to leverage the NB model to aid in damage control by offering preventive support measures.

For comparative evaluation, [93] analysed the performance of RF, SVM, and NB models. The analysis was conducted using eleven landslide causative factors, with the NB model implemented through the e1071 library in R. The NB model classified pixels into landslide and non-landslide categories, achieving an accuracy range of 96% to 97%.

In another case study, SVM and NB models were utilized alongside fractal theory for landslide susceptibility mapping (LSM). This study considered 10 landslide causative factors. Naïve Bayes prediction is given by the equation (2.15).

$$Y_j = \{argmax P(y_i), y_i = [ls, non - ls], \prod_{i=1}^{10} P\left(\frac{x_i}{y_i}\right)\} \quad (2.15)$$

The integration of fractal theory (FRT) to select non-landslide locations improved prediction accuracy, with AUC values reaching 0.96 for SVM and 0.98 for NB. When non-landslide locations were selected randomly, AUC values decreased to 0.708 for SVM and 0.727 for NB, underscoring the importance of carefully selecting non-landslide locations for accurate LSM.

Additionally, [94] implemented the NB algorithm using R 3.0.2 and the R miner package to model LSM. Model performance was assessed through AUC and RMSE, with results showing an AUC value of 0.91 and an RMSE of 0.464.

Artificial Neural Network

Artificial Neural Network (ANN) mimics the functioning of biological neurons and their interconnections, enabling parallel information processing. The fundamental architecture comprises interconnected neurons organized into layers: an input layer, one or more

hidden layers, and an output layer. Each connection between neurons is assigned a numerical value called weight. The output (h_i) of the hidden neuron i can be computed using the equation (2.16) below, where σ represents the activation function, N is the number of input neurons, V_{ij} denotes the weight, X_j is the input to the input neuron, and the threshold of the hidden neuron is defined by equation (1.7) [95].

$$h_i = \sigma \left(\sum_{j=1}^N V_{ij} x_j + T_i^{hid} \right) \quad (2.16)$$

Sigmoid function is popularly used as the activation function to add non-linearity to the hidden layer. Sigmoid function is given by ;

$$\sigma(u) = \left(\frac{1}{1 + \exp(-u)} \right) \quad (2.17)$$

Several studies have utilized artificial neural networks (ANN) for landslide susceptibility mapping (LSM). In a study conducted by [96], 14 causative factors with assigned weights were used as input variables. The model was trained using the backpropagation algorithm, aiming to determine the weight of each causative factor. Validation was performed using data from Korea's Youngin, Janghung, and Boeung regions. The evaluation revealed the highest accuracy of 81.36% and the lowest accuracy of 71.72% across nine test locations. Factors such as slope, land cover, and lineament distance were assigned the highest weights, while soil type and forest density had the lowest weights.

In another comparative study, [97] examined the Mamdani Fuzzy Inference System (FIS) alongside ANN for LSM. This research was noteworthy as it was among the first to apply FIS to LSM. The findings indicated that ANN exhibited less uncertainty compared to FIS, with satisfactory prediction accuracy for both models. The AUC value for ANN was 0.94, while FIS achieved 0.88. Similarly, [98] employed ANN with the backpropagation algorithm for LSM. Weights were assigned small numerical values between 0 and 1. The network output was calculated and compared with the expected output. If discrepancies were found, the weights were adjusted using a correction rule. This iterative process resulted in an AUC value of 0.84, demonstrating good predictive accuracy.

Comparative studies have further established ANN's performance. For instance, [99] compared ANN with the Analytic Hierarchy Process (AHP), with ANN outperforming AHP. Likewise, [100] compared ANN with the Weight of Evidence (WOE) method, and

ANN outperformed WOE in terms of both success rate and AUC prediction accuracy. A study by [101] investigated two ANN models, Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF), for LSM. The findings highlighted that the MLP model trained with the Broyden-Fletcher-Goldfarb-Shanno learning algorithm delivered better results than RBF.

Research by [102] explored two ANN structures: single and double hidden layers. Four training algorithms were analysed, namely batch backpropagation, quick propagation, conjugate gradient descent (CGD), and the Levenberg–Marquardt algorithm. Although CGD was slower than the others, it yielded the highest AUC success rate and best prediction accuracy. Another study by [103] classified landslides into four types complex, slide, rockfall, and slow-moving—and employed the MLP-ANN model for LSM. ArcGIS was used for mapping and MATLAB for supervised learning. The study achieved an AUC value exceeding 0.9 for the MLP-ANN model. These results underscore ANN's exceptional prediction accuracy and its effectiveness for landslide susceptibility mapping.

Hybrid Models for LSM

The accuracy of machine learning (ML) models in generating landslide susceptibility mapping (LSM) is significantly influenced by the selection of landslide causative factors. A thoughtful and strategic selection of these factors is essential for improving the accuracy of LSM. Hybrid techniques, which combine different ML models or ML models with feature selection and optimization methods, have gained prominence in this domain. Traditionally, multicollinearity analysis has been performed to select causative factors by calculating Variance Inflation Factor (VIF) and Tolerance (TOL) values [104-106]. Other commonly used correlation methods include Pearson's correlation analysis [107], Correlation Feature Selection (CFS) [108], Conditional Attribute Evaluation (CAE) [109], and Spearman's rank correlation coefficient [110].

For instance, [111] explored a novel hybrid model combining Rough Set (RS) theory with SVM to map landslide susceptibility on a regional scale using multisource data. RS was employed for selecting causative factors, while SVM was used for prediction. The combination of RS-SVM demonstrated better performance compared to SVM alone. Similarly, [78] incorporated local environmental factors into their analysis using the Geographically Weighted Regression (GWR) method for improved accuracy. The hybrid

GWR-PSO-SVM model, employing SPSS Clementine 12 software to define causative factors, achieved an AUC value of 0.978.

Another study by [112] investigated a hybrid model integrating Sequential Minimal Optimization (SMO) with SVM. The SMO approach offered benefits such as simplicity, speed, and suitability for large datasets, while reducing problem complexity. When compared with other hybrid models, such as cascade generalization optimization-based SVM and standalone ML models like NB-tree and SVM, SMO-SVM outperformed with an AUC value of 0.824.

In [106], quantitative landslide prediction was conducted using a hybrid model that combined Fractal Dimension, Index of Entropy (IOE), and SVM. The hybrid models demonstrated superior performance, with Fractal-IOE achieving an AUC value of 0.85 and Fractal-SVM reaching 0.97. These results emphasized that hybrid models outperform standalone IOE or SVM models.

To address uncertainties and improve spatial agreement in ML-based LSM, [113] combined K-Nearest Neighbours (KNN), Multi-Layer Perceptron (MLP), RF, and SVM models. The hybrid models exhibited better spatial agreement with higher correlation coefficients. Although these models demonstrated improved prediction accuracy, the use of secondary input data resulted in lower RMSE values.

The study by [114] employed a hybrid model consisting of GeoSOM, RF, and an ensemble model (ANN-SVM-Gradient Boosting Decision Tree [GBDT]). GeoSOM was used for clustering study locations, RF and Pearson correlation coefficient for feature selection, and ANN-SVM-GBDT for generating LSM. While this model showed promise, random selection of non-landslide locations introduced potential errors, and the time required was longer compared to other hybrid models.

In their study [88] integrated Convolutional Neural Networks (CNN) with SVM, RF, and Logistic Regression (LGR). CNN was utilized for feature selection, while LSM was generated using various ML models. However, overfitting was identified as a challenge in such hybrid models. To enhance predictive performance, Fractal Theory (FRT) was employed to select non-landslide locations [115], resulting in improved accuracy for SVM and Naïve Bayes models.

Additionally, [116] explored the integration of Bayesian Optimization (BO) with RF and GBDT, which enhanced the predictive accuracy of both models. [109] studied the

combination of multi-boost with Radial Basis Function Neural Network (RBFN) and Credal Decision Tree (CDT). CDT achieved the highest AUC value of 0.77 when combined with multi-boost, although the combined model's AUC was lower compared to other hybrid approaches.

Several other studies have documented the use of various ML model combinations and integrations with feature selection and optimization techniques, further demonstrating the potential of hybrid methods to enhance LSM accuracy.

Ensemble Models for LSM

The method of combining different conventional machine learning (ML) models using averaging or voting systems is referred to as *Ensemble Methods*. These methods compensate for the limitations of individual models by leveraging their collective strengths, resulting in improved prediction accuracy and enhanced landslide susceptibility mapping (LSM) [117]. Numerous studies have demonstrated the utility of ensemble techniques in improving the accuracy of LSM.

Decision Tree-Based Ensemble Methods

One commonly used ensemble method is Light-GBM, a decision tree-based model known for its efficiency and scalability [118]. The Random Forest (RF) algorithm, a foundational ensemble model, has been widely employed for LSM due to its simplicity, low computational requirements, and robustness [110, 118–123]. Another frequently utilized ensemble model is Random Subspace, which leverages different features in the entity space for classification, as proposed by [108]. Despite its advantages, overfitting remains a key challenge with this model.

The Canonical Correlation Forest (CCF) is an advanced decision tree-based ensemble model that uses multiple decision trees and majority voting for predictions. It improves upon RF and Rotation Forest (RotFor) techniques [104]. Another decision tree-based algorithm is Chi-Squared Automatic Interaction Detection (CHAID), which is useful for automatic classification by analysing numerous landslide causative factors [124].

Gradient Boosting-Based Ensemble Methods

A relatively new model, Natural Gradient Boosting (NG Boost), was developed by [125] to address limitations of existing gradient boosting methods. Similarly, XGBoost, a popular supervised learning model, prevents overfitting through bagging-bootstrap

aggregation, making it a preferred choice for LSM [126, 127]. The Multi-Boost model, combining AdaBoost and wagging techniques, was developed by [128].

The Bagging algorithm, proposed by [129], is widely used in natural hazard studies. This model involves three steps:

1. Creating subsets of the training dataset through random resampling.
2. Generating classification models for each subset.
3. Aggregating sub-models to create the final classification model [130].

Dagging, introduced by [131], combines models generated using subsets of the training dataset, which may use the same or different classification algorithms. Cascade Generalization, proposed by [132], removes dataset biases and improves accuracy through hierarchical model integration.

DECORATE, another ensemble technique, generates artificial datasets to enhance diversity. Proposed by [133], this model calculates the standard deviation of the training dataset and applies Gaussian rules to create artificial samples.

Heterogeneous Ensemble Methods

Heterogeneous ensemble methods combine diverse models for better non-linear representation and generalization. Stacking [134-136] integrates multiple ML models to leverage their diversity, while Blending, a variation of stacking proposed by [137], divides the dataset into training and validation sets to prevent information leakage.

Simple Averaging models output the average prediction from base classifiers, while Weighted Averaging employs AUC-based weighted averages to integrate predictions [136]. Qualitative Ensemble Methods use decision matrices to combine predictions for comprehensive decision-making [138].

Region-Based and Probability-Based Ensemble Techniques

The Semi-Quantitative Partition Method divides the study area into sub-regions based on features like topography. The best predictions from each sub-region are combined for LSM. In the Probability Weighted Model, different weights are assigned to base classifier predictions [138].

The Median Probability Model calculates the median probability across classifiers for final predictions. Committee Averaging converts base classifier probabilities into binary

data and applies a simple voting system. The Weighted Mean of Probability computes the mean weight of probabilities from selected base classifiers to produce the final prediction [139].

Deep Learning Model for LSM

Deep learning (DL) models have proven capable of identifying complex structures in high-dimensional data, allowing them to achieve superior performance [140]. These models are essentially representation learning techniques composed of multiple layers, such as batch normalization, flatten, dense, and dropout layers. The batch normalization layer ensures consistent data distribution for effective learning, while the dense layer handles class distribution. To improve model generalization and reduce overfitting, dropout layers are employed. Another critical component of DL architectures is the activation function, which must be differentiable and selected based on the specific requirements of the model. For landslide susceptibility mapping (LSM), the sigmoid activation function is commonly used due to the non-linear nature of the task [141].

Popular DL methods for LSM include convolutional neural networks (CNNs), long short-term memory networks (LSTMs), deep neural networks (DNNs), and recurrent neural networks (RNNs). CNNs, for example, comprise three main components: convolutional layers, down-sampling layers, and fully connected layers. Convolutional layers extract feature information using multiple kernels, enabling training with fewer parameters compared to fully connected networks. Following the convolutional layer, activation functions such as the widely used rectified linear unit (ReLU) are applied. Down-sampling layers help reduce feature sizes and mitigate overfitting tendencies [142].

DNNs have gained popularity in natural hazard studies due to their versatility. These networks include fully connected layers, dropout layers, hyperparameters for optimization, activation functions, and optimization algorithms [143]. On the other hand, RNNs excel in processing temporal data through recurrent hidden states that learn from both past and present inputs [136]. However, RNNs can struggle with long-term data dependencies, which LSTMs address with their advanced recurrent structure. The basic unit of an LSTM is the memory cell, equipped with gates to manage stored information. The input gate regulates cell states, while the forget gate determines the retention or deletion of previous states [141].

Table 2.1 : Merits and Demerits of ML Models for LSM

Method	Type	Merits	Demerits
Random Forest Method	Modern/Machine learning Ensemble model	Overfitting is reduced as compared to other methods.	Increases computational complexity with a greater number of trees and chances of overfitting reduces with increased number of trees.
Support Vector Machine	Modern/Machine learning Standalone model	Gives High accuracy compared to other models in the segment.	Requires large data sets which makes process lengthy.
Artificial Neural Network	Modern/Machine learning Standalone model	It offers good prediction accuracy.	It is complex to interpret the data generated by this model and is a complex task.
Naive Bayes	Modern/Machine learning Standalone model	Simple and fast method as compared to other methods in the same context.	Features are assumed to independent to each other and presence or absence of any feature does not affect the model.
Hybrid Models	Modern/Machine Learning	Combines strengths of multiple models	Complexity in model compiling and upto certain extent overfitting is observed in the mode.
Ensemble Models	Modern/Machine Learning	Reduces bias and variance.	Overfitting is the major issue in ensemble models.
Deep Learning Models	Modern/Machine Learning	Gives High accuracy and capable of handling complex patterns of data.	Requires large data sets and complex calculations are involved.

2.3 Evaluation Metrics for Machine Learning Models

Assessing the effectiveness of machine learning models in landslide susceptibility studies requires a comprehensive set of evaluation metrics. In machine learning, performance

metrics derived from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are used to evaluate model effectiveness in terms of accuracy, precision, recall (sensitivity), and overall predictive performance. Metrics including accuracy, precision, recall, specificity, and F1-score help determine how well the model identifies actual landslide areas while minimizing false predictions. Moreover, the Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) provide a way to evaluate the model's discrimination ability independently of any classification threshold [72].

Cohen's Kappa (K) statistic is also used to account for agreement occurring by chance, offering a more robust measure of classification accuracy. Combined, these metrics facilitate thorough validation of machine learning models prior to their practical application in landslide hazard mapping.

The Table 2.2 below presents these key validation metrics along with their formulas and brief explanations.

Table 2.2 : LSM Approaches in Indian Himalayan Region

Metric	Formula	Description
Accuracy	$TP + TN / TP + TN + FP + FN$	Overall correctness of the model.
Precision (PPV - Positive Predictive Value)	$TP / (TP + FP)$	Fraction of predicted positives that are correct.
Recall (Sensitivity) (TPR- True Positive Rate)	$TP / (TP + FN)$	Fraction of actual positives correctly identified.
Specificity (TNR - True Negative Rate)	$TN / (TN + FP)$	Fraction of actual negatives correctly identified.
Negative Predictive Value (NPV)	$TN / (TN + FN)$	Fraction of predicted negatives that are correct.

Metric	Formula	Description
F1-Score	$2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$	Harmonic mean of precision and recall.
False Positive Rate (FPR)	$\text{FP} / (\text{FP} + \text{TN})$	Proportion of negatives incorrectly predicted as positive.
Kappa (K)	$(\text{Po} - \text{Pe}) / (1 - \text{Pe})$	Agreement between predictions and actual values adjusted for chance.
Po (Observed Agreement)	$(\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$	Proportion of correct predictions.
Pe (Expected Agreement)	$[(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) + (\text{FN} + \text{TN}) \times (\text{FP} + \text{TN})] / (\text{TP} + \text{TN} + \text{FP} + \text{FN})^2$	Expected agreement by chance.
ROC (Receiver Operating Characteristic)	Plot of TPR vs FPR at various thresholds	Graphical evaluation of model performance across thresholds.
AUC (Area Under Curve)	Overall ability of the model to distinguish between classes	Value between 0 and 1 indicating overall classification performance.

2.4 LSM approaches and Regional Application

Landslide susceptibility modelling has been widely studied across the globe, with researchers employing diverse methods to identify and map landslide-prone areas. Numerous studies have been conducted in different regions, demonstrating the effectiveness of various techniques.

In India, landslide susceptibility research has largely focused on the Himalayan region due to its geographical vulnerabilities. A comprehensive review by [144] examining publications from 1990 to 2019 confirmed a significant concentration of landslide

research within the Indian Himalayan Region, with a clear bias towards Uttarakhand. The regional application of Landslide Susceptibility Mapping (LSM) techniques reflects this imbalance, as depicted in the pie chart Figure 2.6.

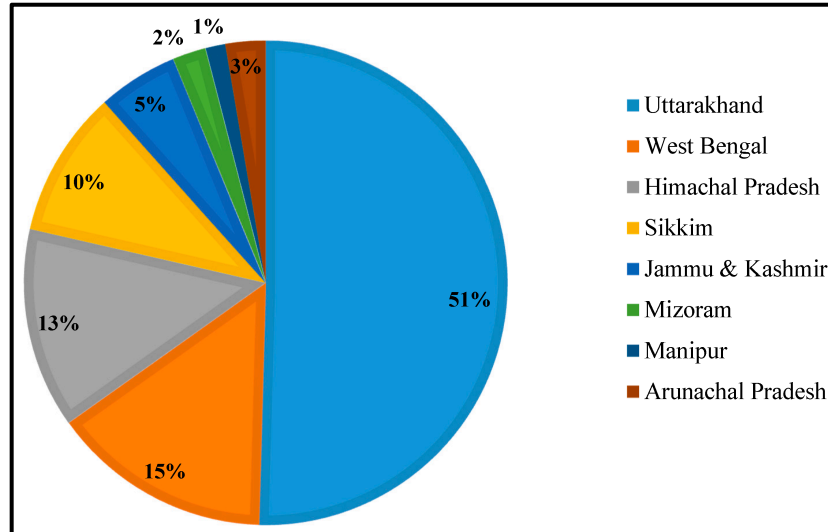


Figure 2.6: Landslide Studies Distribution in India [144]

Uttarakhand alone accounts for 51% of the total LSM-related publications, largely due to its recurring and severe landslide events, complex terrain, and high exposure of human settlements in hazard-prone zones. West Bengal, with a focus on the Darjeeling Himalayas, contributes 15%, followed by Himachal Pradesh at 13% and Sikkim at 10%. Other states like Jammu & Kashmir (5%), Mizoram (2%), Manipur (2%), and Arunachal Pradesh (3%) remain underrepresented in LSM studies. This skewed distribution is primarily driven by the higher population density in the western Himalaya, where landslides pose a greater threat to human life and infrastructure. Supporting this, Pham et al. [145] reported that between 2007 and 2015, landslide events in the region accounted for 5,228 casualties—representing 82.9% of all disaster-related fatalities in the area.

In the Himalayan region, several studies have explored various approaches for landslide susceptibility mapping. Kanungo et al. (2006) [146] conducted a comparative study of conventional, fuzzy, and combined neural-fuzzy methods in the Darjeeling Himalayas, concluding that the combined neural-fuzzy approach was the most effective. Machine learning algorithms, including U-Net, were applied in Nepal's Rasuwa district, although specific results were not detailed [147]. In the Sikkim Himalayas, terrain analysis and spatial assessments have been employed [148], while univariate statistical analysis has been conducted in the Garhwal Himalayas [68]. A knowledge-based approach was

adopted in the Chamoli-Joshimath region of the Garhwal Himalaya [149], and a GIS-based statistical method applicable to the entire Himalayan region was developed [41]. Furthermore, Ray et al. (2007b) utilized a fuzzy-based approach in the Garhwal Himalayas [150]. These efforts demonstrate the range of methodologies applied in the Himalayan region, with ensemble and deep learning techniques being relatively less explored within the scope of this study. There are several studies in Indian Himalayan region with GID based approach and ML approach which are summarized in Table 2.3 : LSM Approaches in Indian Himalayan Region

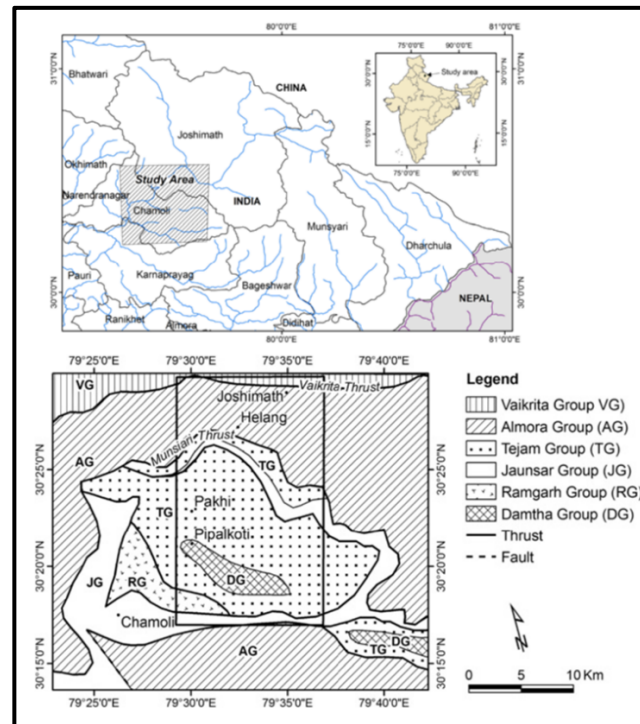


Figure 2.7: Study Area in Garhwal Himalaya and its Geology [149]

Table 2.3 : LSM Approaches in Indian Himalayan Region

Ref	Method/Approach	Study Area
[28]	Bivariate Method (Weighted Overlay Method)	Rudra Prayag district, Garhwal Himalaya
[41]	GIS based Statistical Method	Himalayas
[68]	Univariate Statistical Analysis	Garhwal, Himalayas
[146]	Comparative Study (Conventional, Fuzzy, Combined Neural)	Darjeeling Himalayas

[147]	ML Algorithm and U-Net	Rasuwa District, Himalaya, Nepal
[148]	Terrain Analysis and Spatial Assessment	Sikkim Himalaya
[149]	Knowledge Based Approach	Chamoli-Joshimath, Garhwal Himalaya
[150]	Fuzzy based Approach	Garhwal Himalayas
[151]	GIS Based Weighted Overlay Method	Dehradun and Mussoorie, Uttarakhand
[152]	Bivariate Method (Weight of Evidence)	Rudra Prayag district, Garhwal Himalaya
[153]	Bayesian Logistic Regression Models	Road Corridors in Indian Himalaya
[154]	GIS Approach	Ramganga Catchment, Himalaya
[155]	IVM (Information Value Method)	Sikkim Himalayas
[156]	Rainfall Threshold	Nepal Himalaya
[157]	Probabilistic Approach	Bhagirathi River Basin, North Himalaya
[158]	IVM	Giri Valley, Himachal Pradesh
[159]	BIS-based LHEF	Nainital, Kumaun Himalayas
[160]	WoFE method (Evaluated using BIS)	Darjeeling Himalayas
[161]	IVM (GIS based)	Sikkim Himalayas

2.5 Key Findings and Research Gaps

It is observed that no developed equation or well-proven theory adequately captures the complexity of landslides. Several characteristics collected from in-depth soil analysis will be needed for deterministic susceptibility modelling of a vast area. However, a small number of deterministic methods have also been taken into consideration for the analysis of individual landslides at sites. However, landslide features vary by location, therefore a model created for one place would not be appropriate for another with distinct site

characteristics. As a result, the quantitative machine learning approach is adopted for landslide susceptibility assessment.

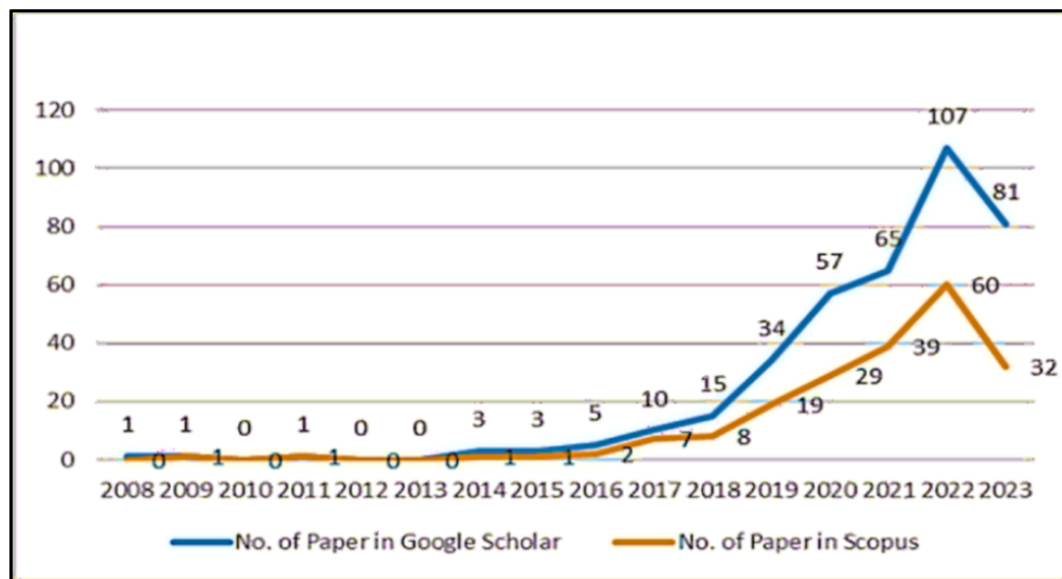
Numerous elements, which can fluctuate depending on the site, must be taken into account while mapping landslide susceptibility. Due to high-resolution data and effective procedures, the use of GIS and remote sensing has increased the quality of landslide susceptibility mapping, which relies heavily on landslide causative variables. The most common independent feature in the majority of LSS modelling is the slope.

For LSM, AHP is the most used Multi Criteria Decision Making technique. However, because of their dynamic change, assigning weight to input parameters—like causal factors is difficult. Map validation is impacted by the subjectivity of weight assignment in heuristic and semi-quantitative approaches. Quantitative (probabilistic) approaches, such as FR and ML, employ statistics and mathematics to unbiasedly allocate weights according to the correlation between various factors and the incidence of landslides.

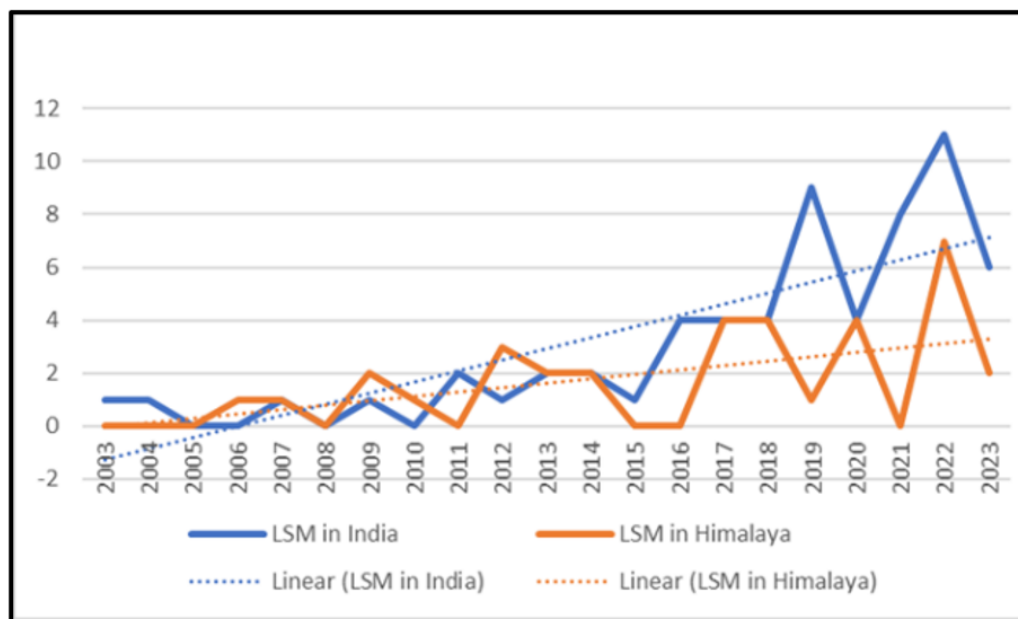
A review of landslide susceptibility mapping (LSM) methods justifies the critical importance of identifying and analysing landslide causative factors in developing accurate susceptibility maps. A comprehensive analysis of numerous articles and in-depth searches through reputable databases such as Scopus and Google Scholar reveal a clear upward trend in the application of machine learning techniques for LSM over the years. Notably, there has been a significant surge in the adoption of these methodologies post-2017, reflecting an increasing interest in utilizing advanced tools to enhance the accuracy and reliability of landslide susceptibility assessments.

Figure 2.8 (a), (b), and (c) have been generated based on an literature review conducted using databases such as Scopus and Google Scholar. Keywords including “Landslide Susceptibility Mapping,” “Landslide Susceptibility Mapping in India,” and “Landslide Susceptibility Mapping in Himalaya” were used to retrieve relevant publications. To understand the trend, publications in last 15 years 2008 to 2023 was taken into consideration. The graphical representation of the results was developed using the Microsoft excel to reflect key trends and patterns in the field . The Figure 2.8 (a) shows the number of papers published within 2008 to 2023. A regional analysis shown in Figure 2.8 (b) further highlighted a research gap in the Himalayan region, where studies on landslide susceptibility mapping remain relatively scarce. This observation underscores the necessity for more detailed investigations in this geologically and environmentally

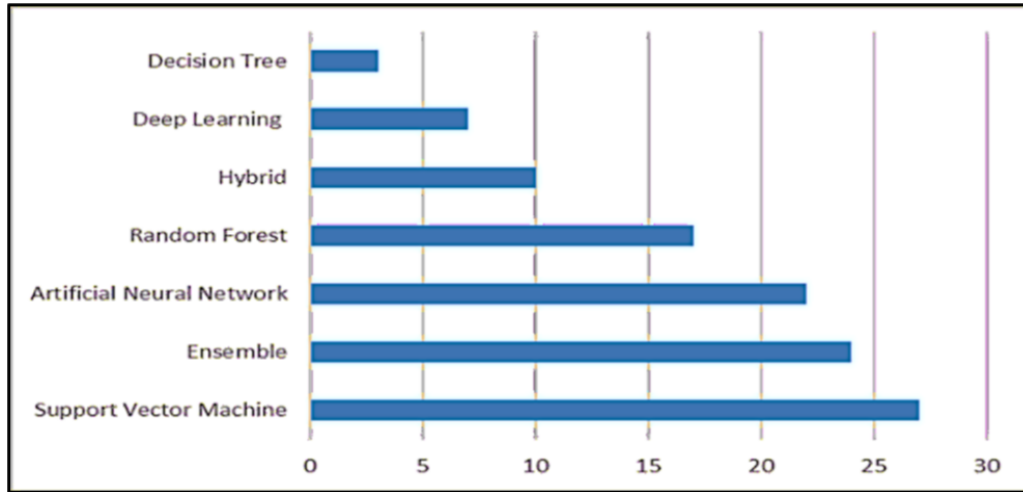
critical area, given its high vulnerability to landslides. In examining various machine learning models employed in LSM, the Support Vector Machine (SVM) emerged as one of the most used methods. Additionally, the adoption of ensemble methods, artificial neural networks, random forests, and hybrid approaches has shown a steady increase, signalling their growing popularity in recent years as shown in Figure 2.8 (c). However, the use of deep learning models and decision trees remains comparatively underexplored in the context of LSM. Based on these findings, the following areas for further research and exploration are suggested to address existing gaps and advance the field.



a.



b.



c.

Figure 2.8: (a) Number of Published papers in Scopus and Google Scholar on LSM, (b) Number of Papers published for LSM in Himalayan region, (c) Number of published papers with different models.

2.6 Problem Statement and Objectives

Landslides represent a significant hazard in mountainous regions, particularly in the Indian Himalayan region, where diverse geological and climatic conditions make the area highly susceptible. Extensive literature reviews were conducted to analyse both conventional and modern approaches to landslide susceptibility modelling (LSM). These reviews also highlighted the most used landslide causative factors, the spatial distribution of studies within the Indian Himalayan region, and the gaps in research. While modern machine learning techniques have gained prominence, limited studies focus on ensemble models or artificial intelligence approaches in this region. Furthermore, no comprehensive landslide susceptibility maps have been developed for certain critical areas within the Himalayas using advanced methods. The primary aim of this study is to enhance landslide susceptibility assessment through the integration of machine learning techniques and geospatial tools. To achieve this, the following specific objectives have been formulated;

1. **To generate landslide susceptibility maps** for the selected study area by employing various machine learning algorithms in combination with Geographic Information System (GIS) technologies.

2. **To identify and select the most relevant landslide conditioning factors**, encompassing geomorphological, geological, hydrological, and anthropogenic variables that contribute to slope instability.
3. **To evaluate and validate the predictive performance** of the developed models using established statistical metrics, ensuring reliability and robustness in real-world applications.
4. **To delineate and classify the spatial extent of landslide-prone zones** based on model outputs, thereby supporting future land-use planning and the formulation of appropriate risk mitigation strategies.

2.7 Original Contribution of Thesis

The present study contributes to the field of landslide susceptibility mapping by integrating advanced machine learning techniques with geospatial analysis in a data-scarce and topographically complex Himalayan environment. Unlike previous approaches, this research emphasizes the use of ensemble models and a comprehensive feature selection strategy to enhance prediction accuracy and interpretability. The findings aim to bridge existing methodological gaps and offer practical insights for hazard assessment and planning in vulnerable mountain regions. Key contributions are as listed below;

1. **Comprehensive Landslide Inventory Creation:** Landslide locations in the study area were digitized and mapped using GIS platforms. A total of 1,559 landslide points were identified and recorded on Google Earth. This inventory was further developed using QGIS to serve as a foundational dataset for modelling.
2. **Acquisition and Preparation of Causative Factor Data:** Fifteen critical landslide causative factors were identified, including curvature, elevation, NDVI, rainfall, slope, Stream Power Index (SPI), Topographic Wetness Index (TWI), aspect, and proximity to lineaments, roads, and streams. Additional factors such as geology, geomorphology, landcover, and soil data were also prepared. All datasets were converted into raster formats at a consistent resolution of $30\text{m} \times 30\text{m}$ using GIS platforms.

3. **Generation of Non-Landslide Points:** To balance the dataset, 1,559 random non-landslide points were created, adhering to buffer distance and slope threshold criteria. This ensures robust modelling by providing both positive and negative samples.
4. **Implementation of Semi-Quantitative and Quantitative Models:** The Analytical Hierarchy Process (AHP) was applied to develop a baseline understanding of landslide susceptibility. Frequency Ratio (FR), a bivariate statistical method, was employed to assess factor contributions to landslide occurrences.
5. **Application of Machine Learning Models:** Six standalone machine learning techniques were implemented, including Logistic Regression (LGR), Support Vector Machines (SVM), Kernel-SVM (K-SVM), K-Nearest Neighbours (K-NN), Naïve Bayes (NB), and Decision Tree Classification (DT). Advanced ensemble methods such as Random Forest (RF), Extreme Gradient Boosting (XG-Boost), stacking, and voting classifiers were used to enhance model performance. Neural networks were also employed to model complex nonlinear relationships in landslide susceptibility.
6. **Validation and Model Comparison:** Model performance was rigorously evaluated using Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) metric. This comparative analysis highlighted the strengths and weaknesses of each method in predicting landslide susceptibility.
7. **Development of Landslide Susceptibility Maps:** Landslide susceptibility maps were generated for the study area, incorporating the outputs of the best-performing models. These maps provide critical insights for future planning, disaster management, and regulatory measures to mitigate landslide risks.
8. **Identification of High-Risk Zones within Study Area:** Landslides susceptibility maps were further used to determine percentage pixel distribution of different susceptibility classes in different parts of the study area. Subsequently, the results were used to derive high risk zones within the study area.

To summarize, work in this thesis integrates diverse machine learning techniques with GIS technologies to produce reliable and actionable landslide susceptibility maps. These contributions not only advance the field of LSM but also offer practical tools for managing landslide-prone regions in the Indian Himalayas.

CHAPTER 3

Modelling Approach and LSM Generation

3.1 Methodology

The methodology adopted in this study follows a systematic approach to analyse and model landslide susceptibility by integrating spatial data, feature selection, and machine learning models. The workflow consists of multiple phases, including data collection, landslide inventory preparation, feature extraction, data preprocessing, model built up and model validation.

The process as shown in Figure 3.1 begins with the selection of a study area where landslide occurrences have been recorded. The landslide inventory is compiled using both point and polygon datasets, which provide precise locations and spatial extents of past landslides. The dataset is enriched with conditioning factors obtained from multiple sources, including topographical, geological, and hydrological attributes. Digital Elevation Model (DEM) data is employed to derive slope, aspect, elevation, curvature, Topographic Wetness Index (TWI), and Stream Power Index (SPI). Additional factors such as land cover, geomorphology, soil properties, and proximity to linear features (roads and lineaments) are also included.

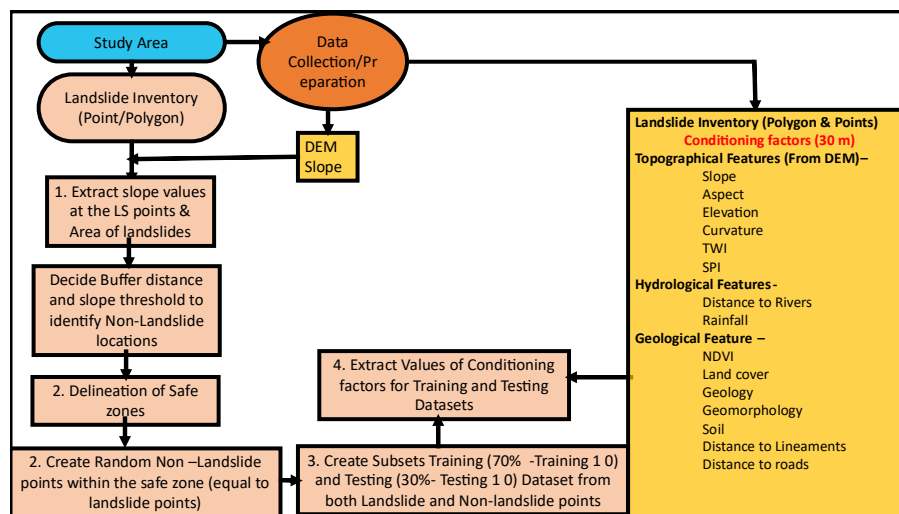


Figure 3.1: Workflow for data collection, feature extraction, and dataset preparation

To ensure a balanced dataset, the Grid Unit Method is employed for feature extraction at both landslide and non-landslide locations. The study area is divided into equal-sized grid cells to maintain uniform spatial coverage, with each cell representing a spatial unit for terrain attribute analysis. Landslide points from the inventory are mapped onto the grid, and an equal number of non-landslide points are generated using systematic random sampling. To select non-landslide locations for the analysis, a buffer distance of 250 meters was applied around all known landslide points to exclude areas directly influenced by landslide activity. Non-landslide sites were then identified beyond this buffer zone on slopes with a threshold value of less than 20 degrees. This criterion is based on the observation that the majority of landslides occur on slopes steeper than 20 degrees, as seen from the slope distribution analysis. Combining the buffer distance with slope thresholds ensures the selection of stable locations, reducing spatial bias and improving the reliability of the landslide susceptibility model.

Once the landslide and non-landslide datasets are prepared, the extracted features undergo a rigorous selection process. Feature selection was performed in two stages: unsupervised and supervised methods. In the unsupervised phase, techniques such as Generalized Variance Inflation Factor (GVIF), variance thresholding, and correlation analysis are used to eliminate redundant and highly correlated features. This ensures that only independent and relevant variables are retained for further analysis.

In the supervised phase, three major feature selection methods filter, wrapper, and embedded approaches are applied. These methods identify the most significant predictors that contribute to landslide susceptibility, enhancing model efficiency and reducing computational complexity. The optimal subset of features is finalized based on their relevance and importance in predicting landslides.

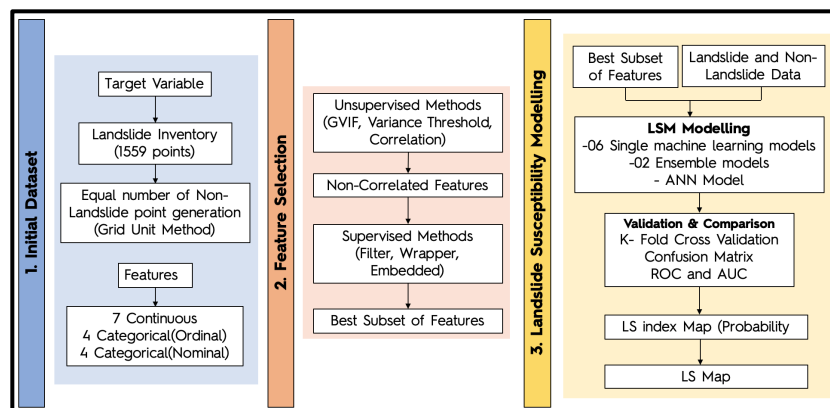


Figure 3.2: Methodological Framework for Machine Learning Modelling

With the refined dataset, multiple machine learning models are implemented to assess landslide susceptibility. The methodological framework describe above is shown in Figure 3.2. The study employs six single machine learning models along with two ensemble models and an Artificial Neural Network (ANN) model. The models are trained using 70% of the dataset, while the remaining 30% is used for testing. The performance of these models is evaluated using validation techniques such as K-fold cross-validation, confusion matrix, Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC) metrics. These assessments ensure robustness and accuracy in susceptibility predictions.

The trained models generate landslide susceptibility maps by calculating Landslide Susceptibility Index (LSI) values for each grid unit. These probability-based susceptibility maps indicate regions prone to future landslides. The final output is a classified landslide susceptibility map, which provides insights into high-risk areas, aiding in disaster management and mitigation efforts.

By integrating spatial analysis, feature selection techniques, and advanced machine learning models, this methodology offers a comprehensive approach to landslide susceptibility assessment, ensuring accuracy and reliability in hazard prediction.

3.2 Study Area

The Northern and North-Eastern Himalayan states, including Uttarakhand, Himachal Pradesh, Arunachal Pradesh, Nagaland, Manipur, and Mizoram, experience significant landslide risks due to their steep mountainous terrain and unstable geological conditions. Among these states, Himachal Pradesh is particularly vulnerable, facing 25 out of 33 identified natural hazards. The region's topography, with elevations ranging from 245 m to 6,610 m, consists of deep valleys and steep slopes, making it highly susceptible to landslides.

The Sutlej Basin, situated within the young and geologically unstable western Himalayan fold mountains, is especially prone to landslides. These events pose severe threats to local livelihoods and infrastructure, particularly in settlements near National Highway-22 (now NH-5) and the old Hindustan-Tibet Road, both of which are critical for transportation and connectivity.

Extreme weather conditions further exacerbate landslide occurrences in the region. According to a study by the Indian Meteorological Department, Himachal Pradesh recorded 36 cloudburst incidents between 1990 and 2001, averaging about three per year, with a peak of seven in the year 2000. Kullu district experienced the highest number of cloudbursts (15), followed by Shimla (6), Kinnaur (4), and Mandi (3). Within Kinnaur district, the Kalpa and Nichar subdivisions are particularly vulnerable to extreme weather events.

This research focuses on the lower Sutlej Basin as shown in **Figure 3.3**, encompassing Kinnaur, Shimla, part of Kullu, and part of Manali districts. This area holds strategic importance due to the presence of key border roads and critical infrastructure. Geographically, the study region extends between latitudes $30^{\circ} 46' 38.942''$ N to $32^{\circ} 05' 9.545''$ N and longitudes $76^{\circ} 40' 48.663''$ E to $78^{\circ} 59' 42.463''$ E, covering an approximate area of **13,433 sq. km**. The study area falls in 39 Survey of India (SOI) toposheets of scale 1:50,000 as shown in Figure 3.4.

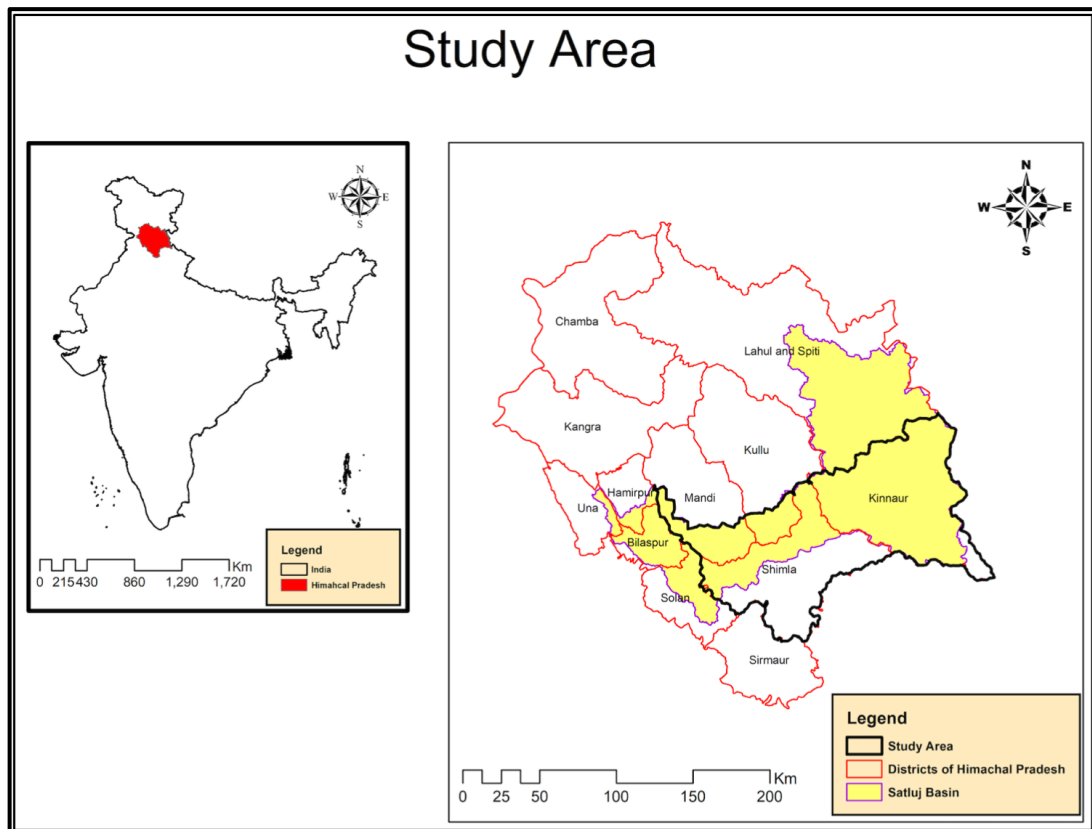


Figure 3.3: Study Area

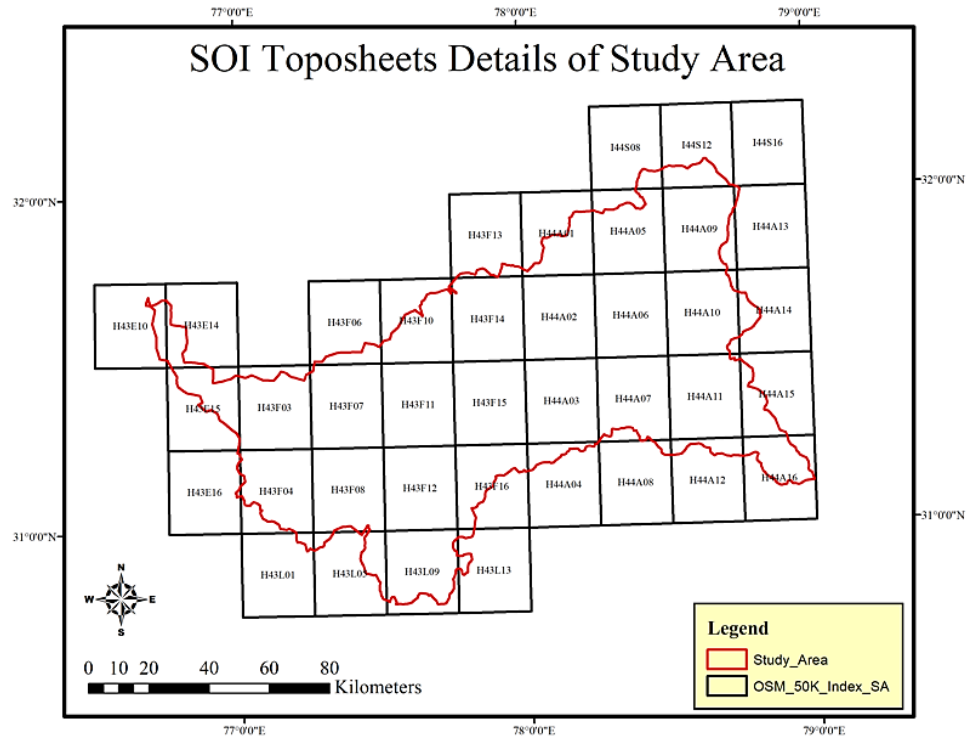


Figure 3.4: SOI Toposheets details of Study Area

3.3 Data Acquisition and Preparation

Apart from compiling a landslide inventory, selecting the appropriate causative factors is essential for accurate landslide susceptibility assessment. While landslide susceptibility modelling (LSM) relies on multiple input parameters, a larger number of factors does not always guarantee higher prediction accuracy. The selection of relevant causative factors depends on various aspects, including the type and mechanism of landslides, characteristics of the study area, scale of analysis, data availability, and the modelling techniques employed. However, no universally accepted guideline exists for determining the most suitable set of factors for a given study. The selection of causative features varies across different research works, with many studies utilizing feature selection techniques to identify the most influential parameters.

In this study, the selection of causative factors was based on the geomorphological, geological, hydrological, meteorological, and anthropogenic characteristics of the study area. Generally, landslide-causing factors are categorized into four groups: (i) topographical, (ii) hydrological, (iii) geological, and (iv) anthropogenic. For landslide susceptibility modelling, fifteen factors contributing to landslides were selected:

curvature, elevation, NDVI, rainfall, slope, Stream Power Index (SPI), Topographic Wetness Index (TWI), aspect, distance to lineaments, distance to roads, distance to streams, geology, geomorphology, landcover, and soil. These data were acquired and prepared into raster format at a consistent resolution of 30m x 30m using a GIS platform. For this research, ten significant factors were selected: elevation, slope, curvature, distance to streams, NDVI, landcover, soil type, geology, geomorphology, and distance to roads. The data preparation and analysis were conducted using ArcGIS and QGIS. The Table 3.1 shows the Features, their sources and tools used in preparation.

Table 3.1 : Data Acquisition and Preparation

Feature Name	Source	Tools Used
Landslide Inventory	Geological Survey of India (GSI) – https://bhukosh.gsi.gov.in/Bhukosh/Public Landsat 8, Google Earth	ERDAS, ArcGIS
Topographical Features		
Slope (Degree), Elevation (m), Curvature	Shuttle Radar Topography Mission (SRTM) (30m spatial resolution)	ArcGIS
Hydrological Feature		
Distance To River (m)	Shuttle Radar Topography Mission (SRTM) (30m spatial resolution)	ArcGIS
Rainfall (mm/year)	Gridded Rainfall Data CRU (Climatic Research Unit) https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.05/cruts.2103051243.v4.05/pre/ IMD (India Meteorological Department) Pune - https://imdpune.gov.in/lrfindex.php (0.25° x 0.25°)	ArcGIS
Geological Feature		
NDVI	Landsat 8	ArcGIS
Landcover	Landsat 8, United States Geological Survey (USGS), Web Map Service (WMS)	ERDAS, ArcGIS
Soil	Food and Agriculture Organization (FAO)	ArcGIS

Feature Name	Source	Tools Used
Geology, Geomorphology, Distance to Lineaments (m)	Geological Survey of India (GSI) – https://bhukosh.gsi.gov.in/Bhukosh/ Public	ArcGIS
Anthropogenic Feature		
Road, Distance to Road (m)	Socioeconomic Data and Applications Centre (SEDAC) of NASA's "Earth Observing System Data and Information System (EOSDIS)," Google Earth	ERDAS, ArcGIS

The process of data collection and analysis was structured into five major categories, as outlined below:

Landslide Inventory: The initial step in data preparation involved compiling a landslide inventory, which serves as the foundation for landslide susceptibility modelling (LSM). This dataset contains detailed information on landslide occurrences and their spatial distribution within the study area as shown in Figure 3.5. A total of 1,561 landslide locations were identified, mapped using Google Earth, and further processed in QGIS to create a landslide inventory map. The landslide inventory was prepared by delineating the Area of Interest (AOI) in ERDAS Imagine using Landsat 8 imagery as shown in Figure 3.6, while simultaneously syncing it with Google Earth for enhanced spatial accuracy. The delineated AOI was then converted into a polygon shapefile, ensuring proper georeferencing. To facilitate a more detailed susceptibility analysis, random points were generated within each landslide polygon using the "Create Random Points" tool available in the Data Management Tools of ArcGIS. This process ensured a balanced and statistically significant dataset, essential for robust model training and validation.

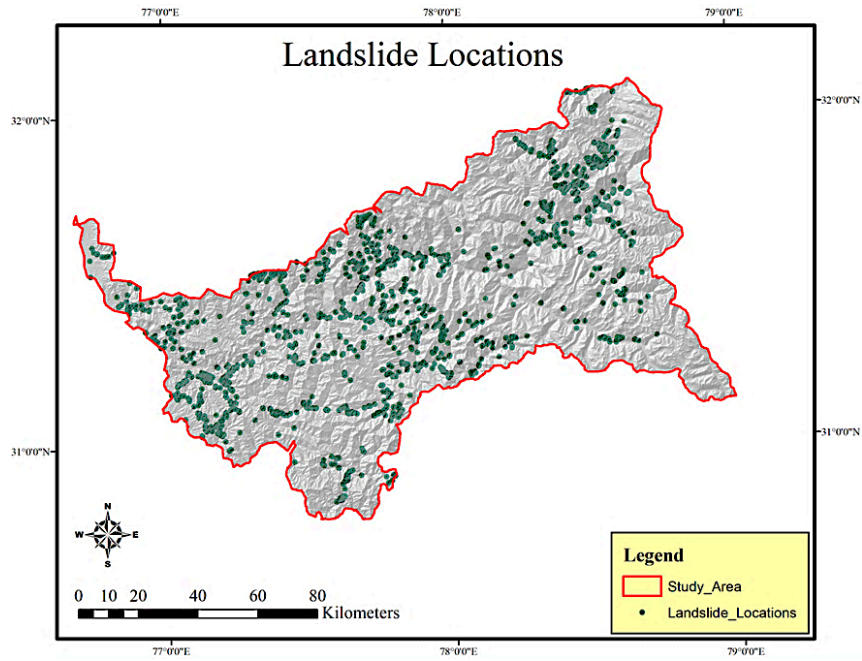


Figure 3.5: Landslide Locations in the Study Area



Figure 3.6: Preparation of Landslide Inventory

Topographical Features: The topographical characteristics of the study area were assessed using key terrain features such as slope, Aspect, elevation, and curvature, Topographic Wetness Index and Stream Power Index. The inclination of a slope serves as a primary control in landslide risk assessment, as it determines the equilibrium between destabilizing gravitational forces and the terrain's inherent resistance to failure. Steeper terrain features, whether convex (bulging outward) or concave (hollowed inward), show a stronger association with landslide activity. Areas with minimal curvature—those that are relatively flat or gently sloped—tend to exhibit greater slope stability. The range of curvature values serves as a key predictor of landslide susceptibility, where more pronounced deviations from zero curvature correspond to increased instability. Elevation plays a crucial role in landslide modelling, as it significantly impacts a range of geomorphological, hydrological, and environmental processes that can lead to slope instability. In landslide prediction, aspect reveals sun/wind exposure patterns, TWI identifies water accumulation zones, and SPI pinpoints erosion hotspots - together improving model accuracy.

These parameters were derived from Shuttle Radar Topography Mission (SRTM) data with a 30-meter spatial resolution, processed in ArcGIS. The extracted features offer valuable insights into the region's terrain morphology, which plays a crucial role in landslide susceptibility.

The study area exhibits significant elevation variations, with the lower reaches primarily consisting of valleys and river basins, while the upper regions are characterized by steep mountain slopes and ridges. To better visualize these variations, a Triangulated Irregular Network (TIN) model was generated, as shown in Figure 3.7. The elevation range spans from approximately 434 meters in the lower valleys to 6,540 meters in the higher mountainous regions, highlighting the stark contrast in terrain. This variation significantly influences slope stability, drainage patterns, and landslide occurrence. The TIN representation effectively captures the three-dimensional terrain structure, providing a high-resolution depiction of elevation distribution across the study area. Subsequently, Slope, Aspect, Elevation, Curvature, Topographic Wetness Index (TWI) and Stream Power Index (SPI) datasets were prepared from SRTM image in GIS platform. The maps for these features are shown in Figure 3.8.

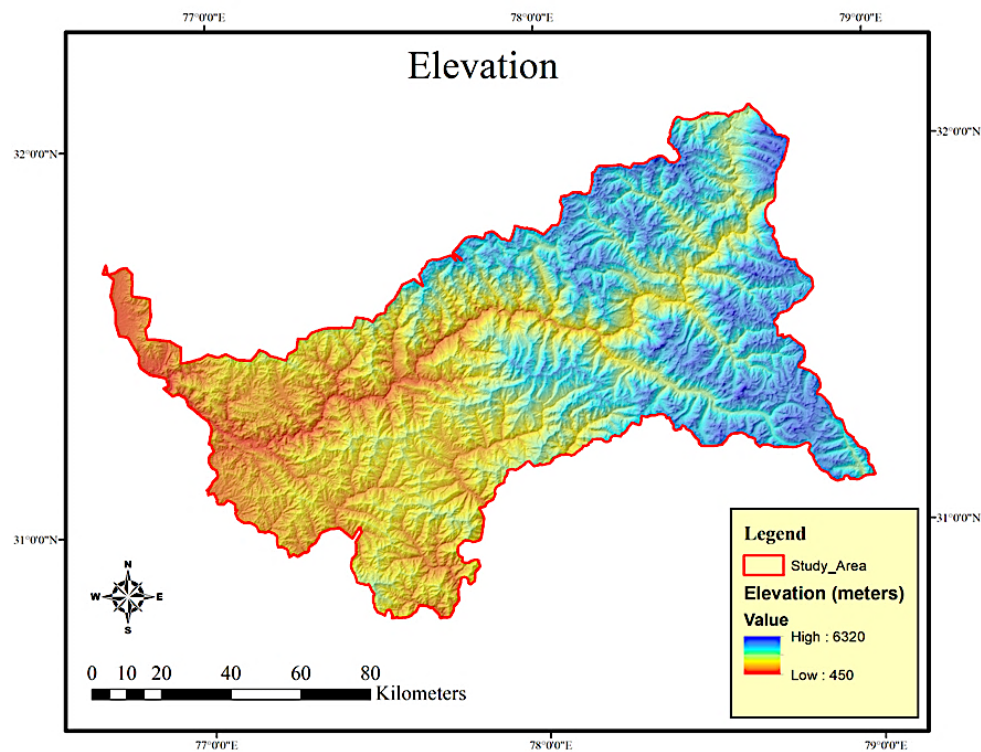
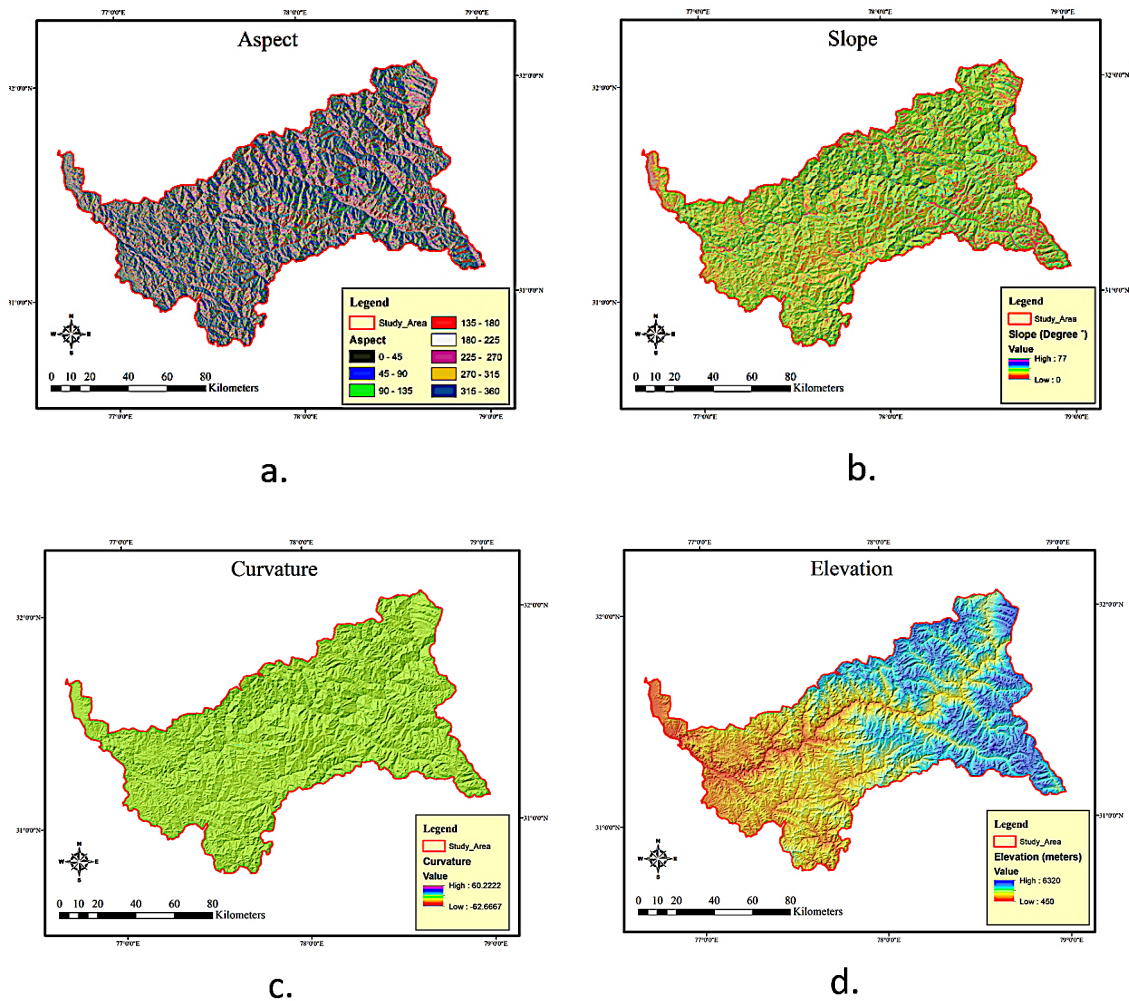


Figure 3.7: TIN Representation for Study Area



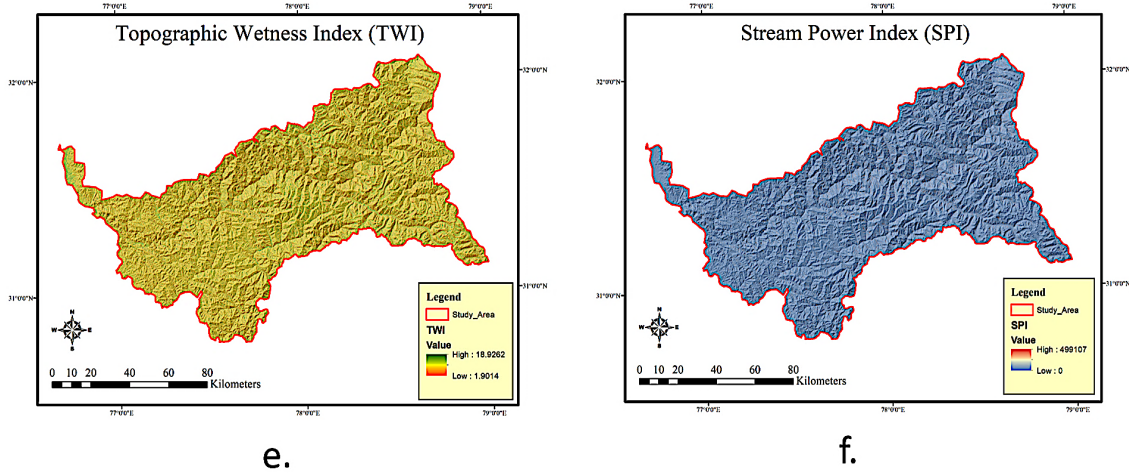


Figure 3.8: Topographical Features: (a) Aspect, (b) Slope, (c) Curvature, (d) Elevation(m), (e) TWI and (f) SPI

Hydrological Features: Hydrological conditions are major contributors to landslide susceptibility, as surface water flow influences soil saturation, erosion, and slope stability. One of the key hydrological parameters, distance to rivers, was derived from SRTM data (30m resolution) using ArcGIS. Understanding the spatial relationship between landslide-prone areas and nearby river systems is crucial for assessing the role of hydrology in triggering slope failures.

The drainage network of the study area, shown in Figure 3.9, was extracted using hydrological tools in ArcGIS. The region exhibits a well-defined dendritic drainage pattern, with major rivers and tributaries flowing through steep mountainous terrain. The presence of high-order streams in regions with rugged topography increases the risk of landslides due to constant fluvial erosion and changes in slope stability. Moreover, proximity to drainage channels influences groundwater seepage, leading to increased pore water pressure—a major factor in landslide initiation.

By analysing the distance of known landslide locations from river networks, a landslide susceptibility index based on hydrological influence was developed. This assessment helps in understanding how river proximity and water accumulation zones contribute to slope instability, making it a vital component in landslide susceptibility modelling. This dataset was further used to derived Distance to Stream layer. Further, annual gridded rainfall data of study area were extracted from Indian Meteorological Department website. Inverse Distance Weightage (IDW) method was used to derive spatial distribution of rainfall over the study area. Maps for Rainfall and Distance to Stream are shown in the Figure 3.10.

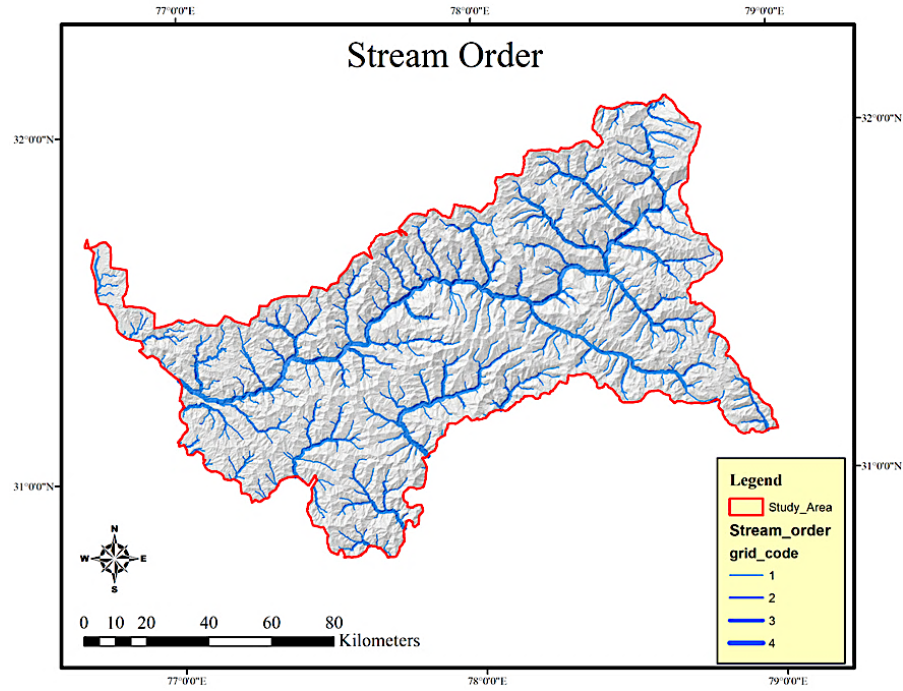


Figure 3.9: Drainage Map for the Study Area

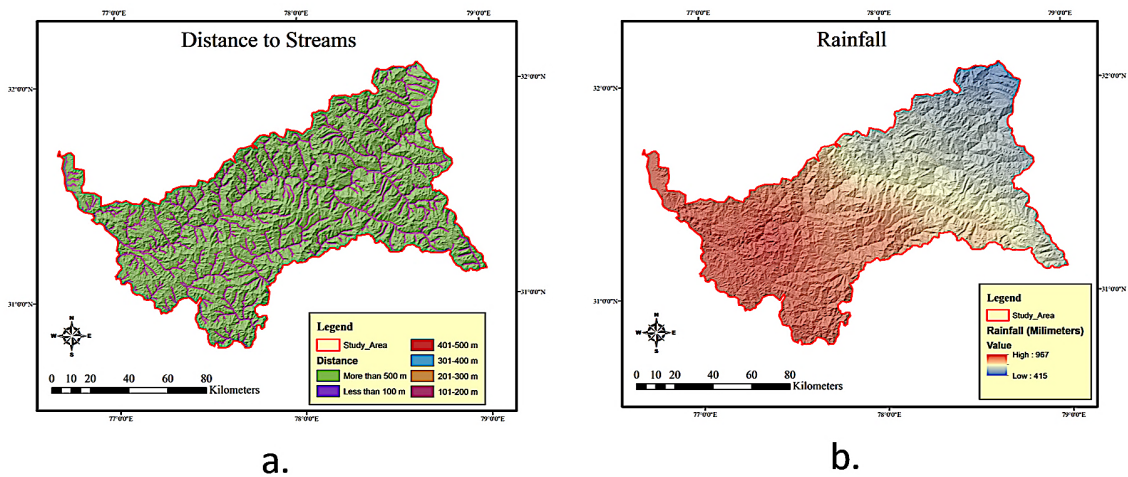


Figure 3.10: Hydrological Features: (a) Distance to Stream, (b) Rainfall

Geological Features: NDVI, soil type, distance to lineaments, geomorphology, and geology are essential factors in landslide modelling because they collectively influence vegetation cover, material strength, structural weaknesses, and the inherent characteristics of the landscape—all of which play a vital role in determining slope stability and landslide susceptibility. The geological characteristics of the study area were analysed using multiple remote sensing and GIS-based datasets. To assess vegetation cover and land-use patterns, NDVI (Normalized Difference Vegetation Index) and LULC (Land Use Land Cover) layers were generated using Landsat-8 imagery in ArcGIS and

ERDAS. Landsat-8 images with a 30 m spatial resolution, acquired from the USGS Earth Explorer platform for different dates were selected based on criteria ensuring minimal cloud cover and exclusion of winter imagery to obtain an accurate land cover classification. Figure 3.11 shows the False-Colour Composite (FCC) image of the study area derived from Landsat 8 Satellite Data..

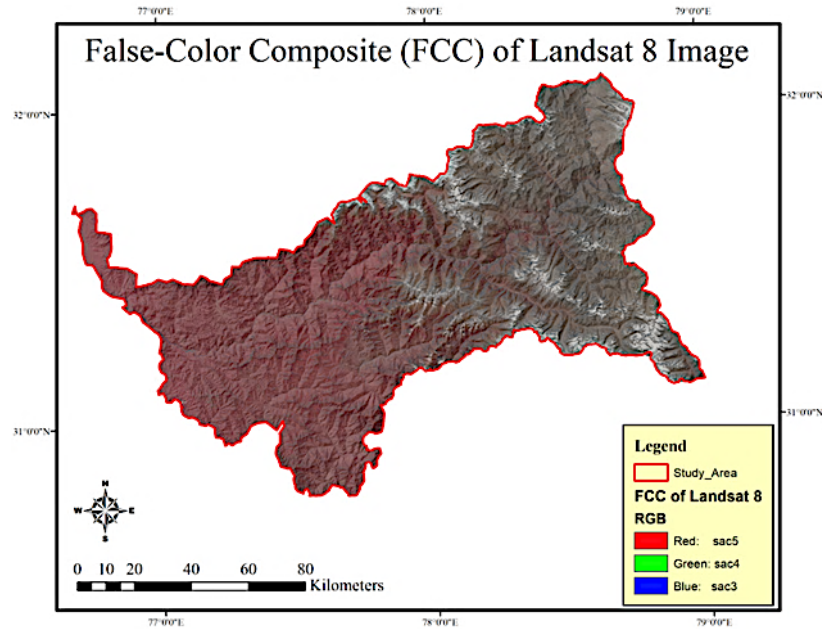


Figure 3.11: Satellite Image of the Study Area

The NDVI layer was derived using the Image Analysis tool in ArcGIS, leveraging the spectral properties of vegetation, where healthy vegetation reflects higher near-infrared (NIR) and absorbs more visible red light. Additionally, soil type, geology, and geomorphology data were collected from various sources and transformed into raster layers for further spatial analysis. These datasets provide crucial insights into the geological composition and structural stability of the terrain, aiding in understanding landslide susceptibility in the region.

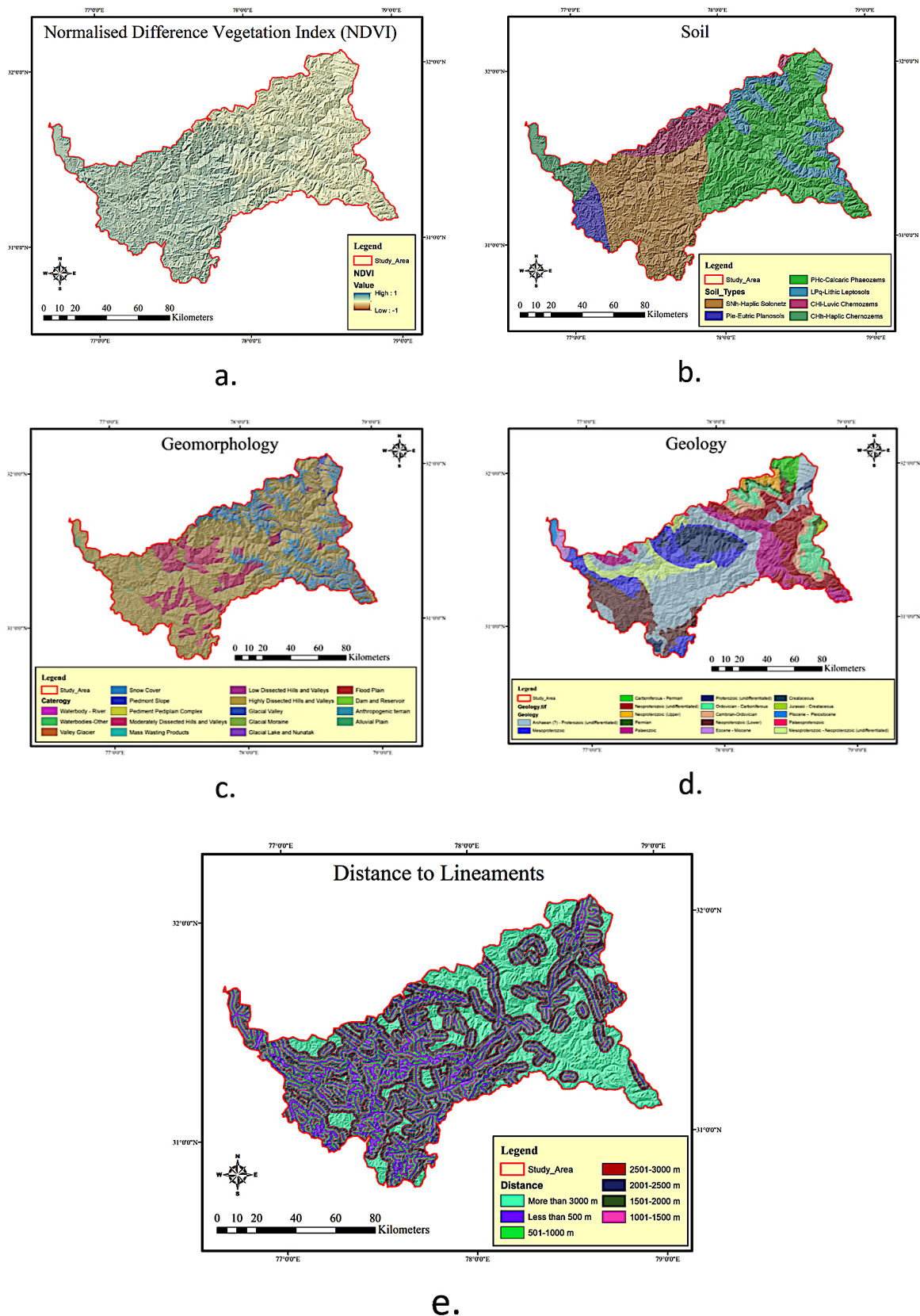


Figure 3.12: Geological Features: (a) NDVI, (b) Soil, (c) Geomorphology, (d) Geology, (e) Distance to Lineaments

The Figure 3.12 above shows the maps for NDVI, Soil, Distance to Lineaments, Geomorphology and Geology. The NDVI is calculated using the formula shown below;

$$\text{NDVI} = \frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}} \quad 3.1$$

To classify the satellite imagery, three widely recognized classification techniques were employed: Supervised Classification, Unsupervised Classification, and a Combined Approach. In the present work, Supervised classification technique has been used to prepare Landcover layer. In Supervised Classification, training samples for various land cover classes were manually defined, ensuring precise categorization based on spectral signatures.

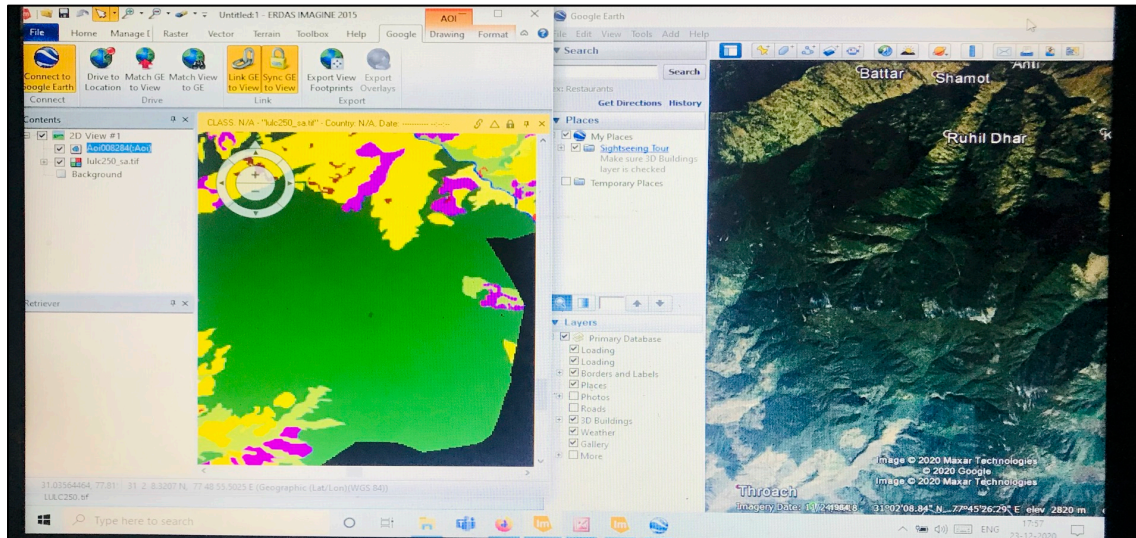


Figure 3.13: Validation of Area of Interest (AOI) for Supervised Classification

To enhance classification reliability, ERDAS IMAGINE was integrated with Google Earth for ground truthing as shown in Figure 3.13 above, and it enabled manual verification and refinement of classified regions. This approach helped in the accurate identification of various land cover types, and a signature file was created for precise classification. The NDVI map played a crucial role in distinguishing vegetation types, such as forests, grasslands, and agricultural areas. Additionally, USGS Earth Explorer and Web Map Service (WMS) land-use data served as reference datasets for generating the Land Use Land Cover (LULC) map as shown in Figure 3.14. Where necessary, pixel-level corrections were applied to ensure an improved classification output, enhancing the accuracy of spatial analysis for landslide susceptibility studies.

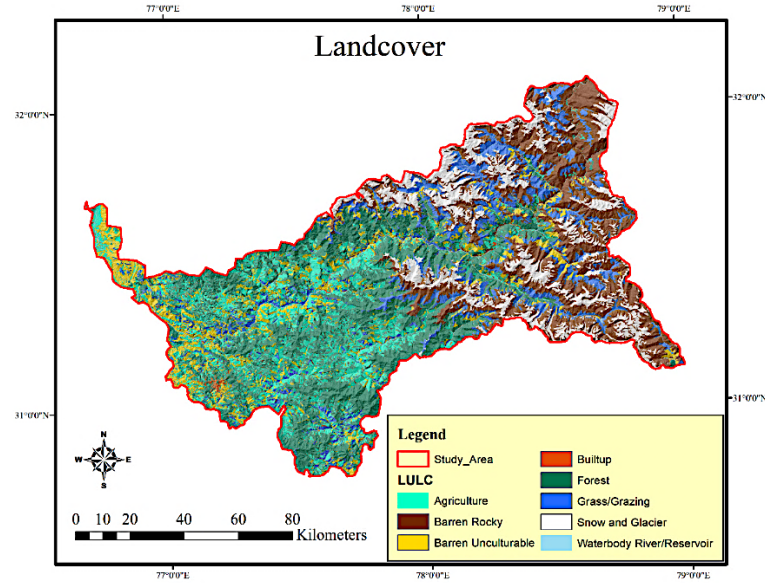


Figure 3.14: Landcover Map

Anthropogenic Features: Human-induced factors also contribute to landslide susceptibility. The existing road network was digitized using Google Earth and ArcGIS. Additionally, open-source road network data was obtained from NASA's Socioeconomic Data and Applications Centre (SEDAC) under the Earth Observing System Data and Information System (EOSDIS). This dataset was also incorporated to generate a distance-to-road layer as shown in Figure 3.15, helping assess the impact of road construction and human activities on landslide occurrence. By systematically integrating these diverse datasets, a comprehensive framework was developed for landslide susceptibility modelling, ensuring a robust assessment of potential hazard zones.

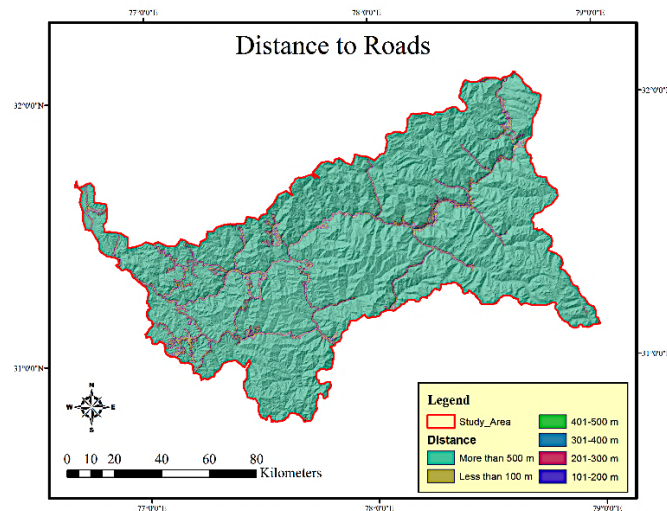


Figure 3.15: Anthropogenic Feature: Distance to Road

3.4 Data Statistics and Visualization

The landslide susceptibility analysis in this study is grounded on a well-structured, geospatially-referenced dataset comprising 3,118 instances 1,559 representing landslide-affected locations and 1,559 denoting stable, non-landslide locations. As part of the dataset preparation prior to machine learning analysis, two landslide locations were excluded from the original inventory because they fell within the buffer zone defined to represent safe (non-landslide) areas. This balanced sampling ensures parity across the dependent variable (landslide occurrence: yes/no), thereby mitigating the class imbalance often encountered in susceptibility modelling and enhancing the robustness of machine learning outcomes. The final independent feature set included 15 variables, comprising 07 continuous and 08 categorical predictors. Each data entry includes one dependent variable (target variable) of dichotomous nature and fifteen independent variables (causative features), carefully categorized into three data types:

- a. **Numerical (Continuous, float):** Curvature, Elevation, NDVI, Rainfall, Slope, Stream Power Index (SPI), and Topographic Wetness Index (TWI)
- b. **Categorical (Ordinal, object):** Aspect, Distance to Lineaments, Distance to Roads, Distance to Streams
- c. **Categorical (Nominal, object):** Geology, Geomorphology, Landcover, Soil

The dependent variable is a dichotomous nominal feature indicating landslide presence or absence. The dataset exhibits excellent integrity, with no missing values recorded across any variables, as shown in Figure 3.16, and Figure 3.17. This high level of completeness enables reliable statistical and predictive modelling.

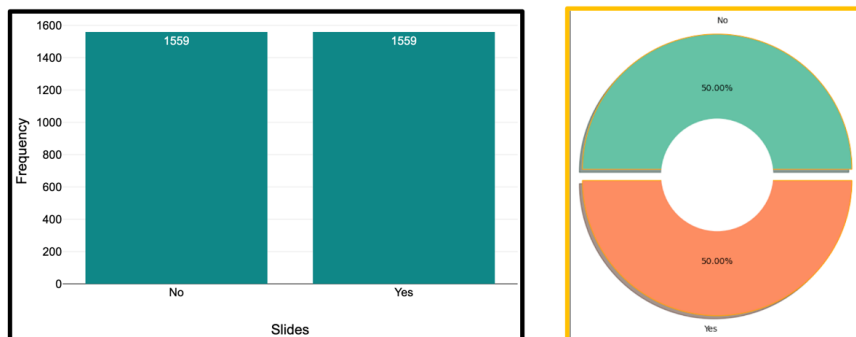


Figure 3.16: Dependent variables Statistics

Descriptive statistics—minimum, maximum, mean, median, standard deviation, and quartile ranges—were computed for all continuous features to assess central tendencies,

dispersion, and skewness. Additionally, pixel-level verification and minor corrections were conducted where automated classification exhibited anomalies, particularly in complex terrain patches. The statistical outputs formed the basis for correlating environmental characteristics with landslide distribution patterns.

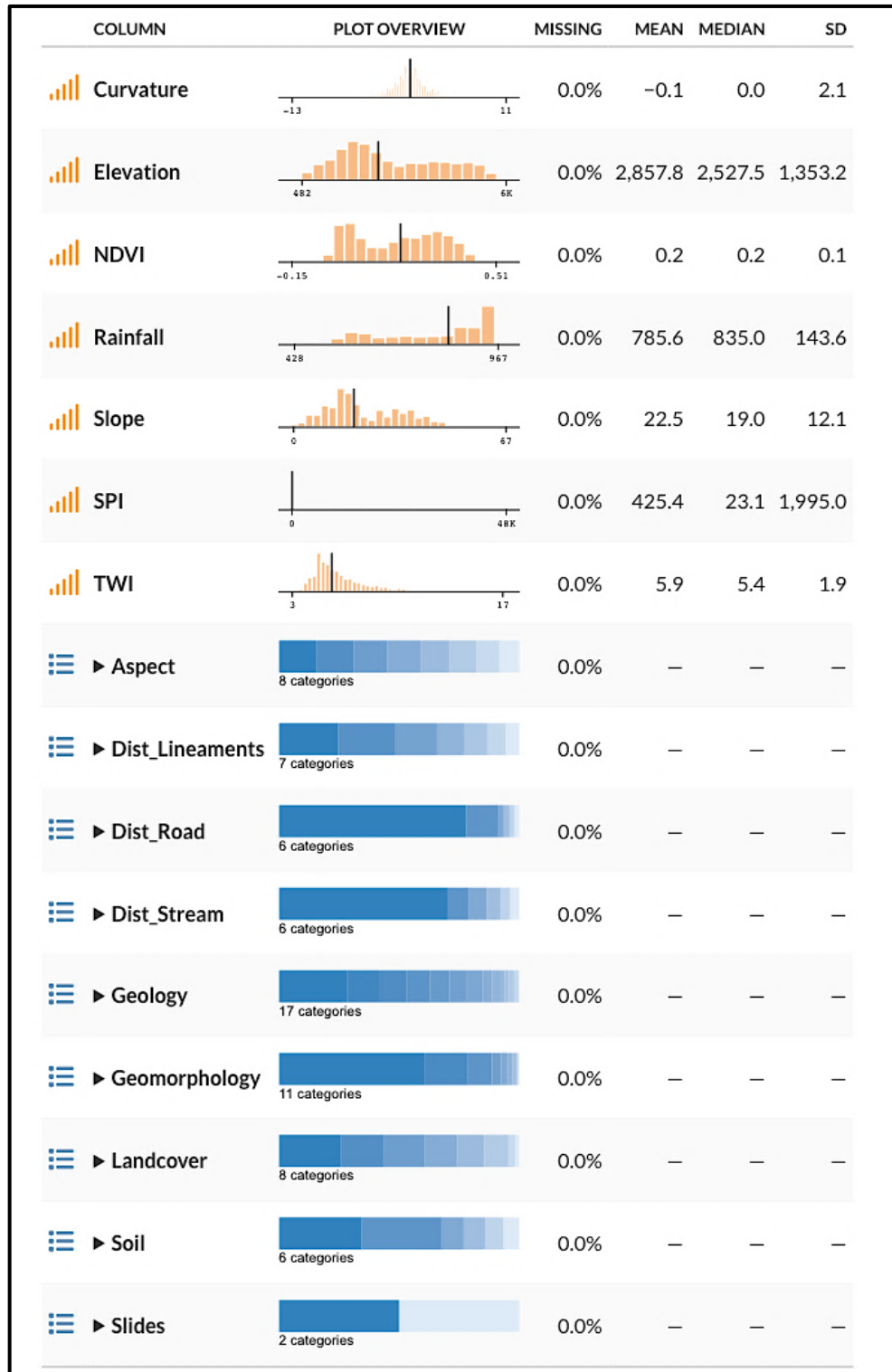


Figure 3.17: General Statistics for Variables

The continuous variables demonstrate distinct and meaningful patterns when stratified by landslide presence. These variations provide critical insights into topographic, hydrological, and vegetative influences on slope stability.

- **Curvature** displays a near-symmetric distribution (mean: -0.11, median: 0.0), with landslide zones concentrated in slightly concave regions (median: -0.1), in contrast to stable sites (median: 0.0). This suggests most stable areas are flat, but with some concave features as against consistently concave terrain in landslide prone regions. As can be seen in the range, it spans from -13 to +10.56, reflecting diverse slope profiles. The Figure 3.18 shows the graphical representation for the statistics.
- **Elevation** exhibits a right-skewed pattern (mean: 2,857.78m, median: 2,527.5m), with landslide-prone areas occurring predominantly at lower elevations (median: 2,257m) compared to non-landslide sites (median: 2,948m). This might be because of higher rainfall, human activities and fluvial erosion at lower elevations and stable bed-rock and reduced anthropogenic activities at higher elevations. The Figure 3.19 shows the graphical representation for the statistics.

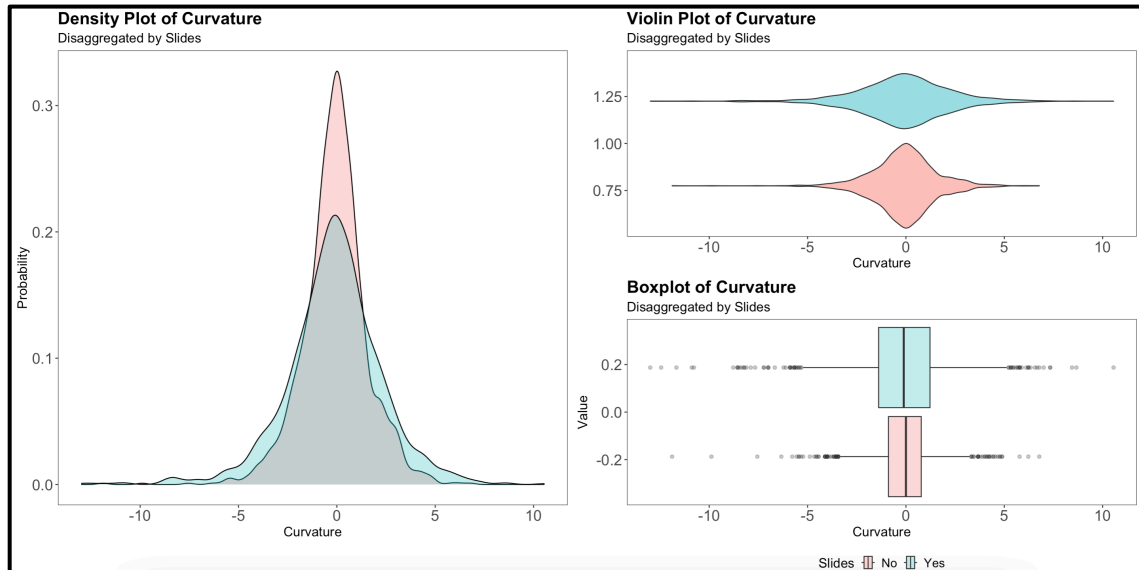


Figure 3.18: Density plot, Violin plot and Boxplot for Curvature with Landslides perspective

- **Slope** angles present the most significant separation, where landslide sites exhibit a notably steeper gradient (median: 32°) than their stable counterparts (median: 14°), confirming slope angle as a primary geomorphic control. Slope angle

emerges as one of the most discriminative factors in the dataset, showing clear separation between landslide and stable sites. This clear difference in the distribution of slope is due to the process by which non-landslide points were created. Based on the slope statistics at landslide locations, slope threshold was decided to delineate safe zones to generate random non-landslide points. The Figure 3.20 shows the graphical representation for the statistics.

- **Annual Rainfall** measurements reveal distinct patterns between unstable and stable slopes, with landslide-prone areas typically experiencing greater precipitation (median = 852mm, mean = 799mm) compared to stable zones (median = 795mm, mean = 772mm). While the absolute rainfall differences appear modest (~50-80mm), they represent statistically significant and practically important thresholds for landslide occurrence in the study area. The consistent left-skew in both distributions suggests that extreme rainfall events (right tail) may be particularly important for landslide initiation. The Figure 3.21 shows the graphical representation for the statistics.

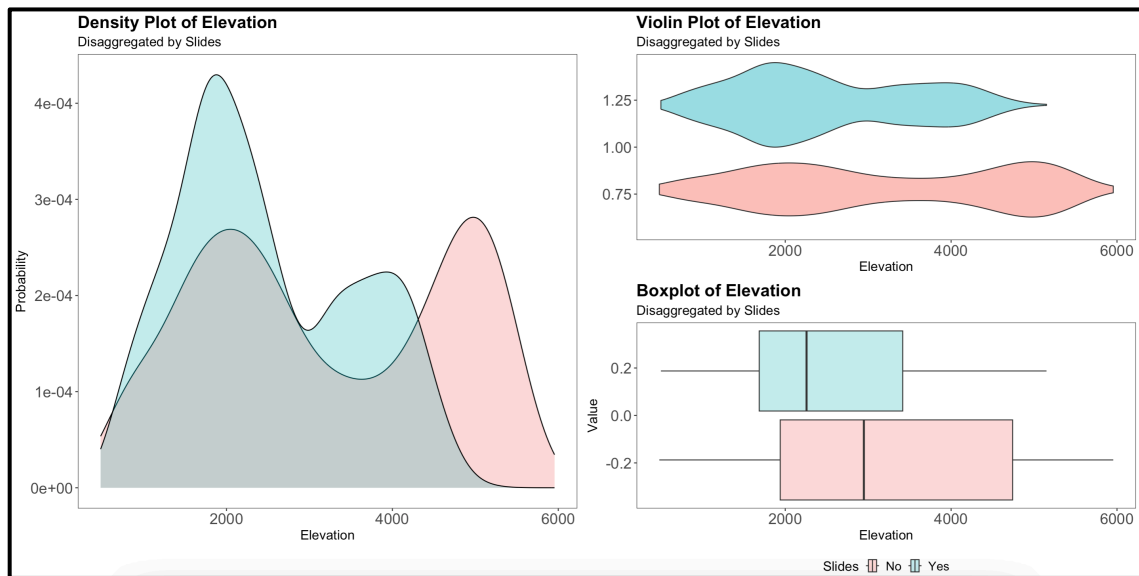


Figure 3.19: Density plot, Violin plot and Boxplot Elevation(m) with Landslides
Perspective

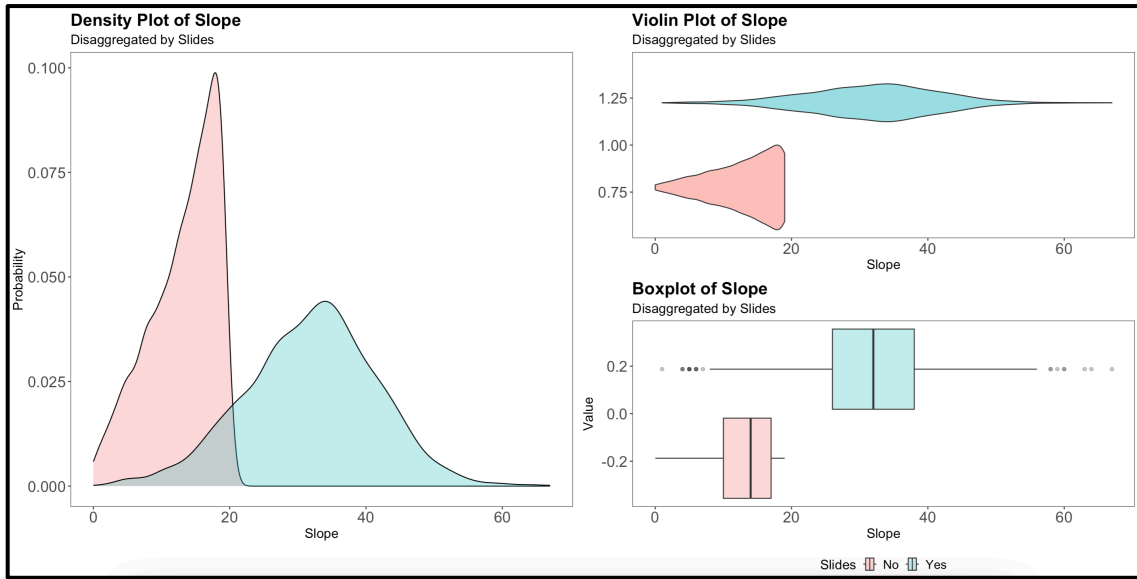


Figure 3.20: Density plot, Violin plot and Boxplot for Slope(Degree) with Landslides Perspective

- **TWI** values are lower in landslide-prone areas (median=5.0, mean=5.4) than in stable zones (median=5.8, mean=6.5), suggesting failures favour transient saturation over persistently wet conditions. The Figure 3.22 shows the graphical representation for the statistics.
- **SPI** analysis reveals that while landslides predominantly occur at moderate stream power levels (median=42.9), extreme global outliers in both landslide (mean=405.5) and non-landslide (mean=445.3) datasets represent critical geomorphic processes that must be retained in ML models to properly capture the full spectrum of slope stability conditions. Modern machine learning algorithms typically handle outliers internally during model training. Therefore, these outlier points were retained in the dataset to observe any potential impact on the overall model performance. The Figure 3.23 shows the graphical representation for the statistics.

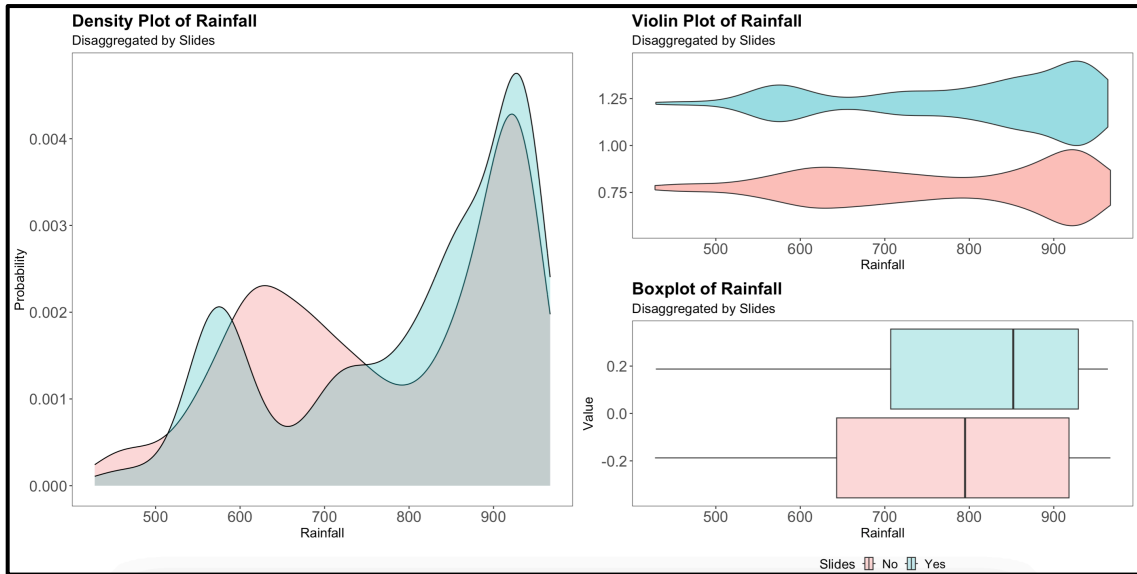


Figure 3.21: Density plot, Violin plot and Boxplot for Rainfall(mm/year) with Landslides Perspective

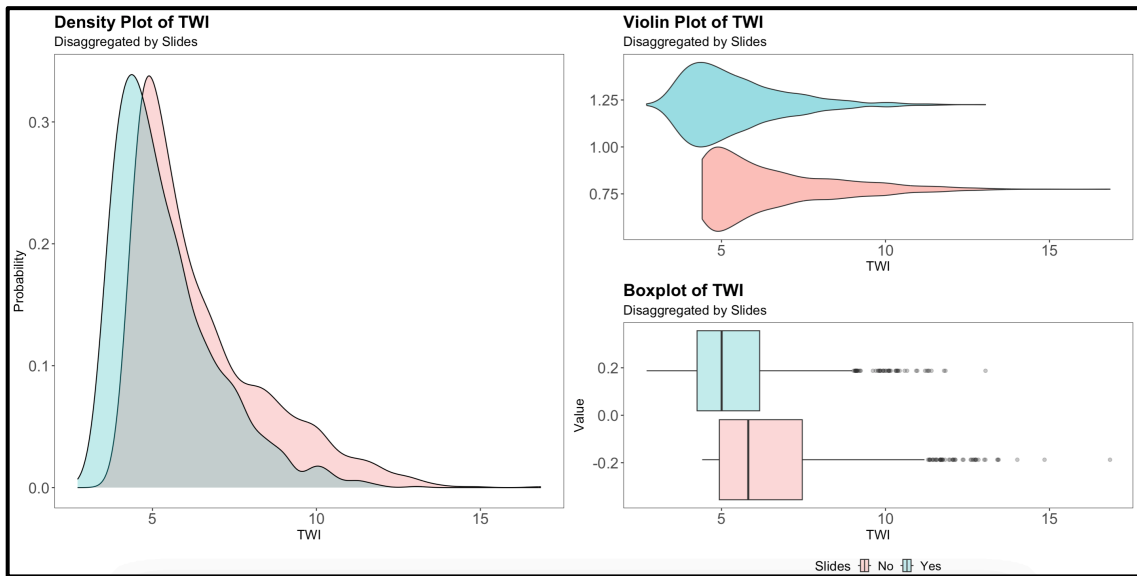


Figure 3.22: Density plot, Violin plot and Boxplot for TWI with Landslides Perspective

- **NDVI** ranges from barren to sparsely vegetated zones, with a mean of 0.19 ± 0.14 . Barren lands ($NDVI \leq -0.15$) are disproportionately associated with slope failures, reinforcing vegetation's stabilizing role. The Figure 3.24 shows the graphical representation for the statistics.

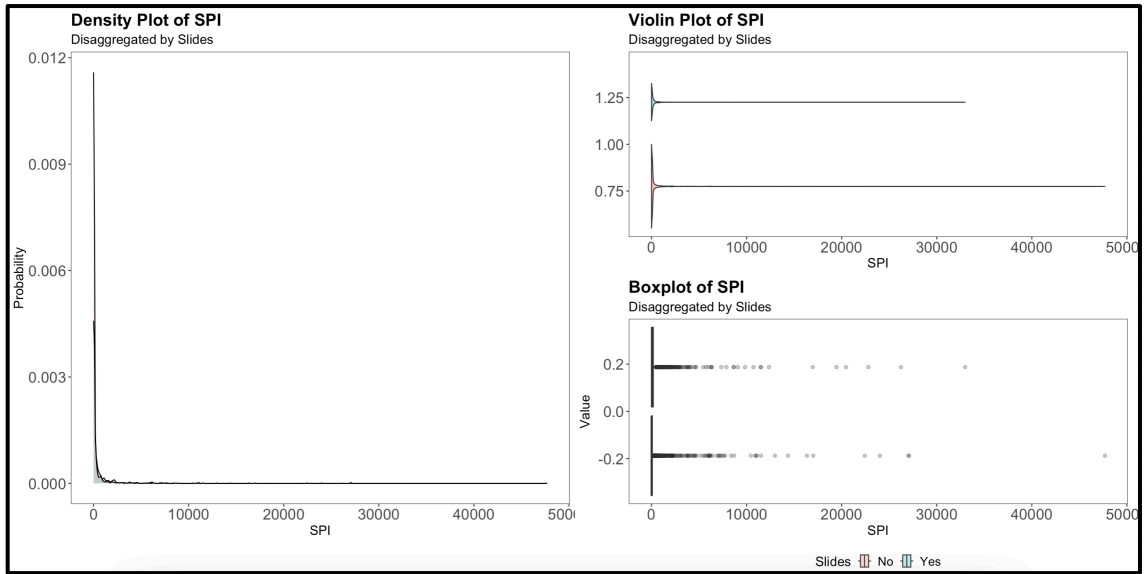


Figure 3.23: Density plot, Violin plot and Boxplot for SPI with Landslides Perspective

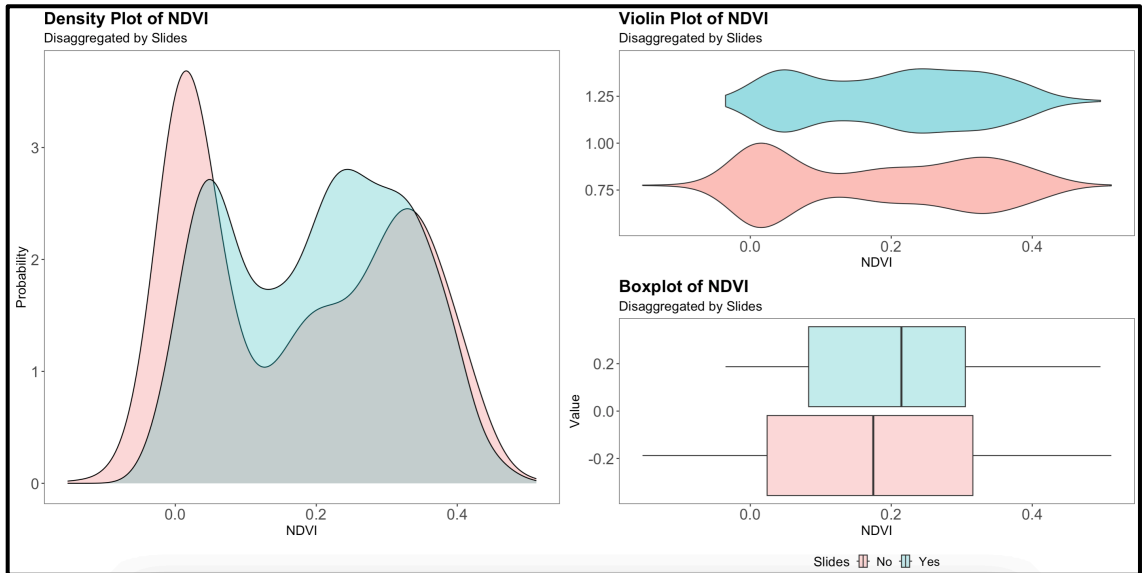


Figure 3.24: Density plot, Violin plot and Boxplot for NDVI with Landslides Perspective

The categorical parameters capture diverse landscape characteristics, with every class represented adequately in both landslide and non-landslide samples. The general statistics as shown in Figure 3.17 shows that there is no missing value in the data and integrity of dataset is ensured. Data distribution through Pie charts and Sankey diagram for ordinal and nominal variables is discussed.

Landcover class distribution with respect to landslide and non-landslide is represented through pie-charts and Sankey diagram which is shown in the Figure 3.25. The Sankey diagram representing landcover types captures the relationship between surface cover

characteristics and landslide activity. Landcover categories include forest, agriculture, barren rocky, grazing land, snow and glaciers, built-up areas, and waterbodies. These are connected to landslide outcomes (“Yes” or “No”), with balanced representation across both classes (50% each). Landcover analysis reveals Grass/grazing areas (18.82% vs 8.27%) and barren unculturable land (14.75% vs 8.27%) show substantially higher representation in landslide-affected zones, suggesting these are the most vulnerable landcover types. Snow and glacier areas demonstrate remarkable stability, constituting only 3.36% of landslide areas but 19.63% of non-landslide areas, making them the least landslide-prone terrain despite their potentially challenging topography. This indirectly confirms the higher rainfall values at lower altitudes of study area is very much influential in landslide occurrences. Forest areas appear slightly more susceptible to landslides (27.45% vs 24.31%), while agricultural lands show similar representation in both scenarios (16.46% vs 16.87%), indicating moderate vulnerability. Built-up areas and water bodies have minimal representation in landslide zones, suggesting either natural resilience or deliberate avoidance of hazardous locations for development.

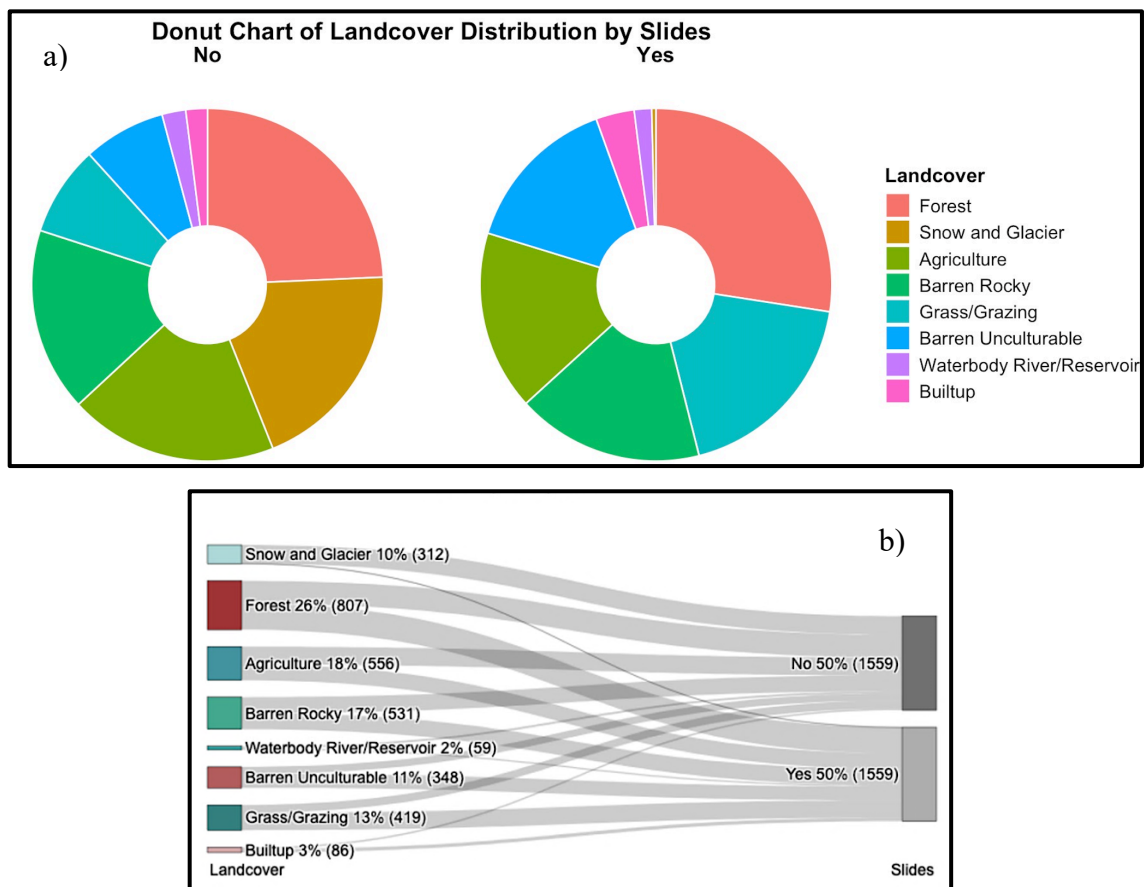


Figure 3.25: Visualization of Landcover Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams

Geological Formation class distribution with reference to landslide and non-landslide locations is shown in Figure 3.26.

Neoproterozoic (Lower) shows significantly higher presence in landslide areas (17.90% vs 11.48%), while Carboniferous-Permian formations exhibit nearly double representation in landslide zones (3.20% vs 1.66%). Quaternary deposits (12.70% vs 10.07%), Mesoproterozoic-Neoproterozoic undifferentiated (11.33% vs 8.11%), and Palaeozoic formations (9.92% vs 7.50%) all show moderate increases in landslide-prone areas. Neoproterozoic (undifferentiated) appears less vulnerable (5.24% vs 7.68%), suggesting greater stability. Archean-Proterozoic undifferentiated shows nearly identical distribution (20.35% vs 20.45%), indicating that other factors likely influence landslide occurrence in these geological settings. This distribution pattern demonstrates that specific geological formations significantly influence landslide susceptibility, with rock type, age, and structural characteristics playing important roles in slope stability.

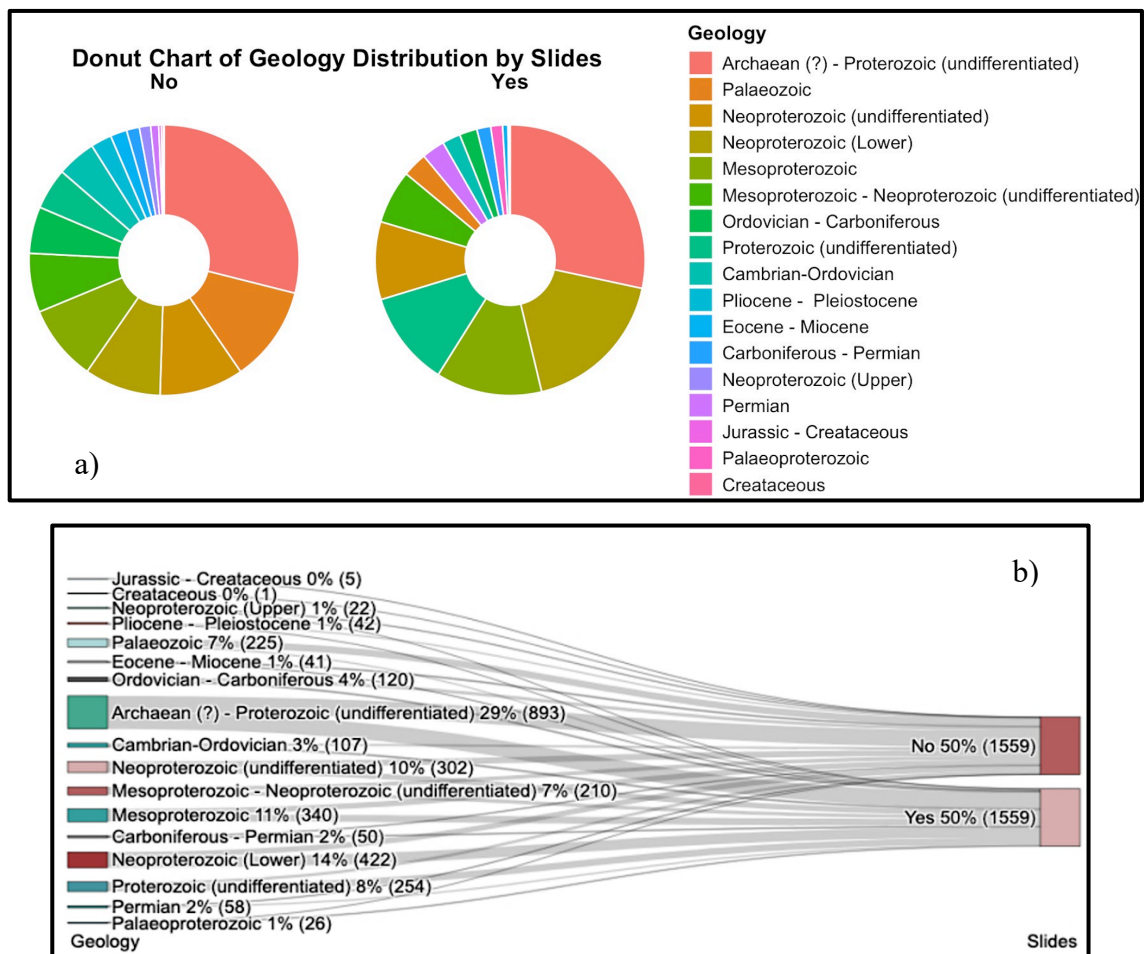


Figure 3.26: Visualization of Geological Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams

Geomorphology class distribution with reference to landslide occurrence is shown in Figure 3.27. Highly Dissected Hills and Valleys show a strong correlation with landslide occurrence, representing 71.46% of landslide areas compared to 49.65% of non-landslide areas. The concentration of landslides in dissected terrain (highly + moderately) is approximately 94% of all landslide occurrences, suggesting these landscape types are highly susceptible to slope failures. Areas with Snow Cover, Valley Glaciers, and Alluvial Plains show significantly lower landslide occurrence, likely due to their different morphological characteristics and slope conditions. This distribution indicates that terrain dissection (which typically creates steep slopes and complex topography) is a critical factor in landslide susceptibility in this area.

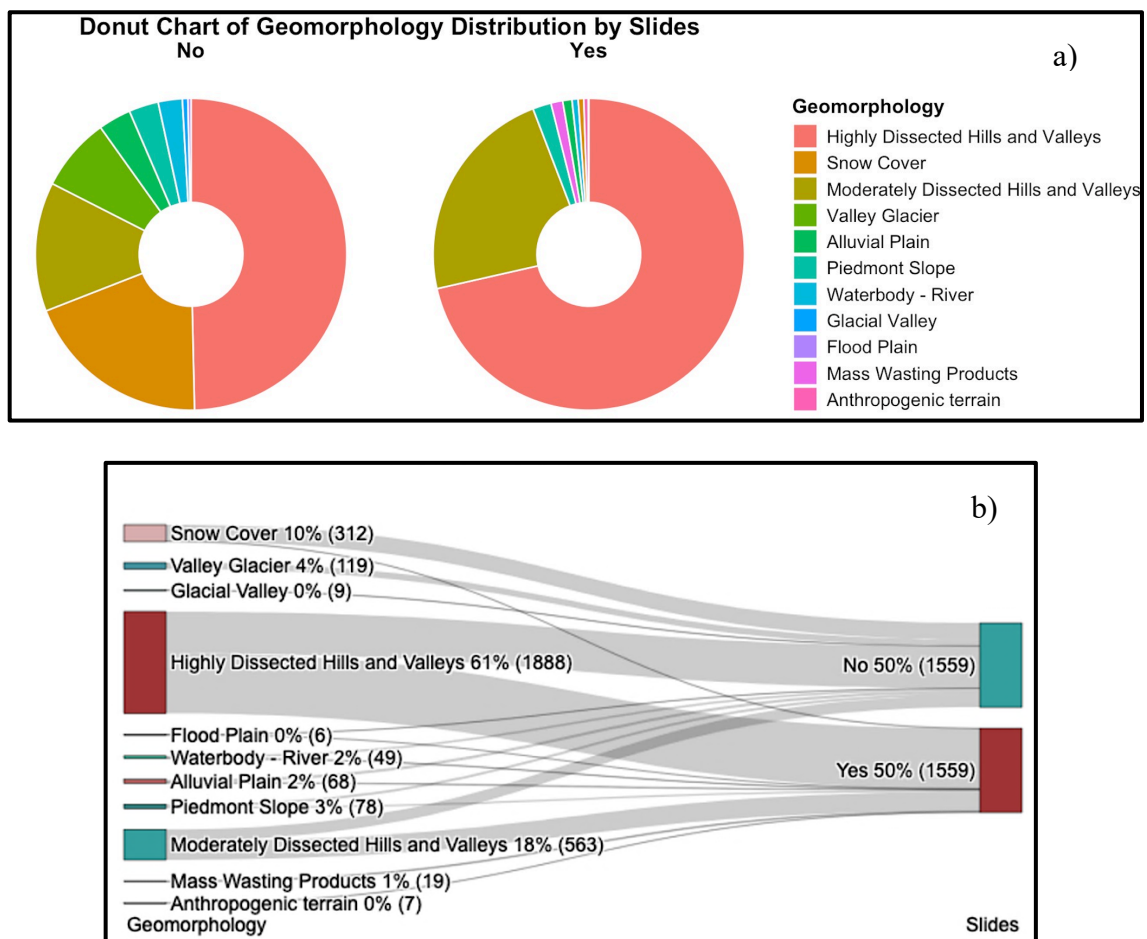


Figure 3.27: Visualization of Geomorphological Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams

Soil class distribution with reference to landslide occurrence is shown in Figure 3.28. The diagram for soil types illustrates how different pedological classes correlate with landslide occurrences. Categories such as SNh–Haplic Solonetz, PHc–Calcaric

Phaeozems, LPq–Lithic Leptosols, and CHh–Haplic Chernozems are linked to landslide classifications. Each category’s contribution is depicted by the width of its flow into “Yes” or “No” outcomes. Haplic Solonetz and Calcaric Phaeozems are the most common, together comprising 68% of the data. Although slope is not directly shown, it indirectly influences soil development and stability certain shallow or poorly drained soils (e.g., Leptosols) are typically found on steep terrain and are more prone to erosion and failure.

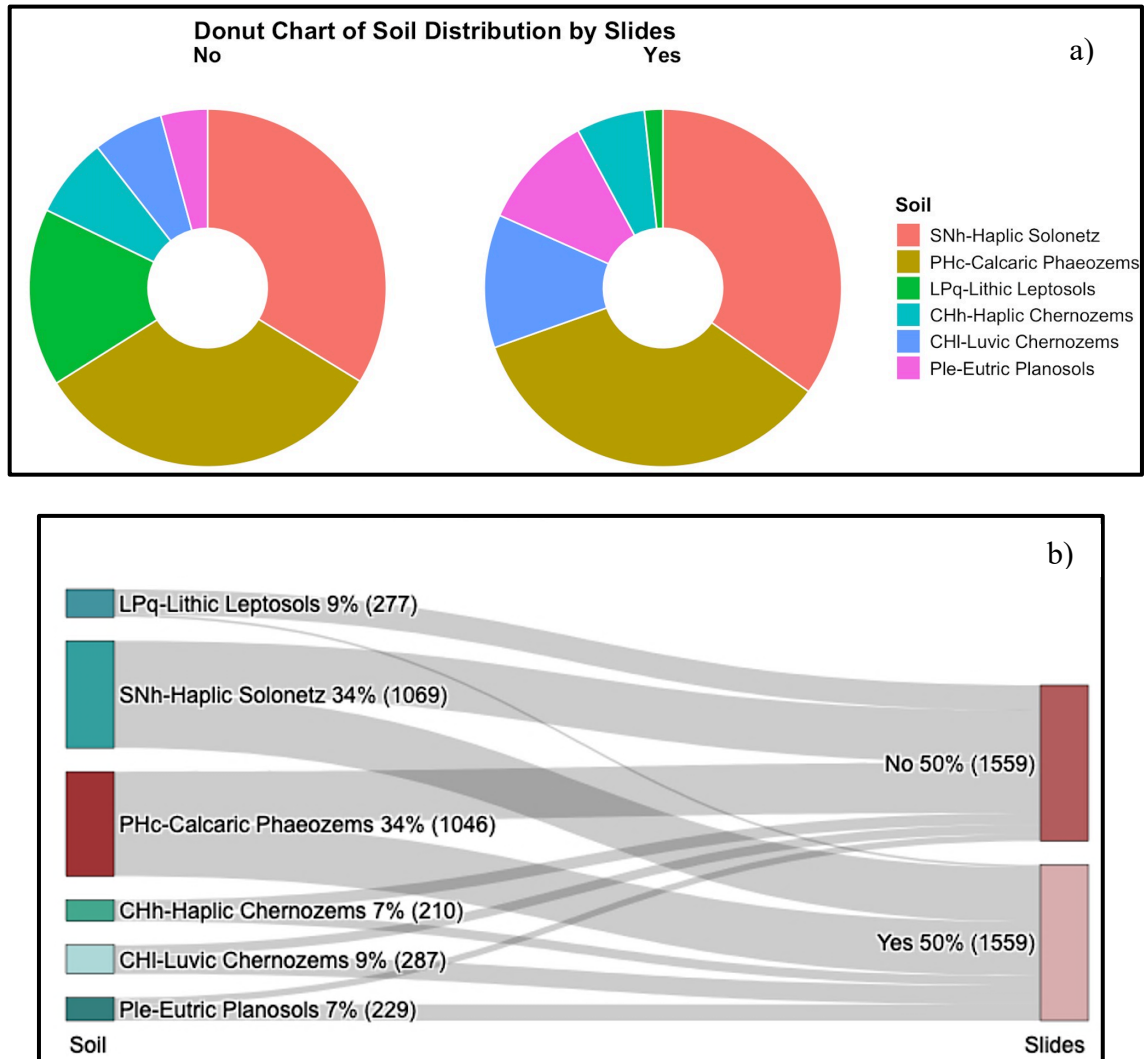


Figure 3.28: Visualization of Soil Category Distribution Across Landslide and Non-Landslide Groups Using a) Donut and b) Sankey Diagrams

Understanding the role of ordinal categorical variables is essential in evaluating terrain and anthropogenic factors that influence landslide susceptibility. These variables, which follow a natural order of classification, help reveal spatial and environmental patterns that contribute to slope instability. In this study, variables such as aspect, distance to lineaments, distance to roads, and distance to streams were analysed in relation to

landslide occurrences. Using a combination of grouped, stacked, and percentage-based bar plots, the distribution and impact of each class within these variables were assessed to identify the conditions most associated with landslide-prone zones. The following analysis provides insights into how these ordered variables correlate landslide events across the study area. The Figure 3.29 to Figure 3.32 shows the graphical representation for ordinal variables and their influence on the landslides.

- Aspect :** The relationship between slope aspect and landslide occurrence is analysed using four distinct bar plots in Figure 3.29. The first plot presents the overall frequency of various aspects such as North, South, Southeast, and others, indicating that the North-facing slopes are the most represented in the dataset. The second plot utilizes a stacked bar chart to differentiate between landslide presence ("Yes") and absence ("No"), where it becomes evident that although the North-facing slopes dominate in number, Southeast-facing slopes show a comparatively higher occurrence of landslides. This trend is further highlighted in the third, grouped bar plot, which places "Yes" and "No" counts side by side, reaffirming that the Southeast aspect experiences more landslides relative to its total count. The final plot, which presents data as percentages, confirms that the Southeast-facing slopes have the highest proportion of landslides when normalized against total samples per aspect class. Collectively, these plots reveal that slope orientation plays a notable role in influencing terrain vulnerability.

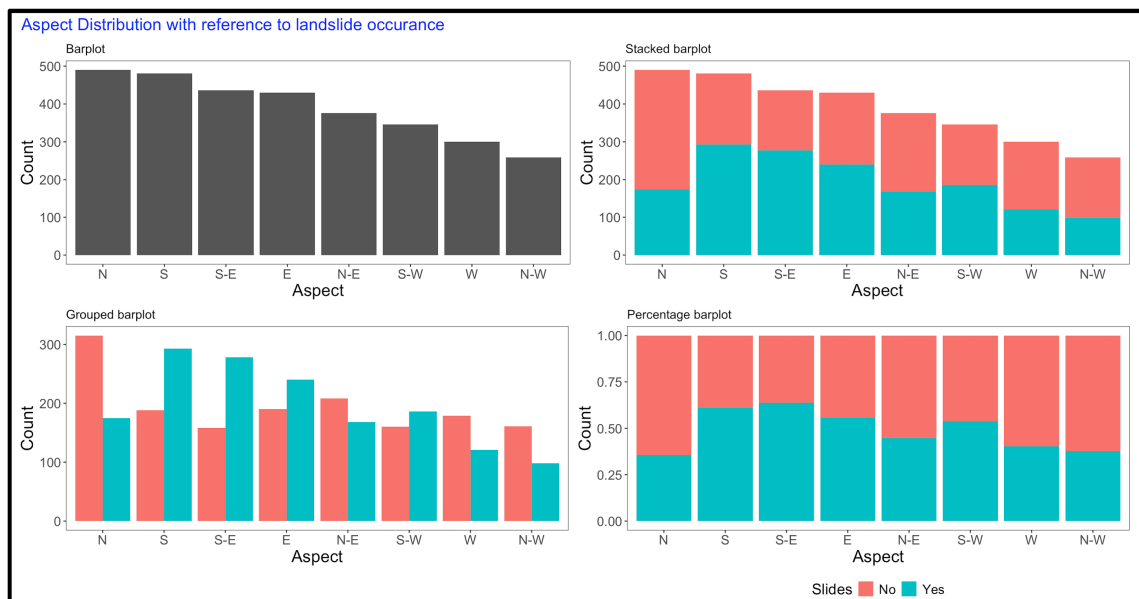


Figure 3.29: Aspect Distribution w.r.t Landslide occurrence

- Distance to Lineaments :** The distribution of landslide occurrences with respect to distance from lineaments is illustrated through a set of four bar plots in Figure 3.30. The first plot displays the overall count of data points across various buffer ranges, showing a general decline in frequency as distance from lineaments increases. The second plot, a stacked bar chart, separates landslide and non-landslide cases within each distance class, clearly showing that a significant number of landslides are concentrated in closer proximity to lineaments. The third plot uses a grouped bar format to allow side-by-side comparison of landslide occurrences in each distance range, emphasizing the disproportionately high landslide count in the “less than 500 m” category. The final plot presents normalized percentages, showing that the risk of landslides is notably higher in areas situated closer to lineaments. These visualizations confirm that proximity to geological lineaments is an important factor in slope failure susceptibility, likely due to the presence of fractures and structural weaknesses in the rock mass.

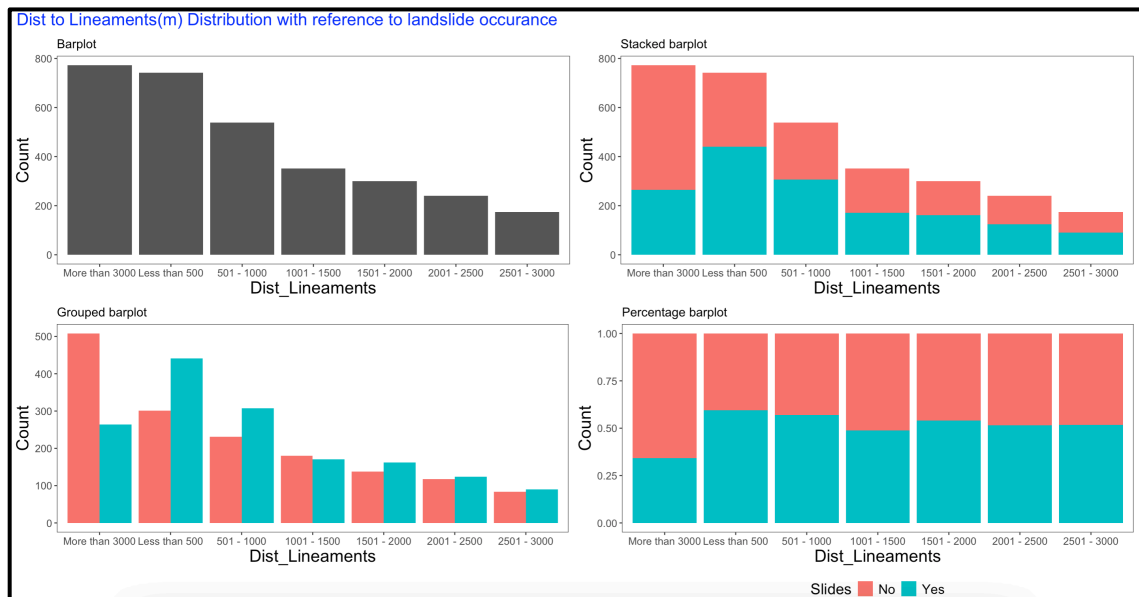


Figure 3.30: Distance to Lineaments(m) Distribution w.r.t Landslide occurrence

- Distance to Road :** The influence of road proximity on landslide occurrence is depicted using four bar plots in Figure 3.31. The first plot shows the distribution of samples across different distance classes, where areas “more than 500 meters” from roads constitute the highest frequency, followed by locations “less than 100 meters” away. The second plot, presented as a stacked bar chart, categorizes each distance group into landslide ("Yes") and non-landslide ("No") events. It reveals

that the majority of “more than 500 m” samples correspond to stable, non-landslide areas, while the “less than 100 m” category includes a notable share of landslide-prone points. The third plot, a grouped bar chart, visually separates the “Yes” and “No” responses for each class, further reinforcing that locations beyond 500 meters tend to be more stable. The final percentage-based chart indicates that areas within 100 meters of roads have a comparatively higher share of landslide events. These patterns suggest that while roads often pass through engineered or modified terrains, slopes in proximity may still experience disturbances, particularly in under-maintained or steep regions.

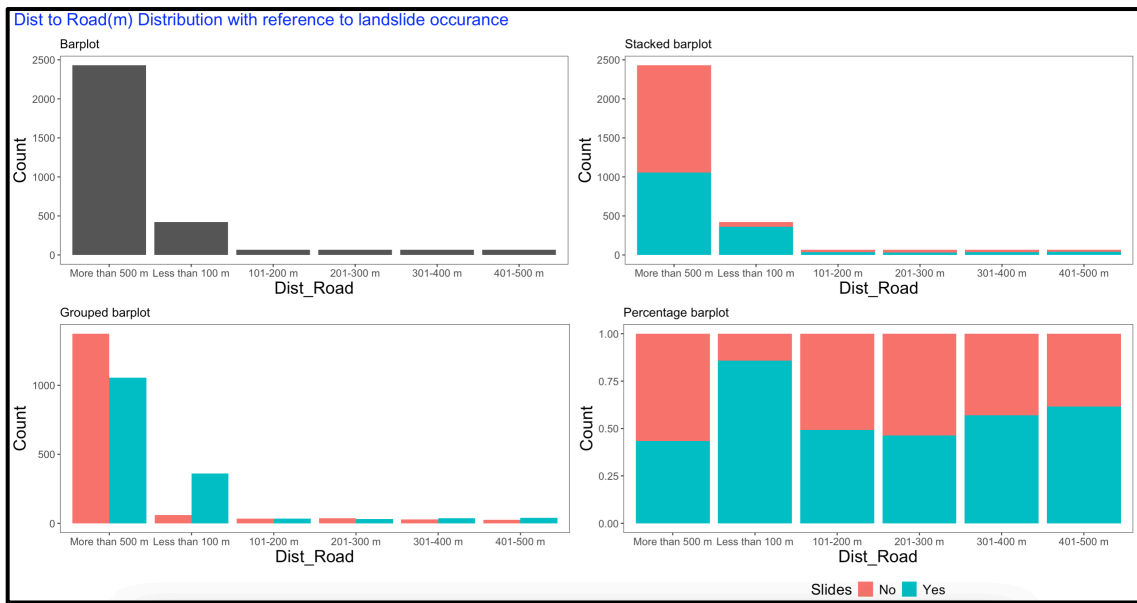


Figure 3.31: Distance to Road(m) Distribution w.r.t Landslide occurrence

- Distance to Streams : Proximity to drainage features and its relationship with landslide occurrence is explored through four bar plots as shown in Figure 3.32. The initial plot presents the distribution of samples across varying distance intervals from streams, with the “more than 500 meters” category being the most prevalent. The stacked bar chart that follows, differentiates landslide from non-landslide events within each class, showing that areas farther from streams have higher counts, predominantly in the stable (non-landslide) category. The grouped bar plot allows direct comparison between “Yes” and “No” across distance classes and further emphasizes this trend. However, the final percentage-based plot reveals that closer proximity to streams, particularly within 100 meters is associated with a relatively higher proportion of landslide occurrences, likely due

to enhanced soil saturation, toe erosion, or streambank undercutting. These findings underline the hydrological influence on slope stability and reinforce the importance of buffer zoning along active drainage lines.

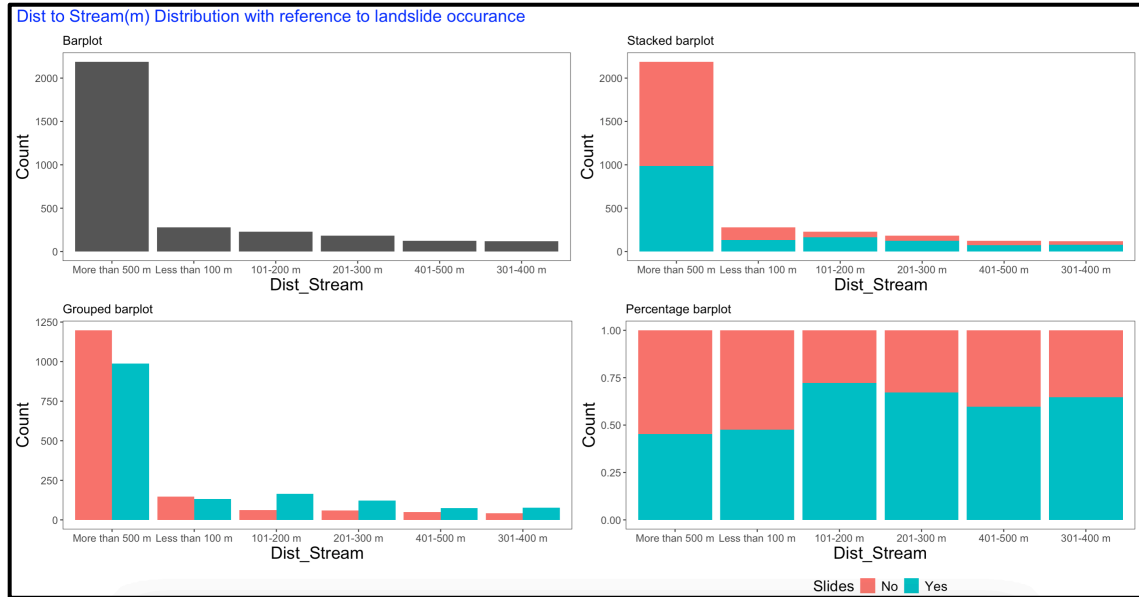


Figure 3.32: Distance to Streams(m) Distribution w.r.t Landslide occurrence

3.4.1 Multivariate analysis of Landslide Occurrence Based on Numerical Variables

To better understand the interplay between environmental and terrain-related factors in landslide susceptibility, scatter plots were utilized to visualize the relationship between multiple continuous variables. These plots incorporate combinations of elevation, rainfall, slope, and NDVI (Normalized Difference Vegetation Index) to identify spatial patterns associated with landslide events. Variables are represented through positional axes, point size, and colour, allowing for the simultaneous observation of key geospatial and climatic indicators.

The first scatter plot shown in Figure 3.33 explores landslide occurrences in relation to rainfall (x-axis), elevation (y-axis), and slope steepness (depicted by point size). The analysis shows that landslides tend to concentrate at elevations between 2000–4000 meters, with fewer events recorded above 4000 meters. Moderate rainfall zones (500–700 mm) show a higher density of landslide points, while larger dot sizes, representing steeper slopes, are predominantly linked with these events highlighting the critical role of gradient in slope failure. The second scatter plot as shown in Figure 3.34 integrates

NDVI (x-axis), elevation (y-axis), rainfall (point size), and slope (point colour) to provide a more nuanced understanding of environmental conditions. The results indicate that landslides are more frequent in areas with lower NDVI values, where vegetation is sparse or degraded. Like the first visualization, mid-elevation regions (below 4000 meters) and areas receiving high rainfall exhibit higher landslide density. Additionally, darker blue-coloured points, indicating steeper slopes, are closely associated with landslide locations, reinforcing the significance of slope gradient.

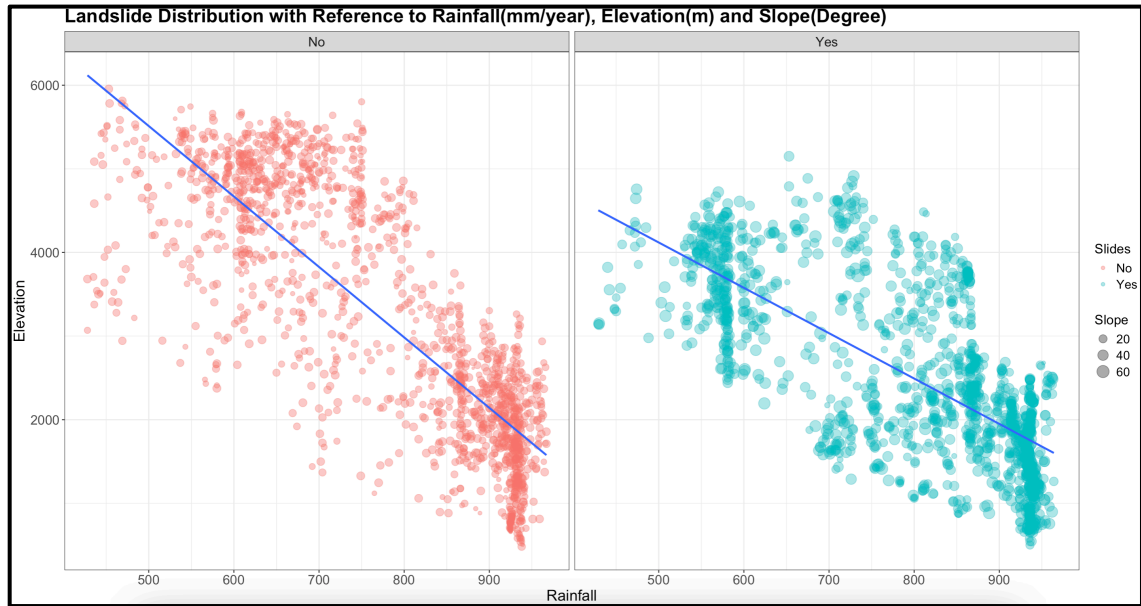


Figure 3.33: Landslide Distribution w.r.t to Elevation(m), Rainfall(mm/year) and Slope(Degree)

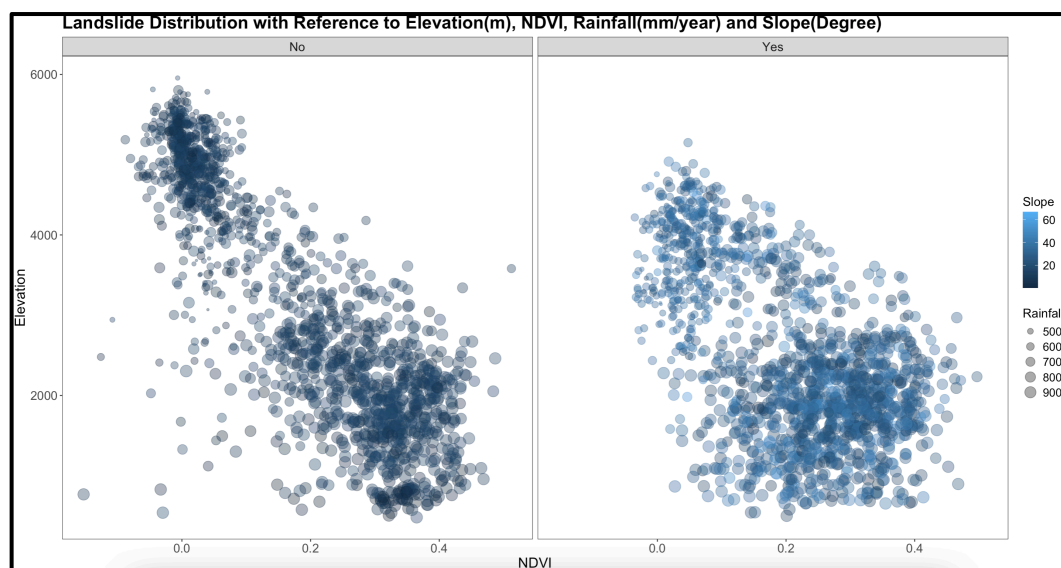


Figure 3.34: Landslide Distribution w.r.t Elevation(m), NDVI, Rainfall(mm/year) and Slope(Degree)

3.4.2 Elevation and Landcover Influence on Landslide Occurrence

This section examines the combined influence of elevation and landcover on landslide occurrences through a series of visual analyses, including density plots, boxplots, and strip plots. These graphical representations in Figure 3.35 provide a comprehensive understanding of how topographic variation and surface characteristics interact to affect slope stability.

A density plot comparing elevation distributions between landslide and non-landslide locations reveals that landslides tend to cluster around mid-elevation zones, particularly near 2000 meters, while non-landslide points are more broadly distributed across the elevation range. This suggests that certain elevation bands may present a higher susceptibility to landslides, likely due to a combination of terrain gradient, soil properties, and moisture accumulation.

When elevation is further disaggregated by landcover type, distinct patterns emerge. Landslides in agricultural and built-up areas are more frequent at lower elevations, likely influenced by anthropogenic activities such as land modification and reduced vegetation cover. In contrast, landcover types like barren rocky regions, forests, and snow/glacier zones exhibit higher landslide frequencies at elevations above 3000 meters, where natural slope conditions are more extreme. Waterbodies and grass/grazing lands also show landslides occurring predominantly at lower to mid-elevations, although their influence appears more localized.

Boxplots comparing elevation for landslide and non-landslide classes further reinforce these trends, with landslide-prone areas showing a slightly higher median elevation. When disaggregated by landcover, boxplots reveal that barren rocky areas and snow/glacier zones experience landslides at significantly higher altitudes, whereas agriculture and built-up zones experience more landslides at lower elevations. These differences highlight the complex interplay between natural topography and land use in driving slope instability.

Complementing these insights, a strip plot offers a detailed visual distribution of landslide and non-landslide points across landcover categories and elevation ranges. The plot shows a relatively balanced distribution of landslide and non-landslide samples across all landcover types, emphasizing that landslides are not restricted to a particular

terrain but are widespread. Elevation-specific patterns within each landcover class provide further depth to the analysis: snow and glacier zones are confined to elevations above 4000 meters, while agriculture and urban areas are primarily located below 2000 meters. Landcover types like forest and barren rocky are observed across a wider range of elevations, suggesting their presence in both low-lying and high-altitude regions.

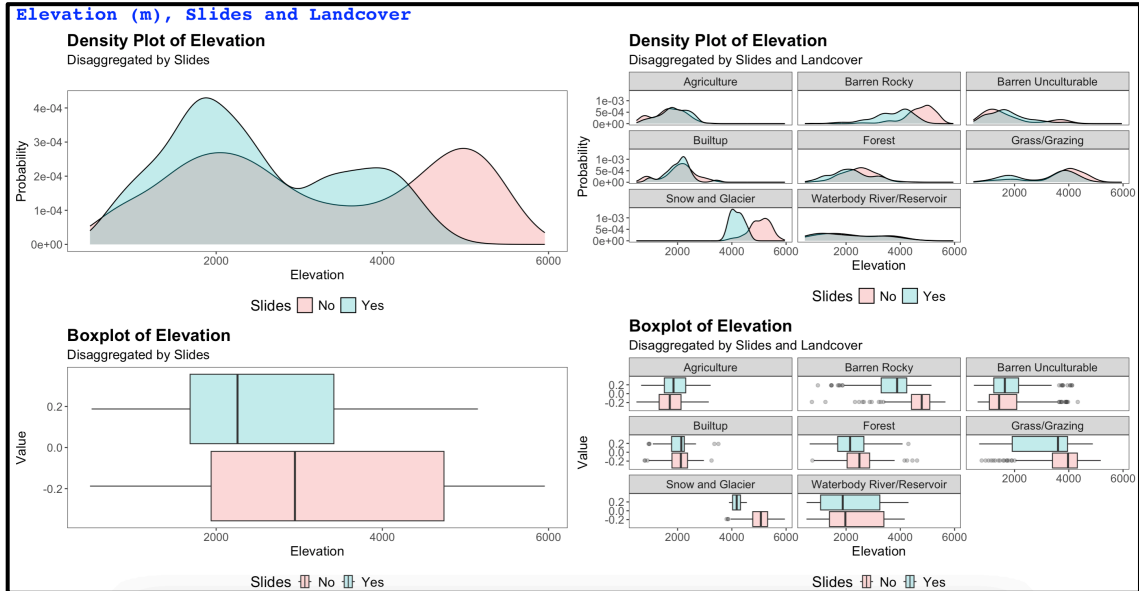


Figure 3.35: Landslide occurrence w.r.t Elevation(m) disaggregated by Landcover

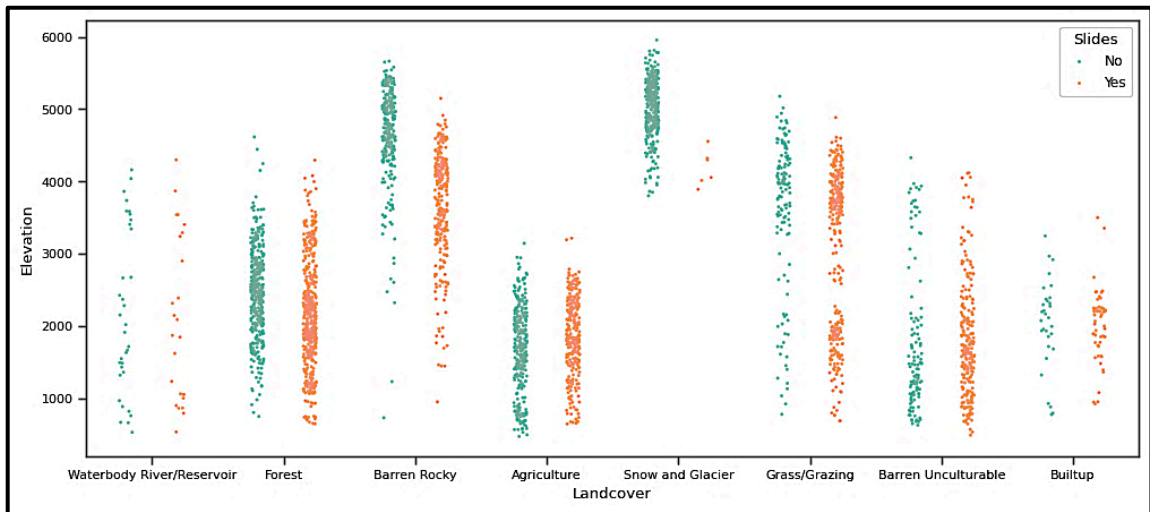


Figure 3.36: Landslide Occurrence w.r.t to Elevation(m) and Landcover

Notably, landslides are recorded across all elevation levels, but their frequency and impact vary depending on landcover type. At higher elevations, landslides are more prevalent in snow-covered and rocky terrains, which are often characterized by steep slopes and unstable geological conditions. At lower elevations, landslides in agricultural

and built-up areas are likely driven by human activities, poor drainage, and land degradation. These findings underline the importance of considering both elevation and landcover jointly when assessing landslide risk, as the interaction between these variables provides a more accurate representation of terrain vulnerability. This integrated understanding is crucial for developing targeted mitigation strategies and planning land use in landslide-prone regions.

3.5 Feature Selection

Feature selection is a fundamental preprocessing step in data-driven modelling, aimed at identifying the most informative predictors while eliminating irrelevant, redundant, or low-utility variables. The process serves multiple purposes: it improves model accuracy by reducing noise, prevents overfitting, enhances computational efficiency, and increases model interpretability by isolating key contributing factors. Importantly, it also mitigates the curse of dimensionality, a scenario where excessive input variables degrade predictive performance despite the availability of sufficient data.

This study adopts a two-stage feature selection approach, comprising unsupervised and supervised methods. The unsupervised stage evaluates features based solely on their internal statistical properties, independent of the target variable (landslide occurrence). This helps detect potential issues such as multicollinearity, low variance, and redundant variables early in the pipeline. The supervised stage, in contrast, measures the predictive power of variables in relation to the dependent variable and determines their contribution using model-driven techniques. The unsupervised feature selection process includes four major steps. First it was made sure that there are no incomplete or null features. After ensuring data completeness, it was followed with the Multicollinearity analysis, low variance filtering and correlation analysis. The three major steps are discussed below.

1. **Completeness of Dataset** : From the general statistics discussed in preceding section it is confirmed that there is no missing data in the selected dataset. The Figure 3.17 in preceding Section 3.3 shows the general statistics.
2. **Multicollinearity Analysis** : Multicollinearity refers to the presence of high linear correlation among independent variables, which can distort model coefficients,

inflate variance, and reduce the interpretability of parametric models. While some machine learning algorithms are robust to multicollinearity, it remains a critical consideration in feature importance estimation and regression-based analyses. To detect multicollinearity, the Variance Inflation Factor (VIF) was computed for continuous variables. VIF (Variance Inflation Factor) is derived from the Equation 3.2 and value obtained by regressing a predictor variable against all other independent variables in the dataset. This measure helps in quantifying multicollinearity, i.e., the extent to which one variable is linearly predictable from the others. While correlation matrices and scatter plots can also be employed to detect multicollinearity, they are limited to revealing only the bivariate relationships between variables. In contrast, VIF is preferred because it captures the combined linear association of a variable with a group of other predictors, offering a more holistic view of interdependence.

$$VIF = \frac{1}{1-R^2} \quad 3.2$$

$$R^2 = 1 - \frac{\text{Sum squared Regression (SSR)}}{\text{Total Sum of Squares (SST)}} \quad 3.3$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad 3.4$$

In above equations, y shows the observed values and $\hat{y}$ shows the predicted values from the regression model. For categorical variables, Generalized VIF (GVIF) was used, adjusting for degrees of freedom.

$$GVIF = VIF^{\frac{1}{2.DF}} \quad 3.5$$

To assess multicollinearity, standard VIF interpretation thresholds were applied. A VIF value below 5 was considered acceptable, values between 5 and 10 indicated moderate multicollinearity that warrants caution, and values above 10 were deemed critical, suggesting the need for corrective action. In this study, all variables demonstrated VIF or GVIF values below 5, confirming the absence of severe multicollinearity among the predictors. The highest VIF observed was 4.3 for Elevation, while Curvature exhibited the lowest value at 1.1, indicating a high degree of independence from other features. These results suggest that the dataset

is suitably conditioned for further analysis without the need to eliminate any variables at this stage. The Figure 3.37 shows the GVIF for all the variables.

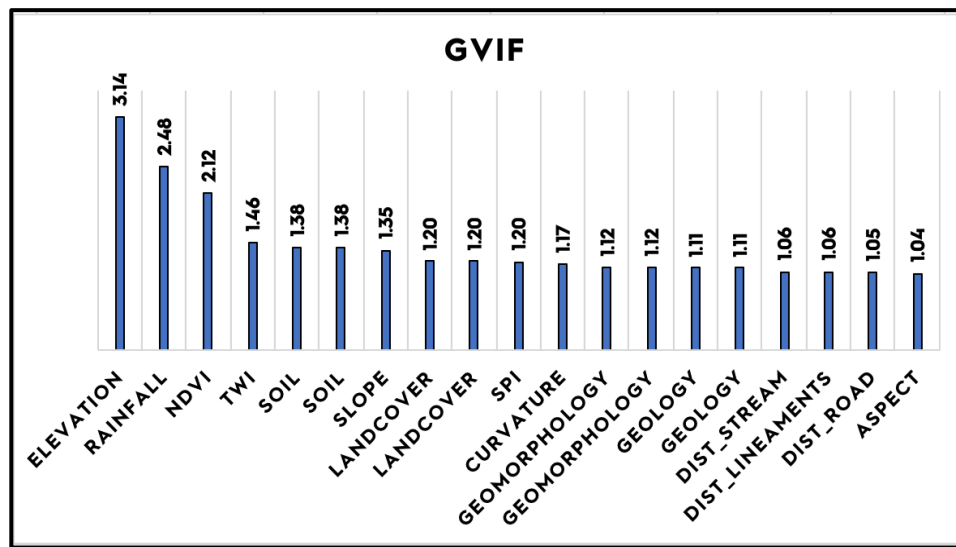


Figure 3.37: GVIF of the Features

3. Variance Threshold : Low-variance features often carry minimal discriminatory information and can introduce noise into machine learning models, thereby reducing model efficiency. To identify such features, a variance threshold of = 0.00 was applied. This step involved evaluating the variance of several numeric variables including Curvature, Elevation, NDVI, Rainfall, Slope, SPI, and TWI.

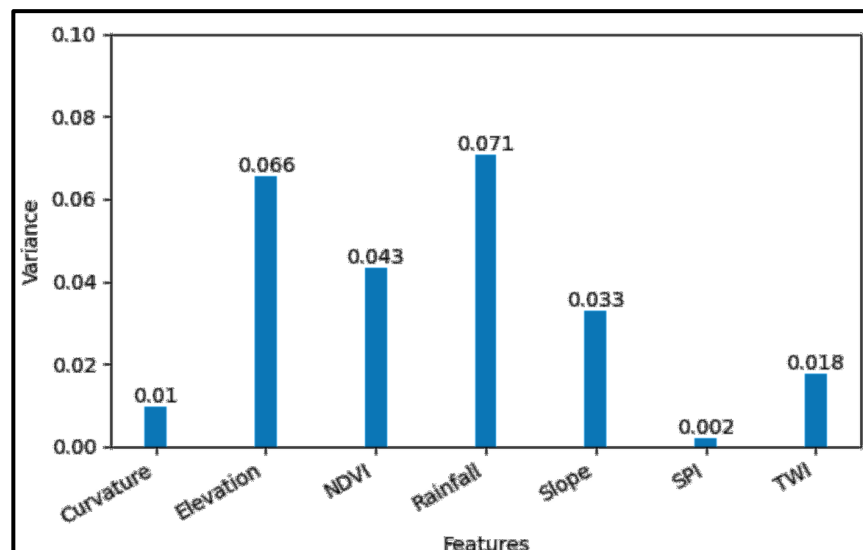


Figure 3.38: Variance of the Features

The mathematical foundation for this assessment lies in the variance formula, which quantifies the average squared deviation from the mean. Among these, SPI, Curvature, and TWI exhibited the lowest variance values, ranging between 0.002 and 0.018. While these values were relatively low compared to other features, none of them fell below the defined threshold. Additionally, no constant-valued features were observed in the dataset. As a result, no variables were eliminated at this stage, and all numeric features were retained for further analysis.

4. Correlation analysis was conducted to detect strong pairwise associations among numerical features that could introduce redundancy into the model. Given the nature of environmental data, Spearman's Rho (r_s) and Kendall's Tau (τ) were employed, as these non-parametric measures are well-suited for datasets that may not follow a normal distribution and are robust to outliers. Spearman's Rho evaluates monotonic relationships and is computed using the formula;

$$r_s = \frac{6 * \sum d_i^2}{n * (n^2 - 1)} \quad 3.6$$

where d_i denotes the difference in ranks of the paired observations, and n is the number of observations.

Kendall's Tau, on the other hand, measures ordinal associations and is defined as;

$$\tau = \frac{C - D}{C + D} \quad 3.7$$

where C and D represent the number of concordant and discordant pairs, respectively. These techniques enabled the identification of meaningful inter-variable relationships without being influenced by extreme values. The analysis revealed a strong negative correlation between Elevation and both Rainfall and NDVI, suggesting that higher altitudes tend to experience lower precipitation and vegetation density. Conversely, a strong positive correlation was observed between SPI and TWI, indicating their shared hydrological significance, which is particularly relevant for modelling landslide susceptibility. These findings affirmed the absence of critically redundant variable pairs (i.e., no correlation coefficient exceeding 0.8), thereby supporting the retention of all numeric features for further modelling.

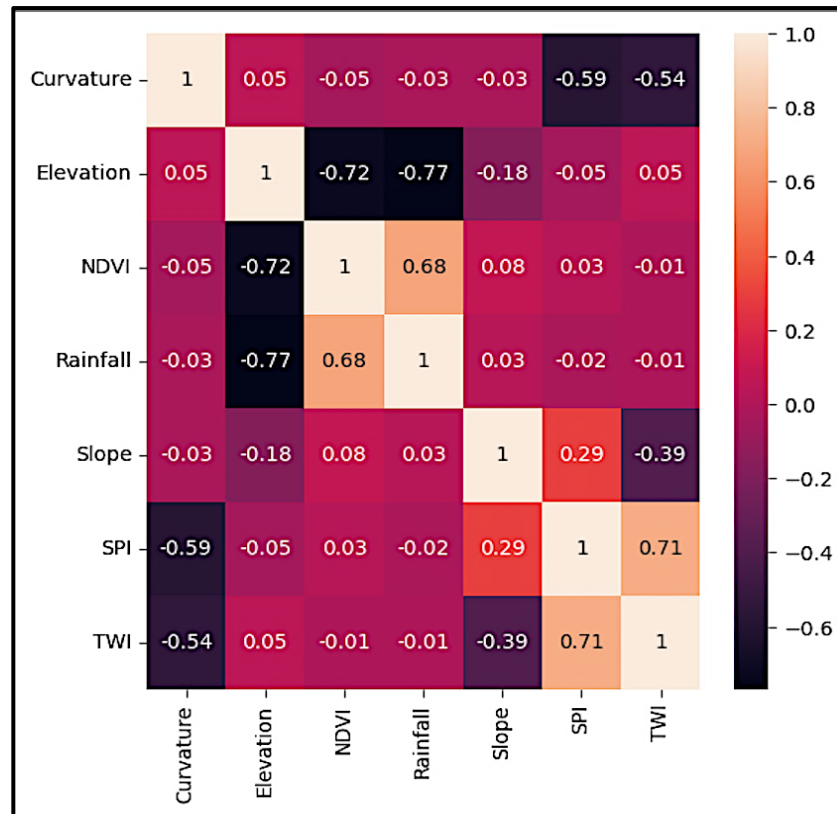


Figure 3.39: Spearman's Rho Correlation analysis

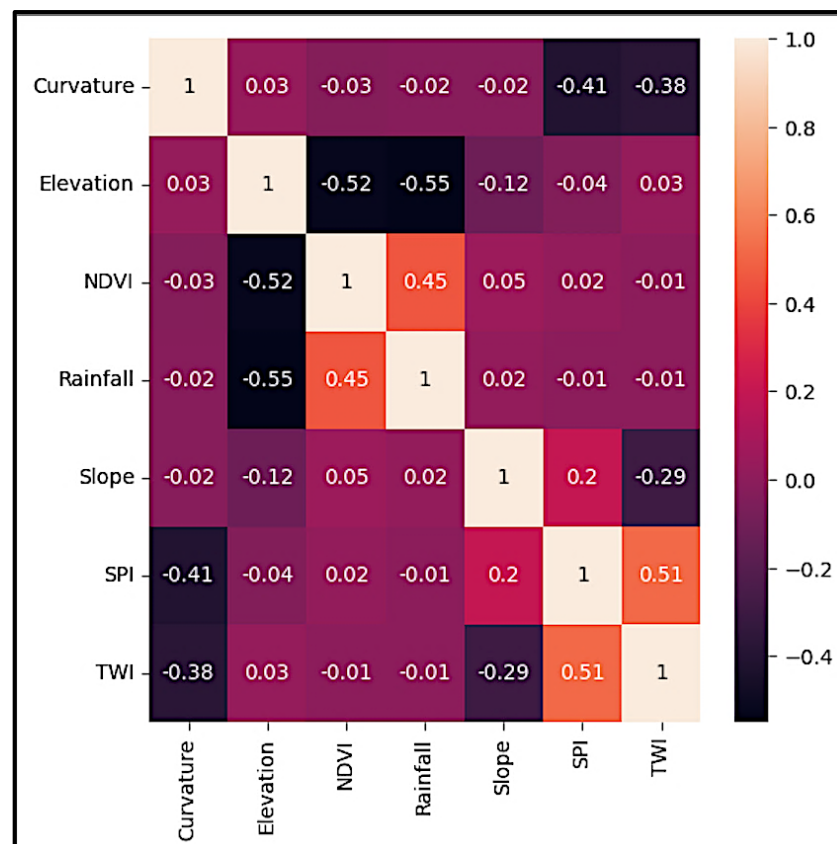


Figure 3.40: Kendall's Tau Correlation analysis

Following the unsupervised selection process, supervised feature selection techniques were applied to further refine the input feature set by directly considering the relationship of each feature with the target variable landslide occurrence. These techniques help in identifying features with the highest predictive potential and reducing model complexity while retaining essential information.

1. **Filter Method** : Filter-based methods assess the relevance of each feature individually, independent of any predictive model. These techniques evaluate the statistical relationship between each input variable and the target variable (landslide occurrence), and features are dropped based on their relative importance to the target. This method offers advantages like lower computation time and avoids overfitting of the data. There is a thumb rule for selecting the methods, which is shown in the Table 3.2 .

Table 3.2 : Filter based feature selection method

Features Target	Categorical	Numerical
Categorical	Chi Squared <i>Mutual Information</i>	t-test Mutual Information
Numerical	t-test Mutual Information	Pearson correlation Spearman's rank correlation Mutual Information

Chi-Squared Test (χ^2): This statistical test was applied to categorical variables to examine whether they are significantly associated with the target class (landslide occurrence). It evaluates the independence between each feature and the class labels. The Chi-square statistic is calculated using the formula:

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} \quad 3.8$$

where O_i and E_i represent the observed and expected frequencies, respectively. The summation ensures that discrepancies across all categories of a feature are aggregated into a single score. A higher χ^2 value implies a stronger dependency between the feature and the class label, making it a valuable metric for identifying influential predictors in classification tasks.

The accompanying bar chart in Figure 3.41 illustrates the results of the Chi-square (χ^2) test applied to categorical environmental variables, highlighting their statistical association with landslide occurrence. This filter-based method evaluates how well each feature contributes to class separation by comparing observed and expected frequencies. The variables assessed include Landcover, Distance to Road, Geomorphology, Distance to Lineaments, Distance to Stream, Soil, Geology, and Aspect. Among these, Landcover demonstrated the highest Chi² score, indicating it has the strongest discriminative power in relation to the target variable.

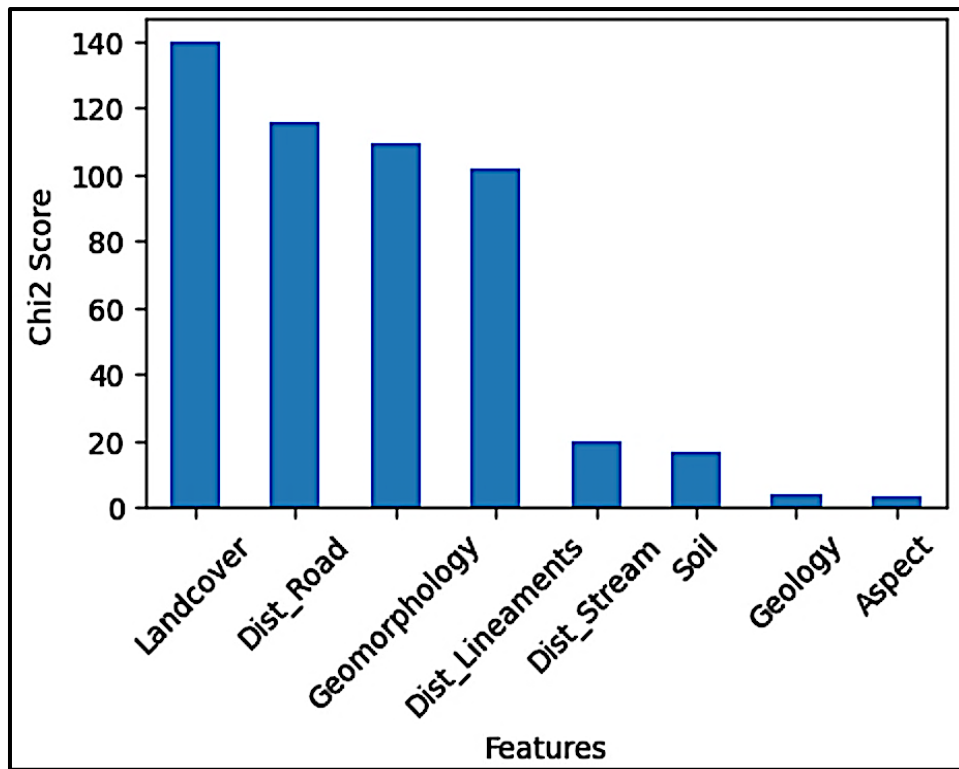


Figure 3.41: Chi-square score of the Features

Mutual Information (MI): Mutual information filter-based feature selection method shows selection for numerical and categorical features. The formula defines mutual information as :

$$I(X ; Y) = H(X) - H(X | Y) \quad 3.9$$

Here, $I(X;Y)$ quantifies the information shared between the feature (X) and the target variable (Y), representing reduction in uncertainty about X given knowledge of Y . $H(X)$ refers to the entropy of X , capturing its overall uncertainty, while $H(X|Y)$ denotes the conditional entropy, representing uncertainty in XX

after knowing Y. A higher mutual information score implies a stronger predictive relationship.

The bar graph shown in Figure 3.42 ranks environmental features by mutual information score, measuring their dependency on the target variable. Slope exhibits the highest mutual information (≈ 0.5), indicating strong predictive relevance, followed by TWI, Elevation, and Geomorphology, and SPI. Features such as Distance to Stream and Distance to Lineaments display minimal scores (≈ 0.01 – 0.02), suggesting limited contribution to classification tasks.

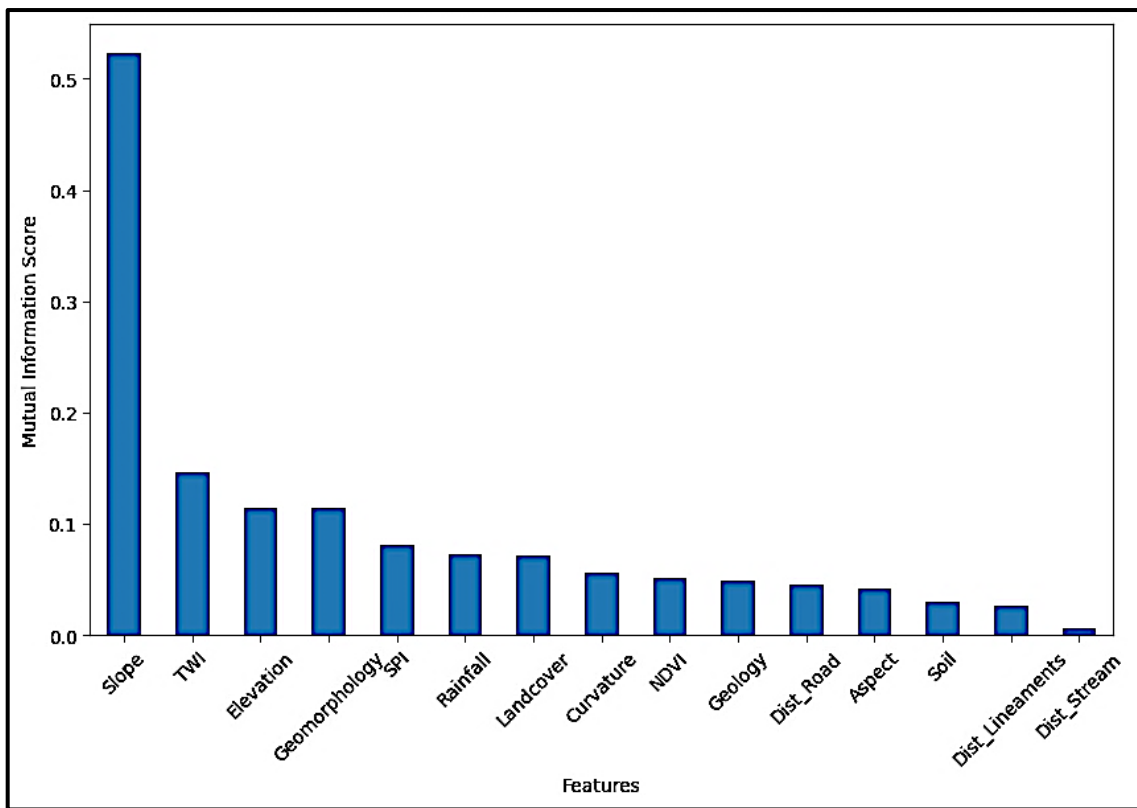


Figure 3.42: Mutual Information Score of the Features

Integrating mutual information-based feature selection optimizes landslide susceptibility modelling by retaining informative variables and reducing redundancy. Selecting high-relevance features such as Slope and TWI enhances model accuracy, ensuring the representation of key geomorphological and hydrological factors influencing landslide occurrence.

2. Wrapper Method : Wrapper-based feature selection methods play a crucial role in optimizing landslide susceptibility modelling by refining predictive

variables through iterative evaluation. These methods assess feature subsets systematically, retaining only the most relevant predictors while discarding redundant ones based on model performance. The most common techniques include Sequential Forward Selection (SFS), which adds features that enhance classification accuracy, and Sequential Backward Elimination (SBE), which starts with all features and gradually removes less significant ones. Recursive Feature Elimination (RFE) ranks features using model-based evaluation, iteratively eliminating the least informative predictors, while Exhaustive Feature Selection tests all possible feature subsets to find the optimal combination. The Boruta Algorithm distinguishes truly important features by comparing them against shadow variables—randomly shuffled versions of the original dataset.

The feature selection analysis employing Backward Elimination with Random Forest and Sequential Classifier, identifies key environmental variables influencing landslide susceptibility. Features such as Elevation, NDVI, Rainfall, Slope, Distance to Lineaments, Distance to Road, Distance to Stream, Geology, Soil, and Geomorphology exhibit varying importance in classification. The results from Gradient Boosting Classifier, demonstrating improved feature ranking accuracy was observed. By integrating wrapper-based feature selection, landslide susceptibility models enhance predictive performance by dynamically selecting only high-impact variables while reducing redundancy. These optimized selections improve interpretability, ensuring key geomorphological and hydrological factors—such as Slope, Elevation, and TWI—are effectively captured in modelling efforts.

3. **Embedded Method :** Embedded feature selection methods integrate the selection process directly into model training, allowing the learning algorithm to dynamically identify and prioritize the most relevant predictors while eliminating those with lower influence. Unlike wrapper methods, which iteratively evaluate multiple feature subsets through repetitive training, embedded methods select features as part of the model's optimization routine. This approach ensures a balance between computational efficiency and model generalization, reducing the risk of overfitting while maintaining high predictive performance.

In this study, several tree-based embedded techniques were employed to assess feature relevance, including Gini Importance, Mean Decrease Accuracy, Permutation Feature Importance, and the Extra Trees Classifier. These models evaluate the contribution of each variable to classification performance during training, offering an interpretable and data-driven hierarchy of feature importance. Multiple visualization outputs were generated to illustrate this process.

The Gini Importance plot, derived from the Random Forest model, identified Slope as the most influential predictor, followed by TWI, Elevation, SPI, Geomorphology, Distance to Road, NDVI, and Rainfall.

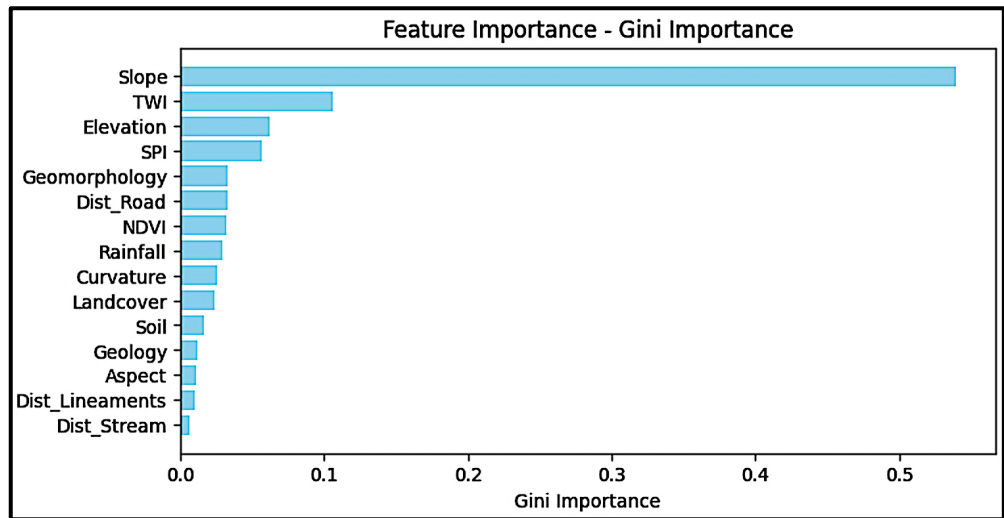


Figure 3.43: Gini Importance Plot

Similarly, the Mean Decrease Accuracy chart reinforced the significance of Slope and TWI, with notable importance also assigned to Distance to Road, Rainfall, NDVI, and Curvature. The Permutation Feature Importance graph highlighted Slope, TWI, and Distance to Road as top contributors, supporting the stability of these variables across models. Finally, the Extra Trees Classifier ranking emphasized a similar trend, with Slope, TWI, Elevation, and Distance to Road consistently emerging as high-impact variables.

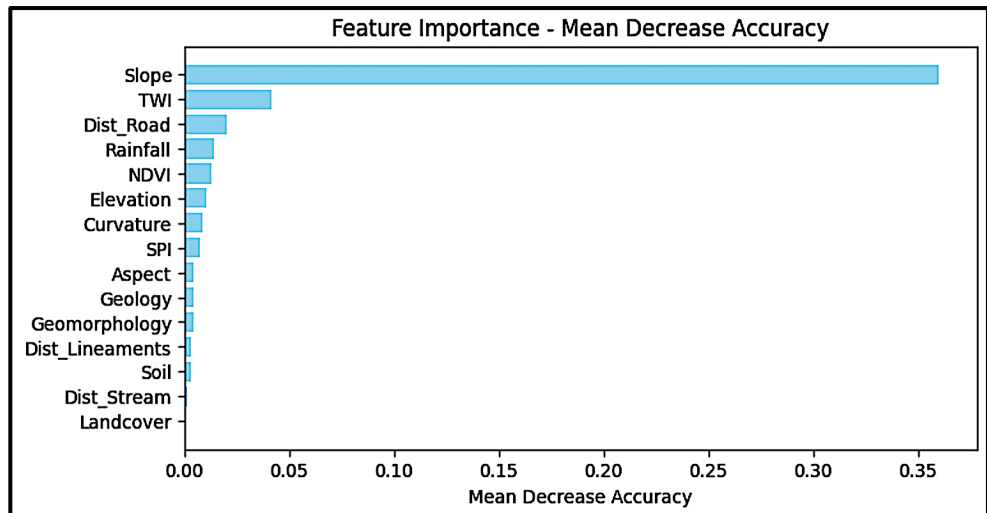


Figure 3.44: Mean Decrease Accuracy Plot

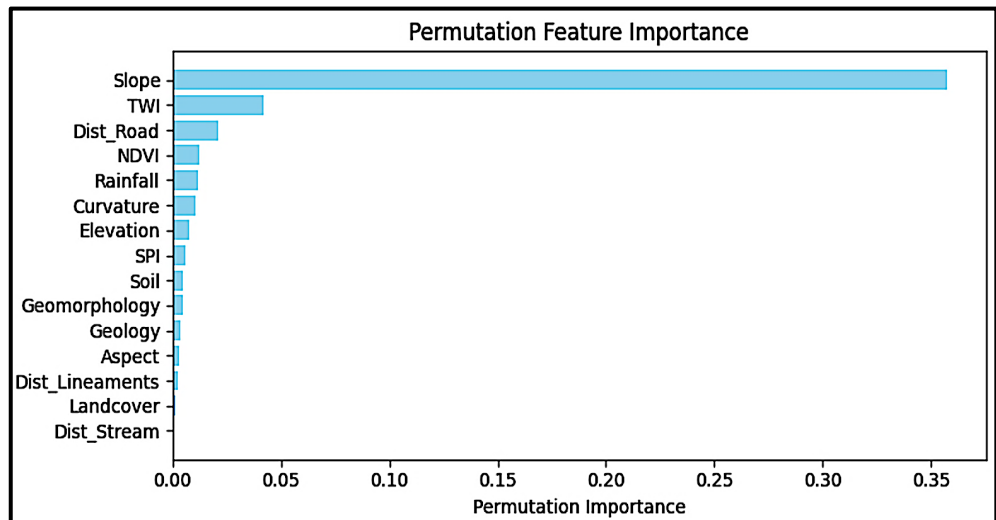


Figure 3.45: Permutation Feature Importance Plot

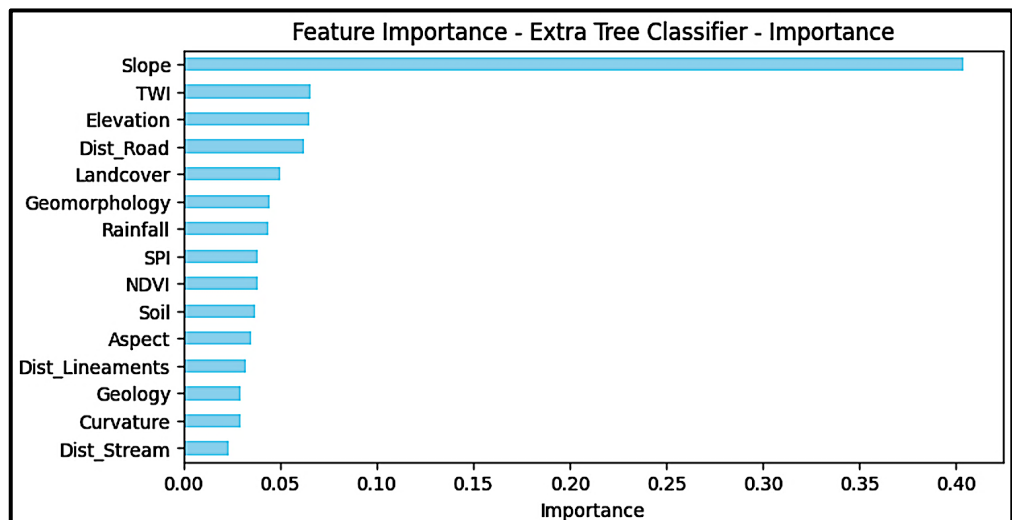


Figure 3.46: Extra Tree Classifier Importance Plot

The feature rankings produced by these methods reflect the intrinsic value of predictors associated with terrain morphology, hydrological dynamics, and geological structure, which are critical to modelling landslide susceptibility. By embedding feature selection within the learning algorithm, the models not only improve computational efficiency but also enhance interpretability by aligning selected features with physically meaningful environmental parameters. Furthermore, the reduction of redundancy and retention of high-impact variables contribute to the robustness of the predictive framework, ensuring that only the most informative predictors are carried forward to the final modelling stage. Looking at the results of all feature selection techniques, Slope has been the most influential feature affecting the model performance. Apart from Slope, the other features like Elevation, NDVI, Geomorphology, Geology, Soil, Landcover, Distance to Roads, Distance to Streams are also the significant features amongst all the features considered for LSM.

3.6 GIS Based Landslide Susceptibility Modelling with Frequency Ratio Method

The Frequency Ratio (FR) Method follows a structured approach to assess landslide susceptibility as shown in

Figure 3.47 below. The methodology consists of several sequential steps as listed below;

1. **Calculation of Frequency Ratio (FR) for Each Class of Variables:** The first step involves computing the Frequency Ratio for each class within every selected variable using a mathematical formula. This step quantifies the relationship between landslide occurrences and causative factors. The table for Frequency Ratio for all variables is appended in Appendix A.

$$FR = \frac{\frac{NL_i}{NL_t}}{\frac{NC_i}{NC_t}} \quad 3.1$$

Where;

NL_i : Number of landslides in class i

NL_t : Total number of landslides in the entire study area

NC_i : Number of pixels of class i

NC_t : Total number of pixels in the entire study area.

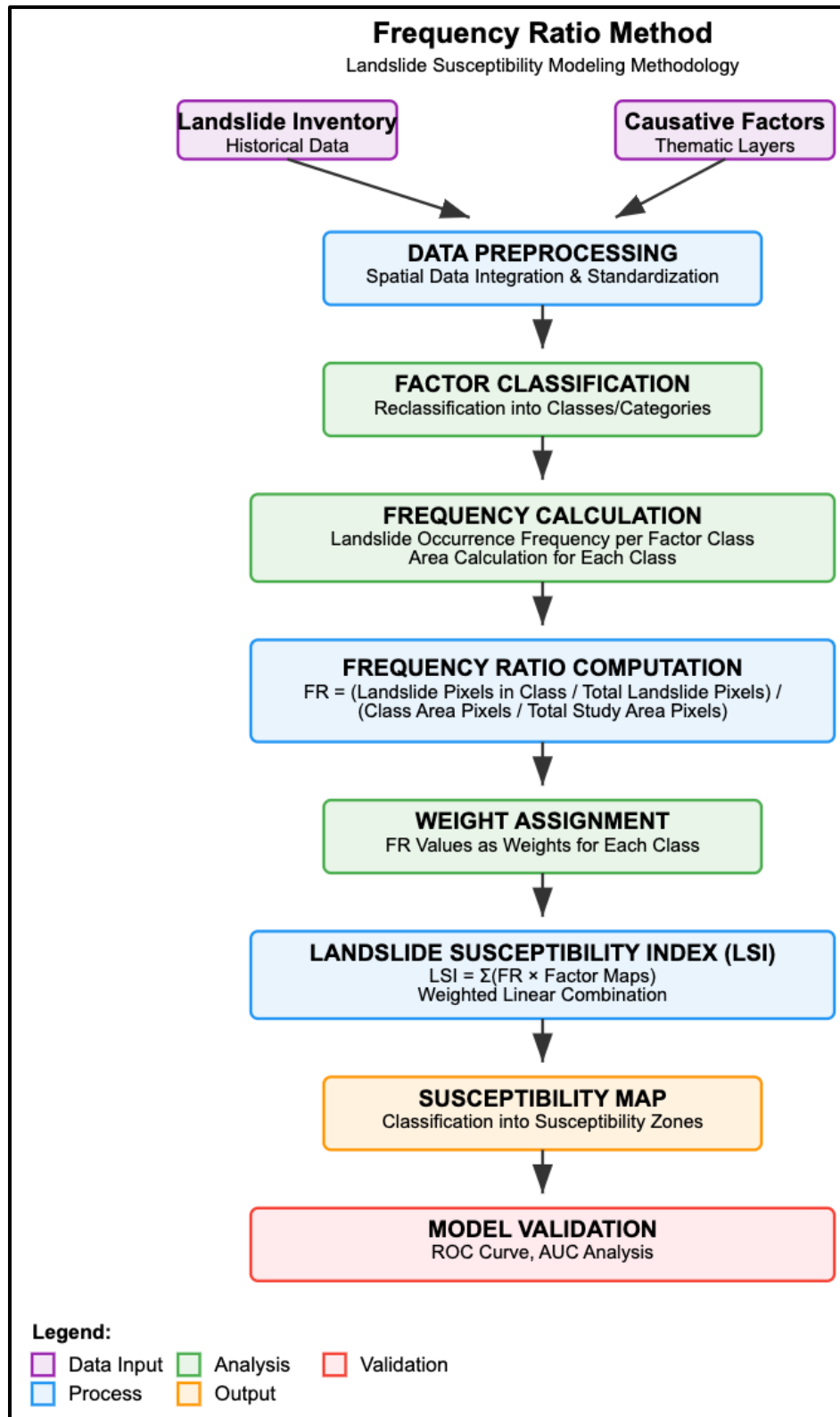


Figure 3.47: Methodological framework for LSM using FR model

2. **Assignment of Relative Frequency (RF) and Layer Reclassification:** Each class category of the contributing factors is assigned a Relative Frequency (RF)

based on the calculated FR values. Following this, all factor layers are reclassified using their respective RF values to standardize the dataset for further analysis.

$$RF = \frac{\text{Frequency Ratio of a Particular Class}}{\text{Summation of Frequency Ratio of every class of a particular variable}} \quad 3.2$$

3. **Determination of Prediction Rate (PR) for Each Variable:** Each variable's predictive influence on landslide susceptibility is determined using a mathematical expression for the Prediction Rate (PR). This step helps assess the significance of individual factors in contributing to landslide events. See Appendix A for detailed results.

$$\text{Prediction Rate} = \frac{\text{Maximum RF} - \text{Minimum RF}}{\text{Minimum}_{(Max\ RF - Min\ RF)}} \quad 3.3$$

Table 3.3 : Prediction Rate for all Variables

Layer	PR
Curvature	1
Distance from River	7.59
NDVI	9.71
LULC	9.87
Slope	10.50
Geology	11.99
Soil Type	12.29
Elevation	14.11
Distance from Road	20.19
Geomorphology	29.64

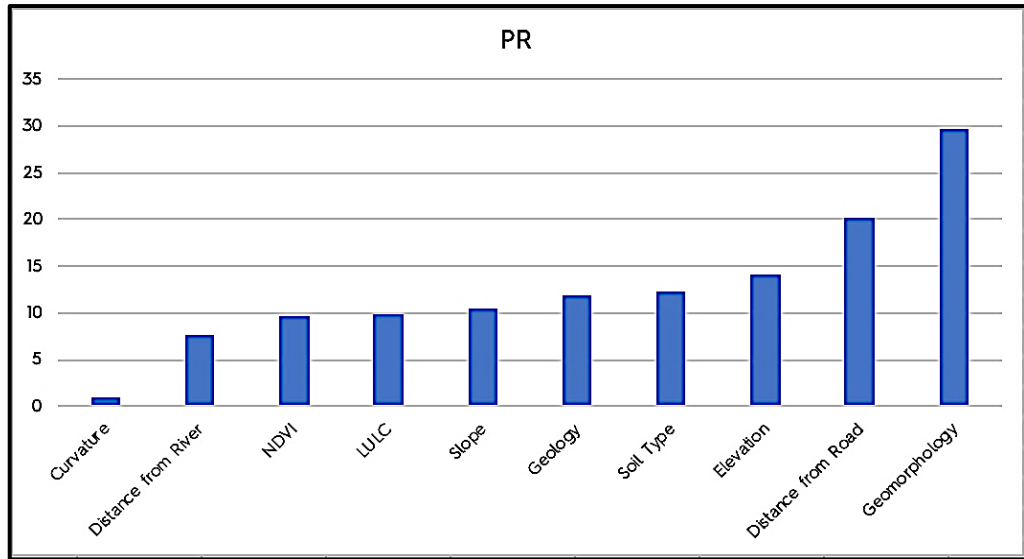


Figure 3.48: Prediction Rate for Variables

4. **Generation of the Landslide Susceptibility Index (LSI) Map:** The Landslide Susceptibility Index (LSI) map is generated by summing the weighted contributions of all factors, where each factor's RF value is multiplied by its respective PR. This process results in an overall susceptibility score for different regions within the study area.
5. **Training and Testing Data Splitting:** The landslide inventory consists of 1,561 landslide locations, which are divided into training (70%) and testing (30%) datasets. The training set, consisting of 1,093 locations, is used to compute the FR values and develop the LSM, while the test set, comprising 468 locations, is employed for model validation.
6. **Validation and Accuracy Assessment:** The final step involves assessing the accuracy and reliability of the Landslide Susceptibility Map. This validation was performed using landslide inventory data and statistical measures, including Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values. The ROC curve for the FR model, illustrated in Figure 3.50, resulted in an AUC value of 0.742, indicating a moderate prediction capability.
7. **Classification of the Landslide Susceptibility Map:** To create a more interpretable Landslide Susceptibility Map, the LSI values are categorized into five susceptibility classes: very low, low, moderate, high, and very high. This classification follows the Jenks Natural Breaks method, ensuring optimal grouping of susceptibility levels.

$$LSI = \sum_{i=1}^n PRi * RFi \quad 3.4$$

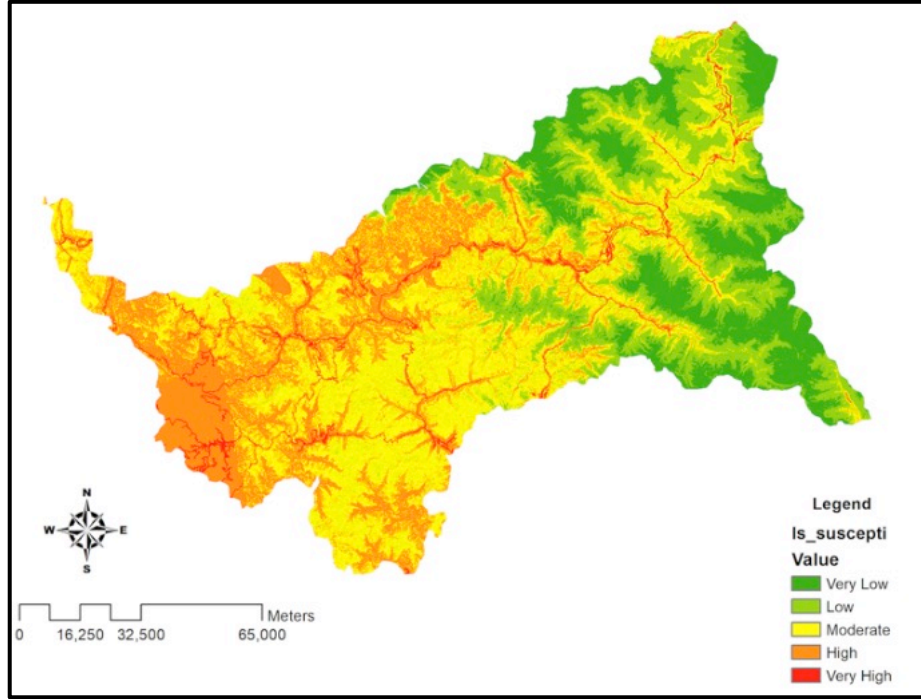


Figure 3.49: LSM using FR Approach

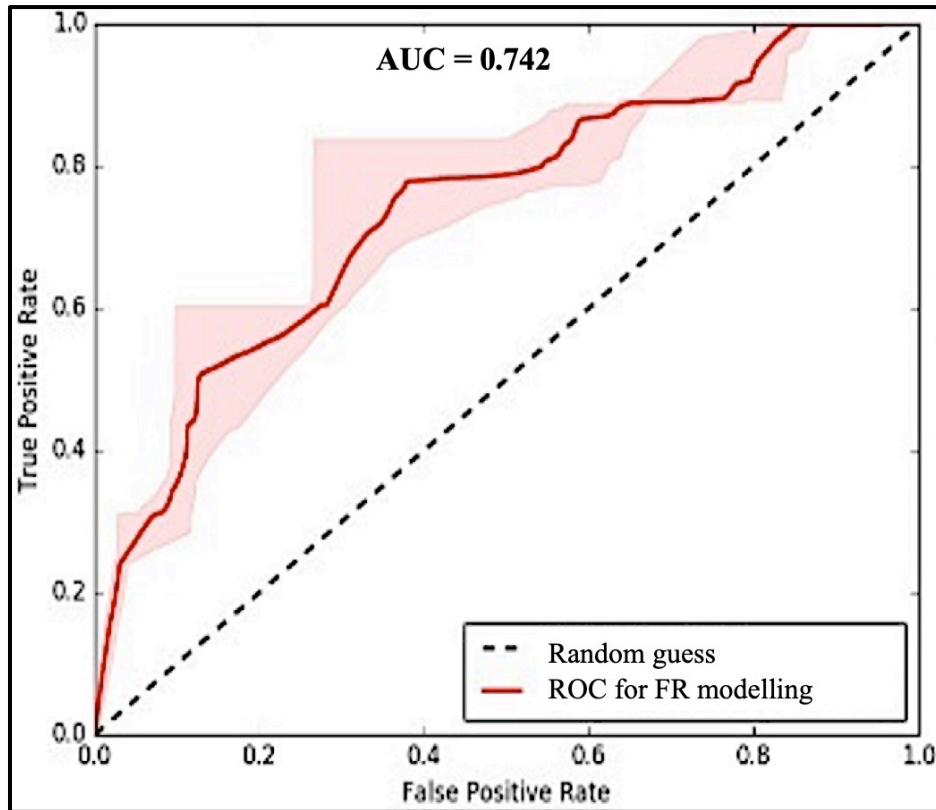


Figure 3.50: ROC for FR Approach

3.7 GIS Based LSM with Analytical Hierarchy Process

The Analytical Hierarchy Process (AHP) is a structured decision-making approach used for landslide susceptibility assessment as shown in the Figure 3.51 below.

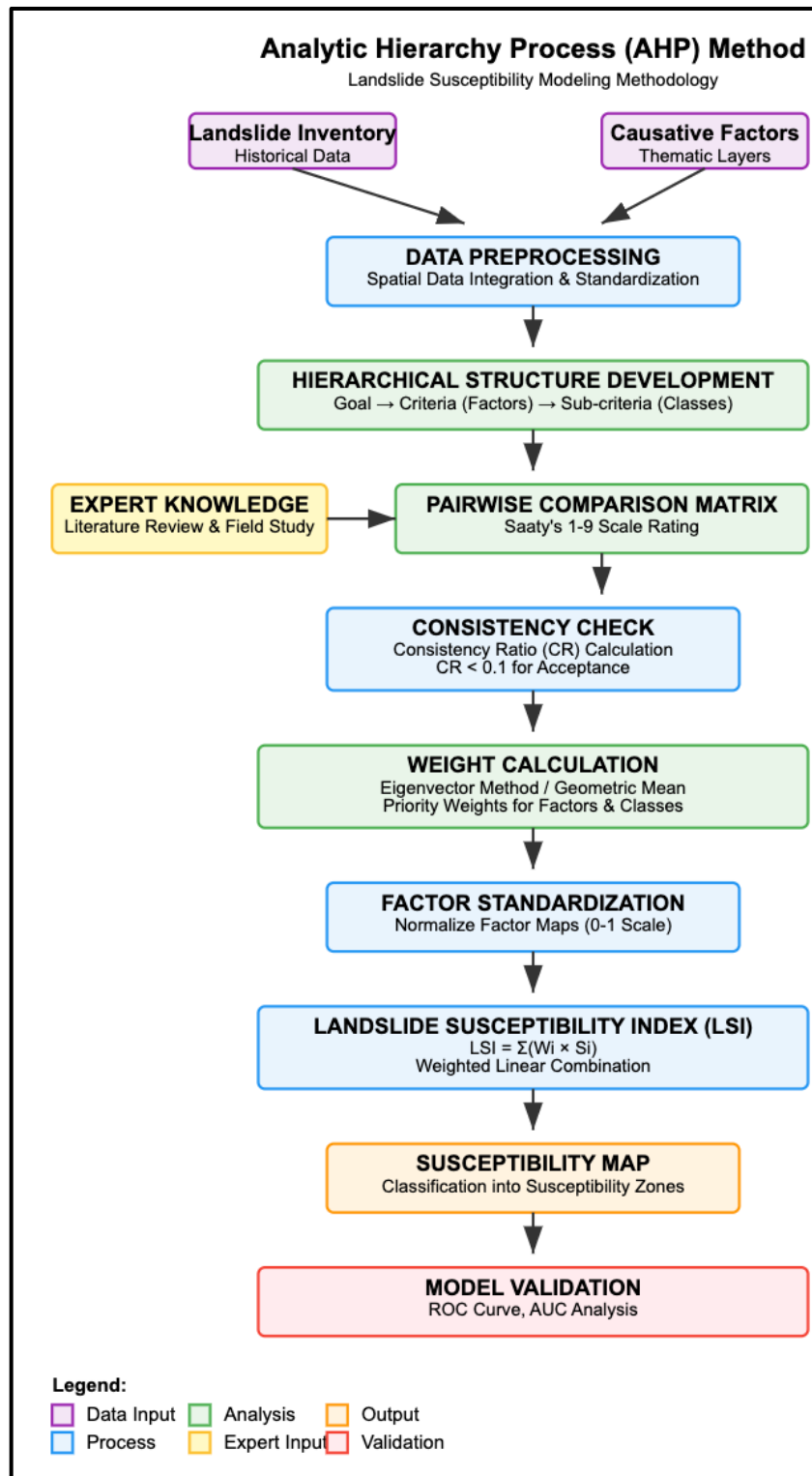


Figure 3.51: Methodological framework for LSM using AHP

The stepwise process followed in the AHP methodology is as listed below and detailed results are attached in Appendix A.

1. **Pairwise Comparison Matrix:** A pairwise comparison matrix is constructed using the Saaty scale, where each landslide causative factor is systematically compared against others to determine its relative importance. This matrix quantifies the influence of each factor in the context of landslide susceptibility.
2. **Normalized Pairwise Comparison Matrix:** The pairwise comparison matrix is then normalized to ensure balanced representation and consistency in decision-making. This step adjusts the initial values to a standardized scale, facilitating accurate weight assignment.
3. **Determination of Criteria Weights:** The normalized values obtained from the previous step are averaged to derive the criteria weights as shown in Figure 3.52. These weights reflect the relative significance of each causative factor in landslide susceptibility assessment.

Matrix	Geomorphology	Distance from Road	Elevation	Soil type	Geology	Slope	LULC	NDVI	Distance from River	Curvature	normalized principal Eigenvector
	1	2	3	4	5	6	7	8	9	10	
Geomorphology	1	2	2	3	3	3	3	3	4	5	22.50%
Distance from Road	1/2	1	2	2	2	2	2	2	3	5	15.52%
Elevation	1/2	1/2	1	2	2	2	2	2	3	5	13.49%
Soil type	1/3	1/2	1/2	1	2	2	2	2	2	4	10.80%
Geology	1/3	1/2	1/2	1/2	1	2	2	2	2	4	9.38%
Slope	1/3	1/2	1/2	1/2	1/2	1	1/2	2	2	4	7.09%
LULC	1/3	1/2	1/2	1/2	1/2	2	1	2	2	4	8.15%
NDVI	1/3	1/2	1/2	1/2	1/2	1/2	1/2	1	2	4	6.16%
Distance from River	1/4	1/3	1/3	1/2	1/2	1/2	1/2	1/2	1	3	4.56%
Curvature	1/5	1/5	1/5	1/4	1/4	1/4	1/4	1/4	1/3	1	2.35%

Figure 3.52: Determination of Criteria Weights

4. **Computation of Weighted Sum:** The weighted sum is calculated by multiplying the criteria weights with the corresponding values in the pairwise comparison matrix. The sum of these weighted values for each row provides the weighted sum vector.

5. **Determination of the Consistency Vector (Ratio):** The consistency vector is computed by dividing the weighted sum values by their corresponding criteria weights. This step helps assess the logical consistency of judgments made in the pairwise comparison process.

Table 3.4 : Fundamental Scale for Pairwise Comparison

Intensity of Importance	Definition	Explanation
1	Equal importance	Two elements contribute equally to the objective
3	Moderate importance	Experience and judgment moderately favour one element over another
5	Strong importance	Experience and judgment strongly favour one element over another
7	Very strong importance	One element is favoured very strongly over another; its dominance is demonstrated in practice
9	Extreme importance	The evidence favouring one element over another is of the highest possible order of affirmation
Intensities of 2, 4, 6, and 8 can be used to express intermediate values, while finer adjustments such as 1.1, 1.2, 1.3 , etc., account for elements with very close importance.		

6. **Generation of Landslide Susceptibility Map:** Using the criteria weights determined in Step 3, the final Landslide Susceptibility Map (LSM) is developed. This is achieved through the arithmetic summation of the weighted values of all causative factors. The resultant susceptibility map is illustrated Figure 3.53.

$$LSI = \sum_{i=1}^n R_i * W_i \quad 3.6$$

7. **Validation using AUROC Analysis:** The predictive performance of the AHP-based Landslide Susceptibility Map is validated using the Area Under the Receiver Operating Characteristic (AUROC) curve. This evaluation, shown in Figure 3.54, provides a quantitative measure of the model's accuracy in differentiating landslide-prone areas.

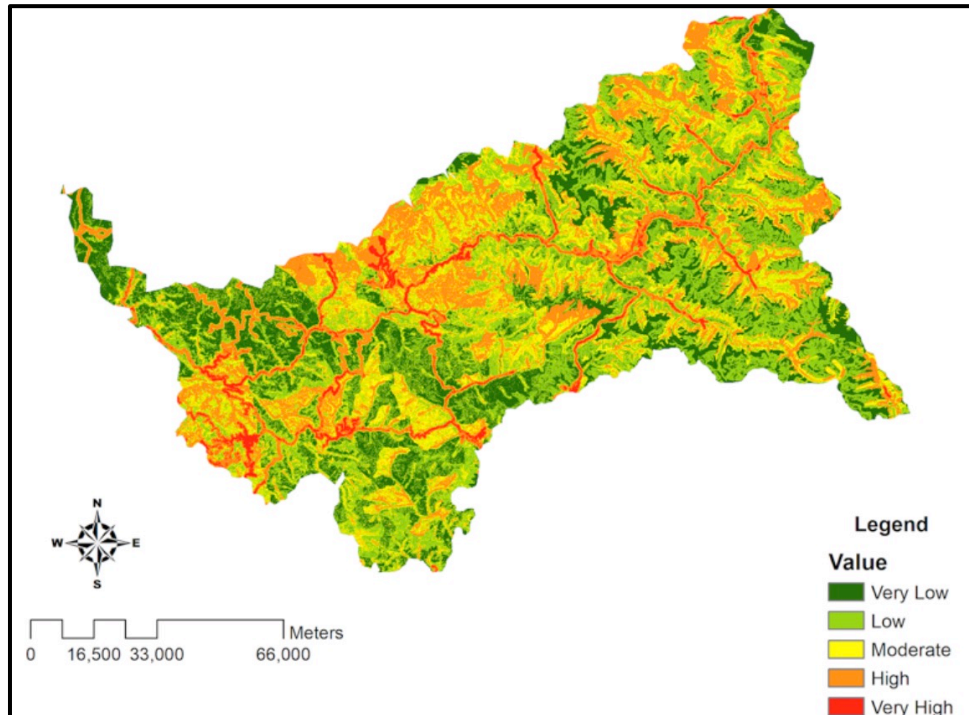


Figure 3.53: LSM using AHP Approach

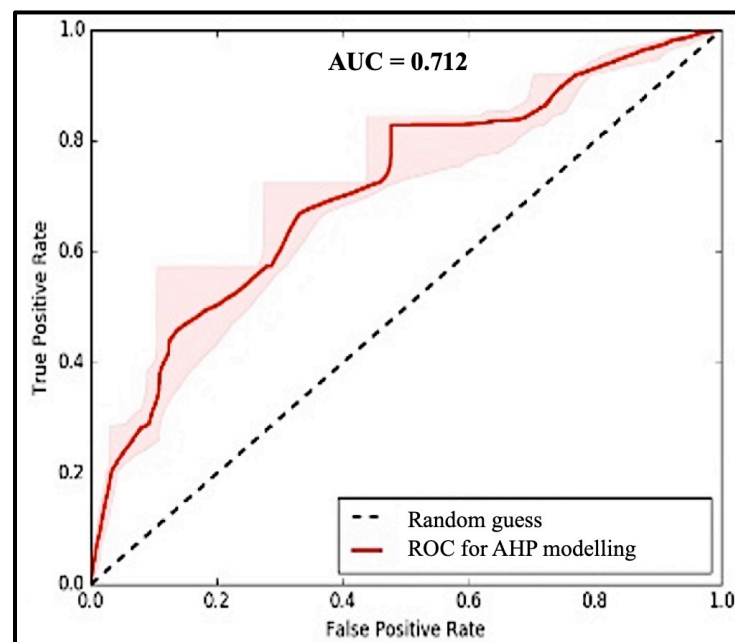


Figure 3.54: ROC Curve for AHP

3.8 Analysis using Standalone Machine Learning Models

The Process for ML modelling is outlined in Figure 3.55. Python (version 3.9.0), R (RStudio version 4.4.0), and ArcGIS (Version 10.6.1) were used in an integrated workflow for landslide susceptibility modeling. ArcGIS was primarily utilized for spatial data preprocessing, including terrain parameter extraction, raster reclassification, and overlay analysis. R, through its extensive package ecosystem available on the Comprehensive R Archive Network (CRAN), was employed for statistical analysis, probability distribution fitting, and data visualization. The ggplot2 package was used to generate plots such as boxplots, histograms, and correlation graphs, while distribution fitting helped assess the statistical characteristics of landslide and non-landslide locations. Analyses in R were conducted using the RStudio environment for its efficient scripting and visualization capabilities. For machine learning implementation and geospatial processing, Python was the primary platform, utilizing libraries such as scikit-learn, xgboost, and rasterio. Model development and analysis were carried out using both Jupyter Notebook for its interactive, notebook-style coding and Spyder for script-based analysis and debugging. This combination of tools enabled robust data analysis, effective modelling, and reproducible results across all stages of the susceptibility mapping workflow.

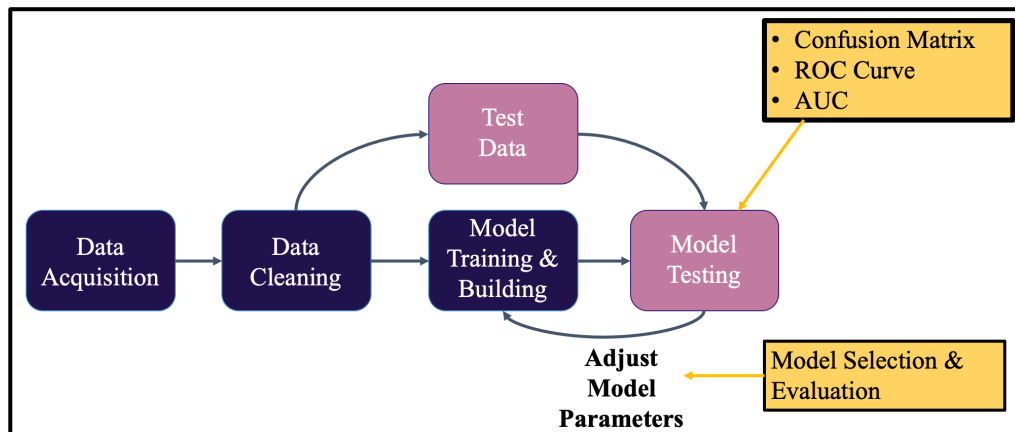


Figure 3.55: Machine Learning Modelling Process

To ensure high-quality data for machine learning (ML) modelling, several preprocessing steps were undertaken, including outlier detection and removal, handling of missing values, encoding of categorical features, and feature scaling. The dataset consisted of 3,118 data points, with an equal split between landslide and non-landslide locations (50% each). For model training and evaluation, the dataset was randomly divided into training

(70%) and testing (30%) sets, which ensures sufficient data for training while maintaining a representative test set for unbiased validation. It also ensures a balanced distribution of landslide and non-landslide points across both subsets.

3.8.1 Model Performance and Statistical Evaluation

Various standalone machine learning models were developed and evaluated to assess their predictive capabilities for landslide susceptibility mapping. The selected models included Logistic Regression (LGR), Support Vector Machine (SVM), Kernel Support Vector Machine (K-SVM), K-Nearest Neighbours (K-NN), Naïve Bayes (NB), and Decision Tree (DT). Each model was trained on the designated training dataset and tested on the unseen test set to evaluate performance using key classification metrics.

To further assess classification accuracy, confusion matrices were generated for each model, detailing the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For example, Logistic Regression correctly classified 453 true negatives and 419 true positives, while yielding 16 false positives and 48 false negatives. These results were analysed in the context of Type I errors (false positives) and Type II errors (false negatives), with the latter being more critical due to the risks associated with undetected landslides. Models such as Decision Tree and Kernel SVM demonstrated lower error rates and better reliability, whereas Naïve Bayes showed comparatively higher misclassification rates, particularly in landslide-prone areas.

Table 3.5 : Confusion Matrix Results

Model Name		Predicted: no (0)	Predicted: yes (1)
Logistic Regression	Actual : no (0)	453	16
	Actual : yes (1)	48	419
K -nearest Neighbour	Actual : no (0)	428	41
	Actual : yes (1)	62	405
Support Vector Machine	Actual : no (0)	455	14
	Actual : yes (1)	61	406
Kernel SVM	Actual : no (0)	459	10
	Actual : yes (1)	54	413
Naïve Bayes	Actual : no (0)	414	55
	Actual : yes (1)	45	422
Decision Tree	Actual : no (0)	433	36
	Actual : yes (1)	27	440

Additionally, K-Fold validation accuracy and standard deviation were analysed across models to evaluate consistency as shown in Figure 3.56 . K-Fold cross-validation is a resampling technique where the dataset is split into K equal subsets (folds); the model is trained on $(K-1)$ folds and tested on the remaining fold, repeating this process K times. Five-fold cross-validation offers a balanced approach by providing more reliable performance estimates than the simple hold-out method and requiring less computational time than ten-fold. It reduces variance associated with a single train-test split while being more efficient than ten-fold, making it a practical and widely used choice for model evaluation. Five-fold cross-validation has been employed to ensure that each subset of the dataset serves as the test set once, enabling a comprehensive assessment of the model's performance across different data segments. The average accuracy across these data segments provides a robust estimate of model performance and helps minimize overfitting. The results revealed that SVM (92.90%) and Kernel SVM (92.80%) achieved the highest average accuracy across the folds, followed closely by Logistic Regression (92.80%). In terms of variability, Naïve Bayes exhibited the highest standard deviation (3.14%), indicating less stable performance, while Decision Tree and Kernel SVM showed lower standard deviations (1.52% and 1.87%, respectively), reflecting greater reliability across different data splits. These findings highlight the significance of both confusion matrix analysis and K-Fold cross-validation in selecting robust models for landslide susceptibility assessment, guiding the choice of classifiers based on predictive accuracy, error minimization, and generalization ability.

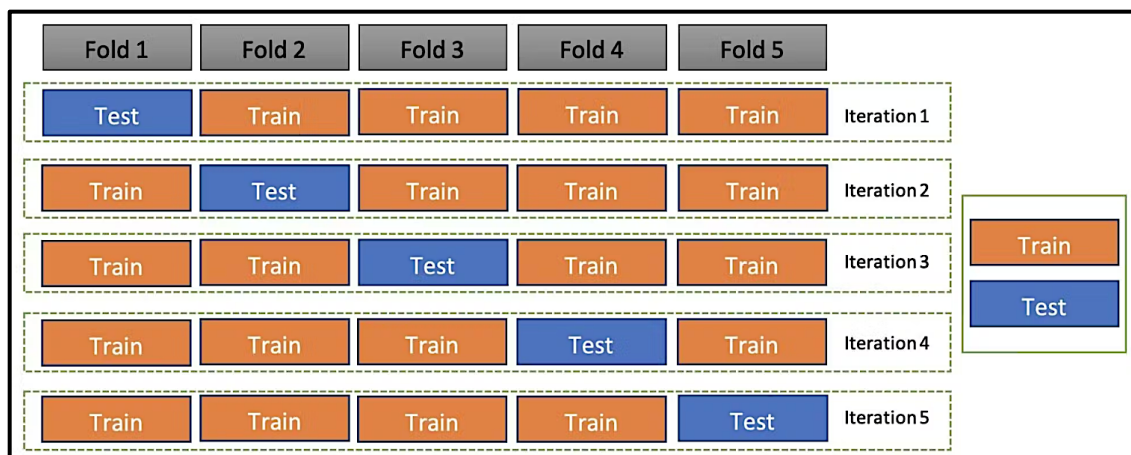


Figure 3.56: K-Fold Cross Validation

The statistical performance metrics of the standalone models on the test set are summarized in, with corresponding results illustrated in Figure 3.57.

Table 3.6 : Statistical Evaluation results for Single ML Models on Testing Datasets

Model	Accuracy	Recall	Precision
KNN	0.88996	0.86724	0.90807
Naïve Bayes	0.89316	0.90364	0.88470
SVM	0.91987	0.86938	0.96667
LR	0.93162	0.89722	0.96322
Kernel SVM	0.93162	0.88437	0.97636
Decision Tree	0.93269	0.94218	0.92437

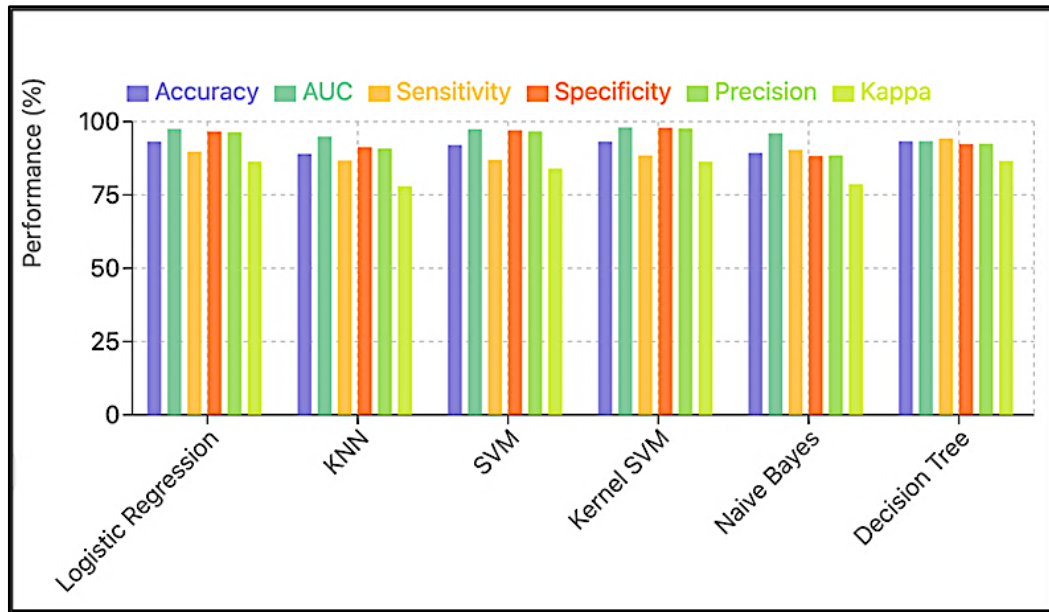


Figure 3.57: Statistical evaluation for single ML Models on Testing Datasets

Following to confusion matrix and cross-validation assessments, Receiver Operating Characteristic (ROC) Curve analysis was performed to evaluate the discriminative power of each classifier. The ROC curve plots the True Positive Rate (sensitivity) against the False Positive Rate, providing a visual and quantitative understanding of each model's ability to distinguish between landslide and non-landslide zones across various threshold levels. AUROC provides insights into a model's ability to differentiate between classes, serves as a key summary metric, where values closer to 1.0 indicate better classification performance.

The comparative AUC values revealed that Kernel SVM achieved the highest score (AUC = 0.9797), followed closely by Logistic Regression (AUC = 0.9749) and SVM (AUC = 0.9741), suggesting their superior ability to distinguish between susceptible and stable regions. Naïve Bayes and K-Nearest Neighbours (KNN) also demonstrated strong performance with AUC values of 0.9604 and 0.9494, respectively. In contrast, the Decision Tree model recorded the lowest AUC (0.9327), indicating relatively lower classification effectiveness among the evaluated models.

The ROC curve analysis as shown in Figure 3.58 reinforces the earlier findings from confusion matrix and K-Fold evaluations, confirming that Kernel SVM, SVM, and Logistic Regression offer the most robust performance in landslide susceptibility classification. For detailed results see Appendix B.

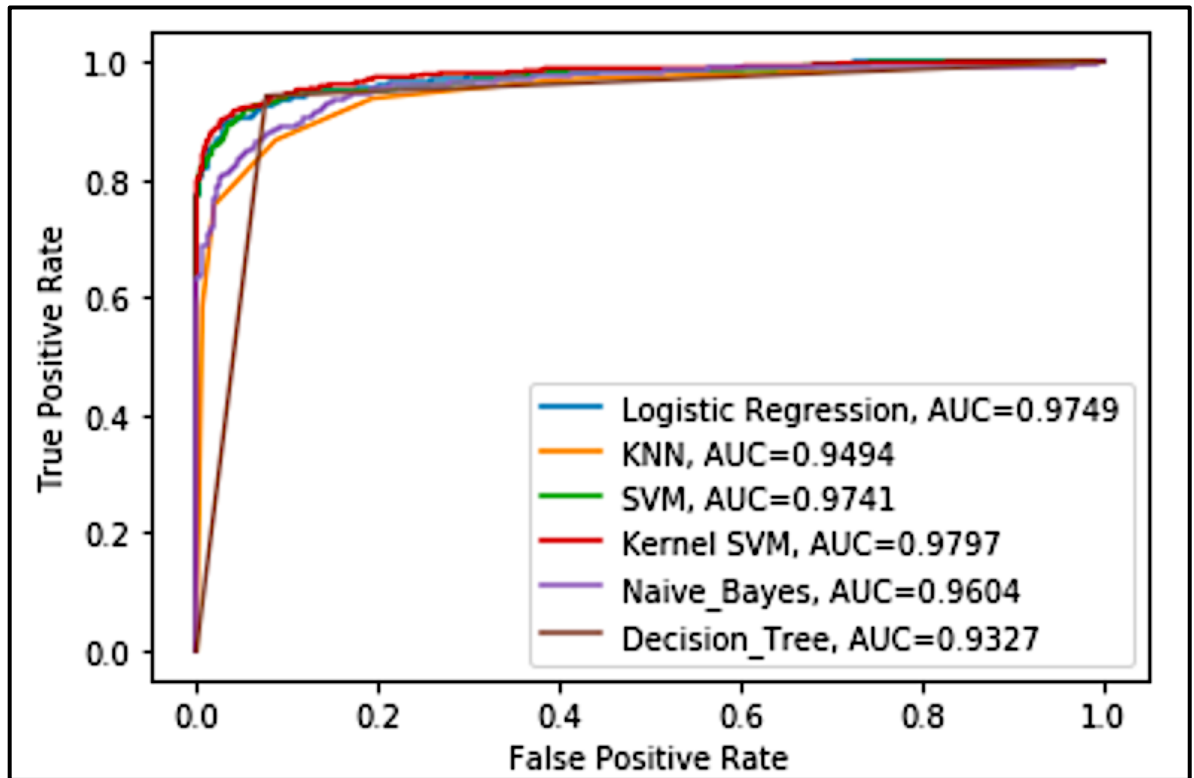
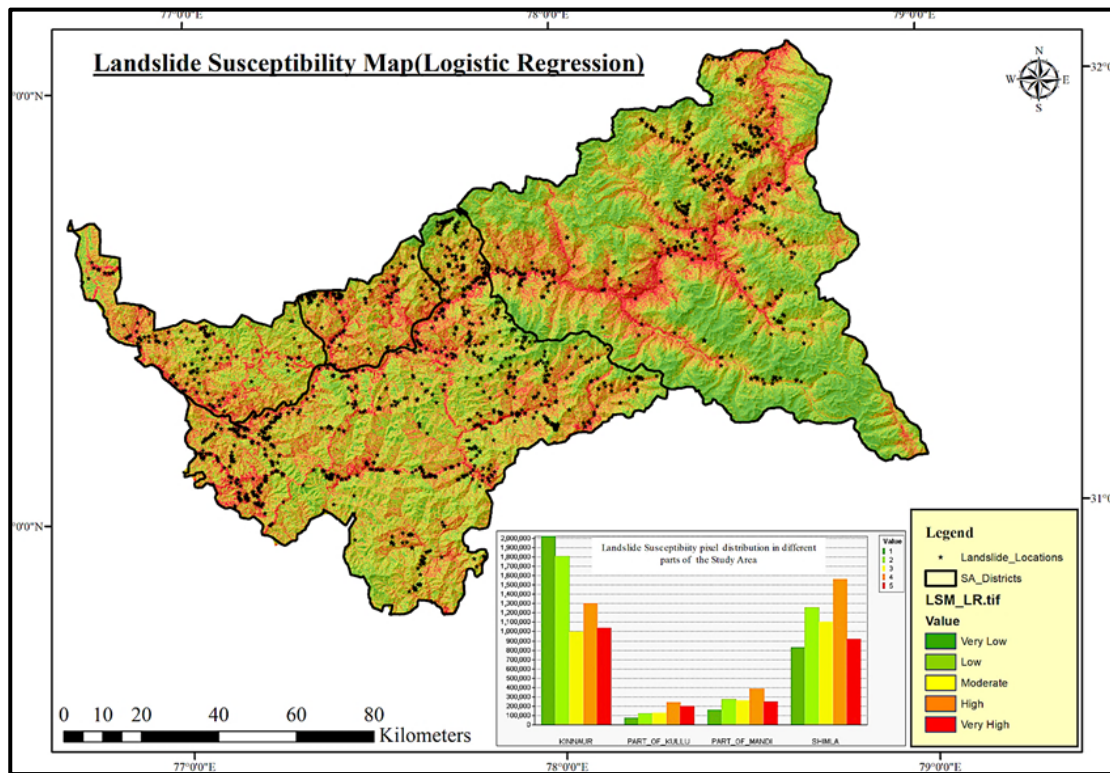


Figure 3.58: ROC Plot for Standalone ML Models

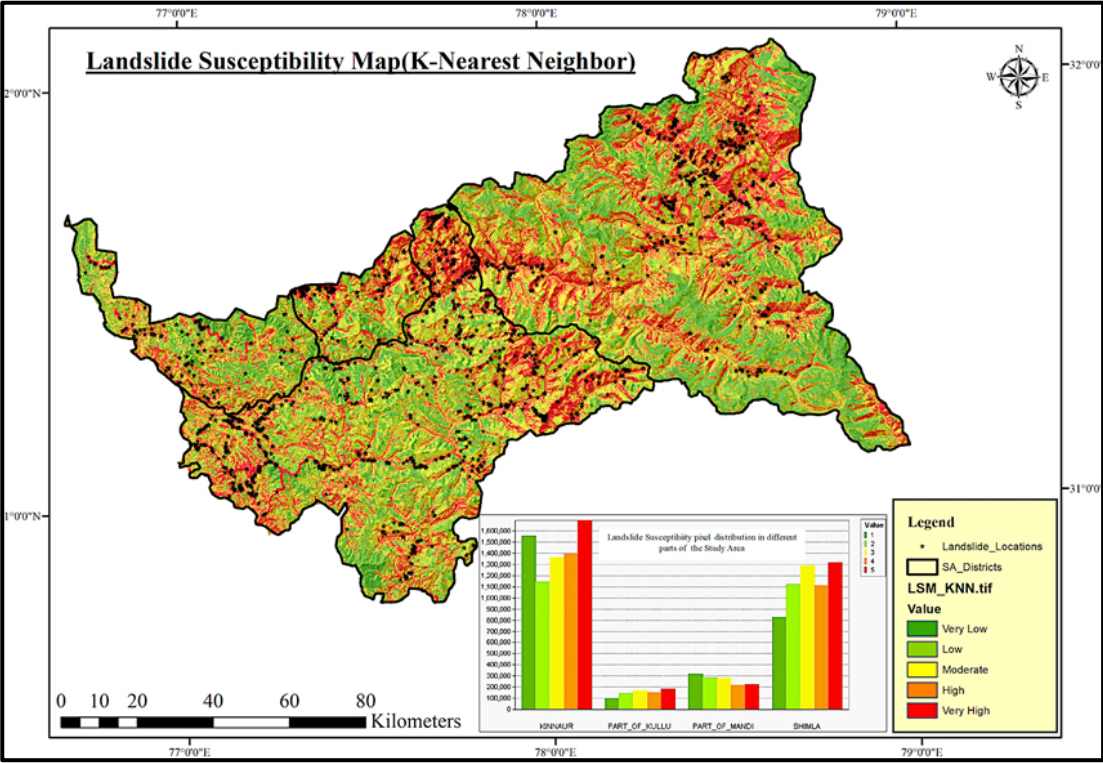
3.8.2 Landslide Susceptibility Maps

Using the trained models, Landslide Susceptibility Maps (LSMs) as shown in Figure 3.59 were generated based on probability-based classification at each pixel. The resulting raster datasets were then reclassified into five susceptibility categories, ranging from very low to very high, using the Jenks Natural Breaks Classification Method. Also, the pixel

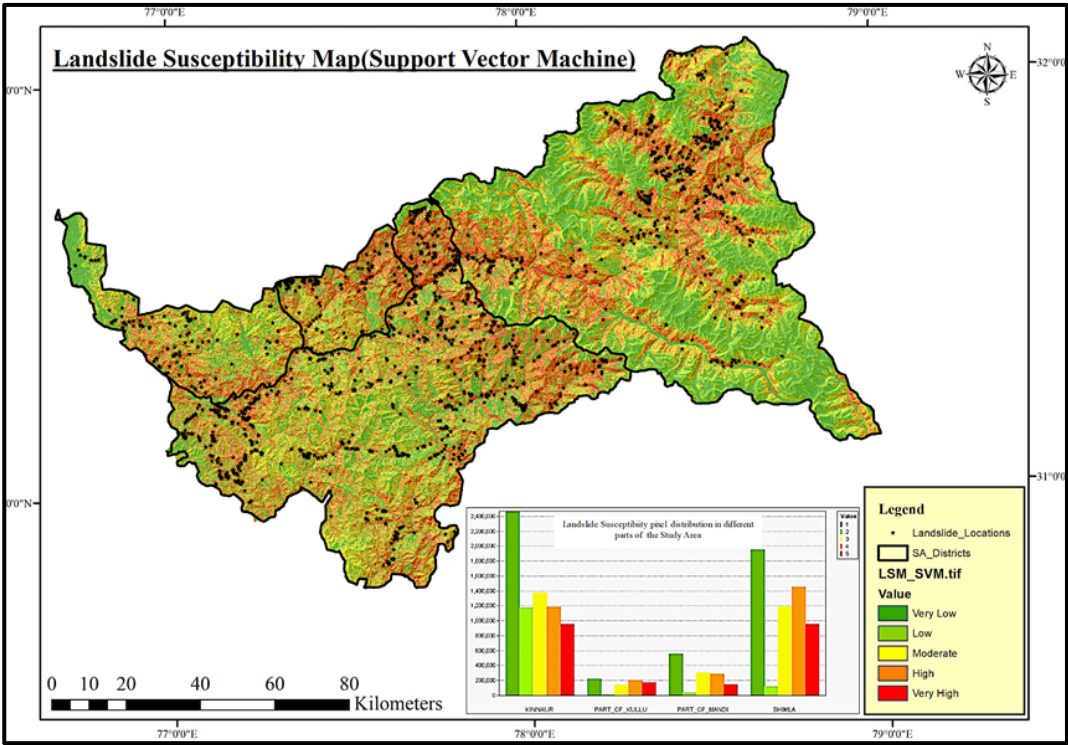
distribution of every susceptibility class in each zone of study area (district wise) has been shown besides the legend of the landslide susceptibility maps of each model. These zonal statistics has been used to identifying the high landslide risk zones within the study area. This has been discussed in length in the next chapter.



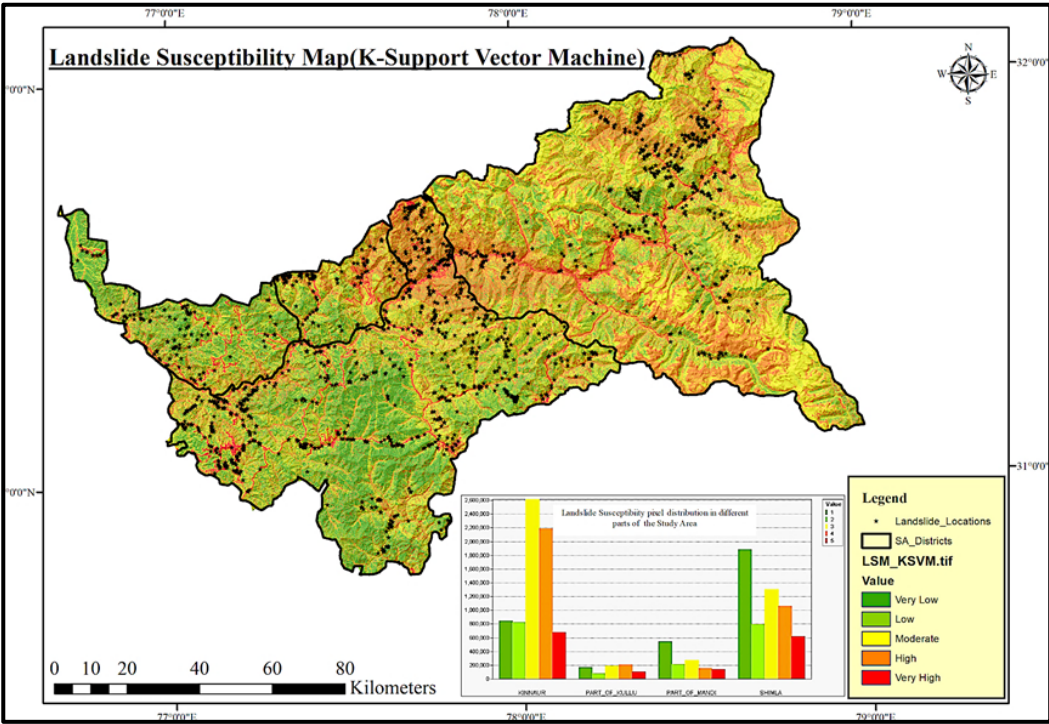
a.



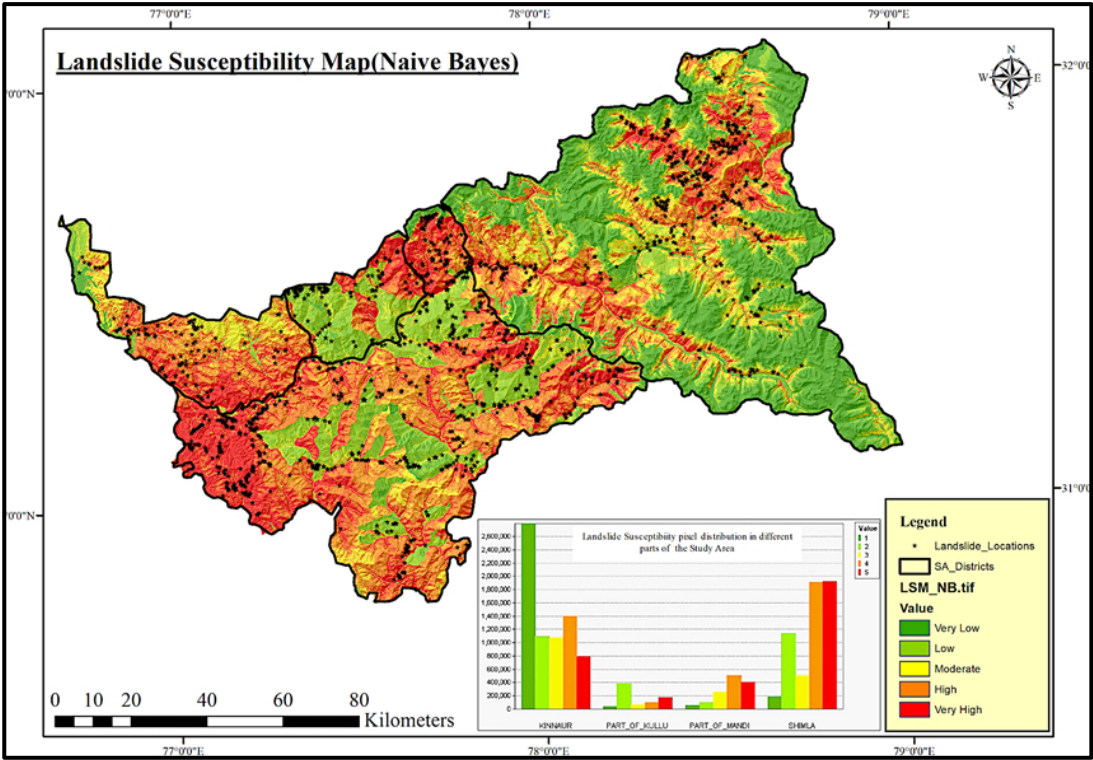
b.



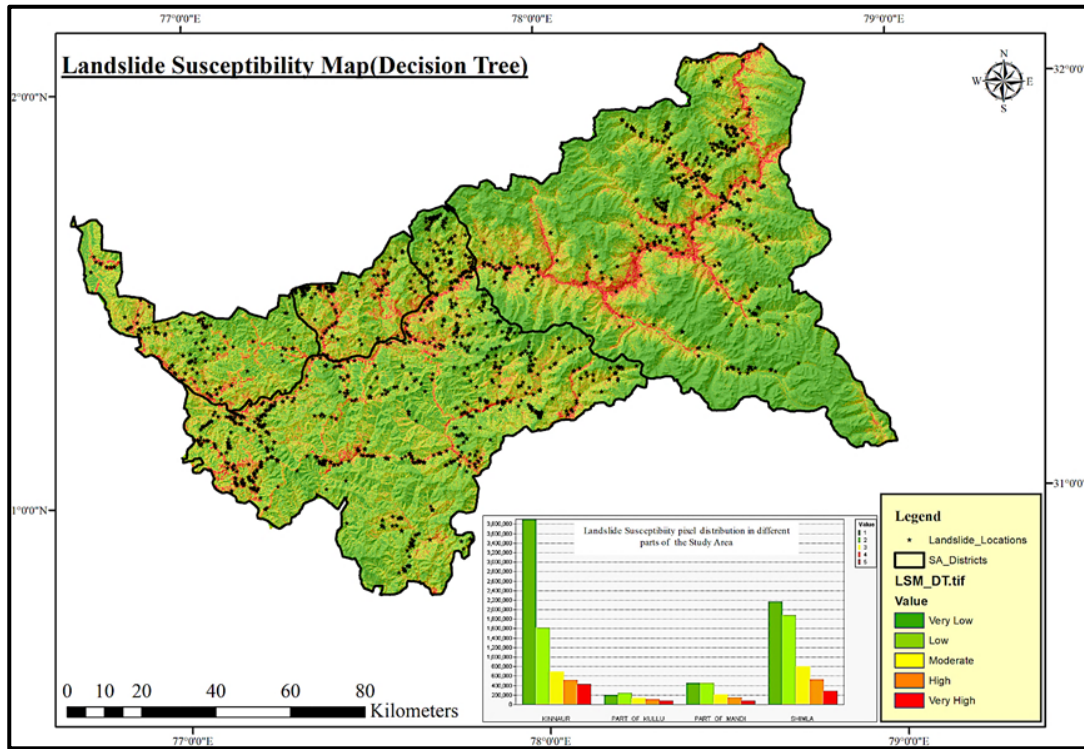
c.



d.



e.



f.

Figure 3.59: Landslide Susceptibility Maps: (a) LR, (b) KNN, (c) SVM, (d) KSVM, (e) NB, (f) DT

To evaluate the spatial reliability of the susceptibility maps generated by the standalone machine learning models, Landslide Density Index (LDI) was employed as a post-classification validation metric. LDI quantifies the concentration of historical landslide occurrences within each susceptibility class and helps assess whether the classified zones reflect actual risk distribution. This evaluation provides additional insight into the effectiveness of the models in spatially identifying landslide-prone areas. The LDI results, along with other spatial validation metrics, are discussed in detail in the next chapter.

3.9 Ensemble Machine Learning Models

Following the implementation of standalone machine learning models, ensemble ML modelling was performed to improve prediction accuracy by reducing variance and bias. Ensemble models integrate multiple individual models to create a more robust predictive framework. The four primary ensemble techniques used in this study include Bagging, Boosting, Stacking, and Voting. The present work specifically employed Random Forest

for Bagging, XG-Boost for Boosting, Stacked models, and Soft Voting as part of the ensemble methodology. The Bagging (Bootstrap Aggregation) approach is a foundational ensemble learning technique used to enhance model performance by reducing both bias and variance. In this method, the original training dataset is resampled into multiple bootstrap subsets each created by randomly selecting instances with replacement. These subsets are then used to train individual weak learners, typically decision trees, in parallel. The predictions from these learners are subsequently aggregated, often through majority voting (for classification) or averaging (for regression), to generate the final output. This aggregation mechanism ensures improved classification stability and reduced overfitting.

Random Forest, a widely used bagging-based ensemble learning technique, constructs multiple decision trees as shown in Figure 3.60 during training and determines the final output by taking the majority vote of individual trees. Prior to model training, one-hot encoding was applied to categorical variables to ensure an appropriate data representation. Backward elimination with sequential feature selection, cross-validated via random forest, was employed to refine the predictor set. Additionally, hyperparameter tuning was performed using grid search to optimize the number of trees and randomly selected predictors, thereby enhancing the model’s accuracy. The optimization results for randomly selected predictors are presented in Figure 3.61, while the grid search results for hyperparameter tuning are provided in

Table 3.7

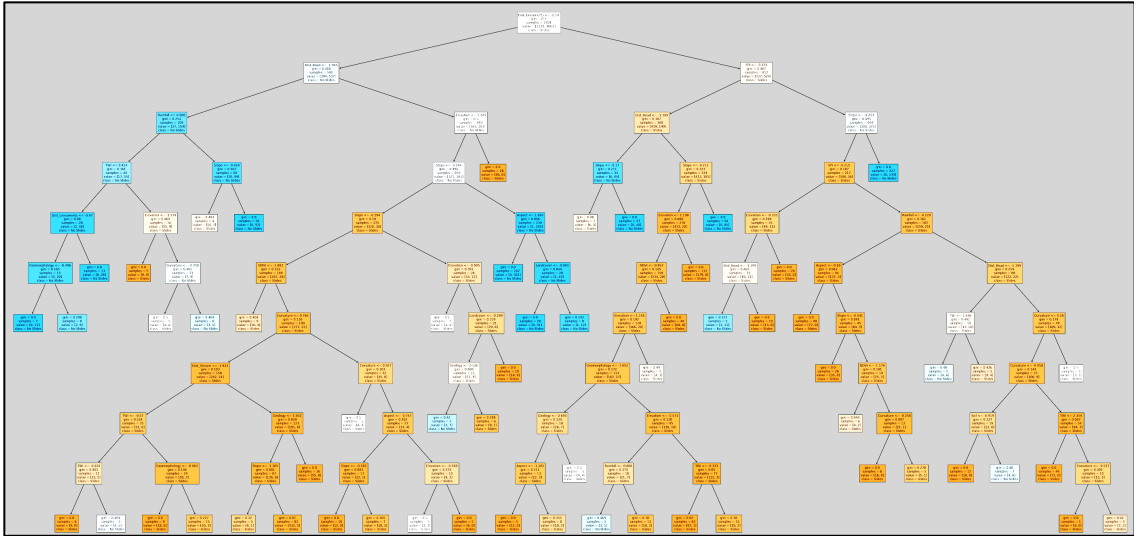


Figure 3.60: Random Forest Trees Visualization

Table 3.7 : Grid Search Results for Hyperparameter Tuning

Hyperparameters	Grid Values
Maximum Depth	10
Minimum Samples Leaf	5
N Estimators	50
N jobs	-1
Random State	42

For model validation, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were computed to evaluate performance, as illustrated in Figure 3.62. Further, additional Bagging algorithms, such as the Extra Trees Classifier and Bagging Classifier, were implemented to compare their accuracy and efficiency against the Random Forest model. The performance comparison of these bagging models is summarized in

Table 3.8.

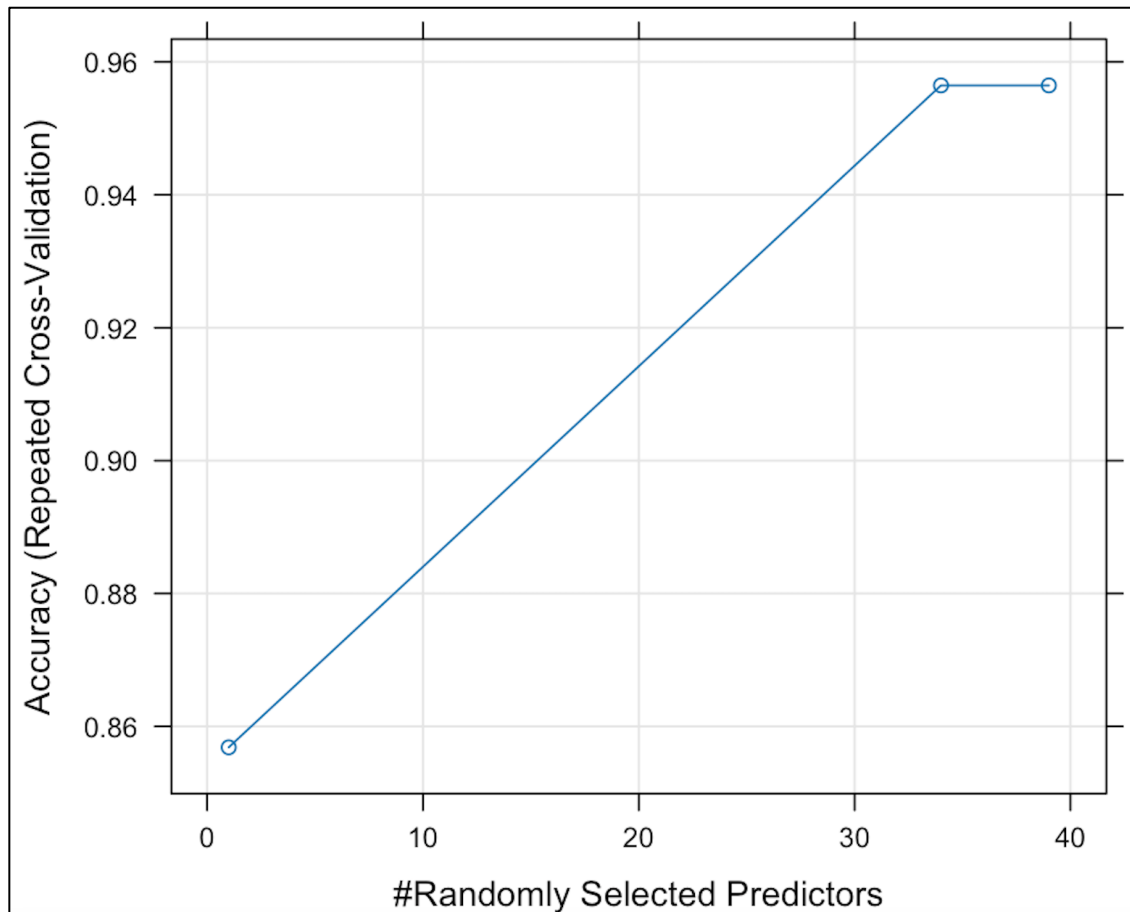


Figure 3.61: Predictor Optimization for Random Forest

Table 3.8 : Performance comparison for different Bagging Models

Model	Accuracy	Recall	Precision
Random Forest	0.95940	0.92291	0.99538
Extra Tree Classifier	0.93376	0.90364	0.96128
Bagging Classifier	0.95727	0.92077	0.99307

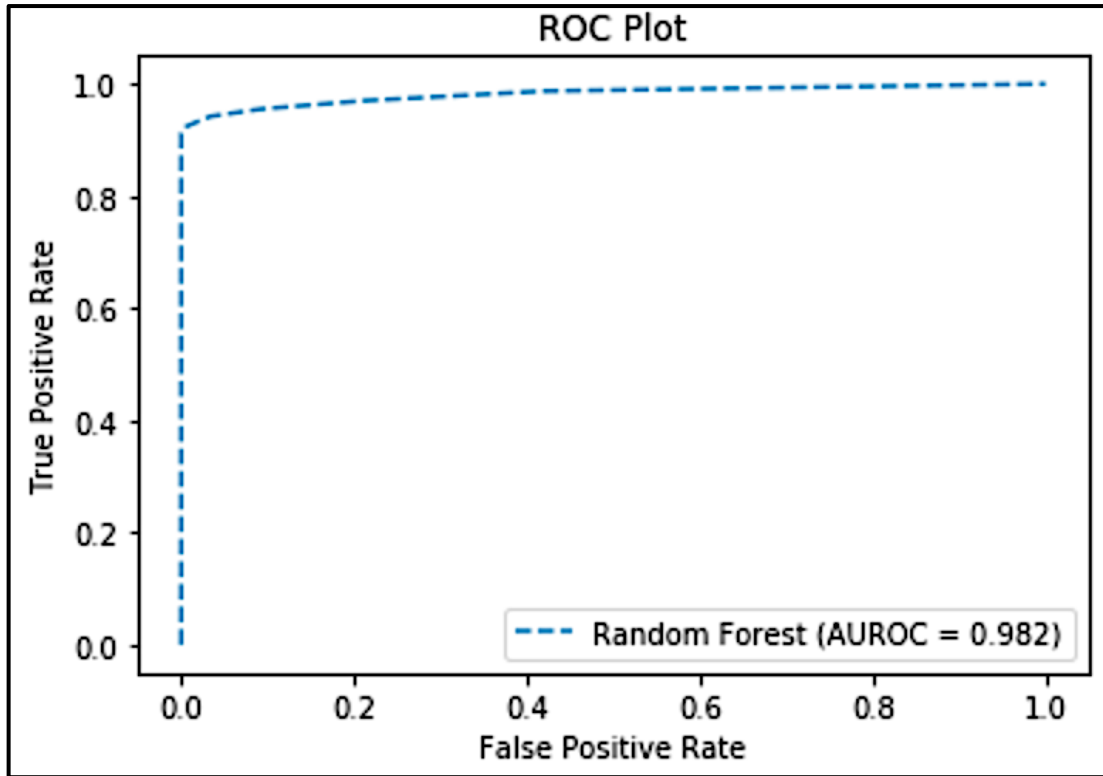


Figure 3.62: ROC and AUC for Random Forest

Further, Landslide Susceptibility Mapping (LSM) was conducted using boosting-based ensemble models, specifically XG-Boost (Extreme Gradient Boosting). XG-Boost is an advanced gradient boosting decision tree algorithm that enhances predictive performance by iteratively constructing weak learners (decision trees) and sequentially improving their accuracy.

The model strengthens the predictive capability by minimizing errors at each iteration, ultimately leading to a highly robust system. To achieve optimal performance, hyperparameter tuning was performed using grid search, identifying the best parameter values to maximize accuracy. The selected hyperparameters and their optimal values after tuning are presented in Figure 3.63 and Table 3.9.

Table 3.9 : Selected hyperparameters for the XGBoost model

Hyperparameters	Grid values
Max depth	(6,18,22)
eta	(0.05,0.3,1)

Hyperparameters	Grid values
gamma	(0,0.01,0.1)
Col sample by tree	(0.75,1)
min_child_weight	(0,1,2)
subsample	(0.5,1)

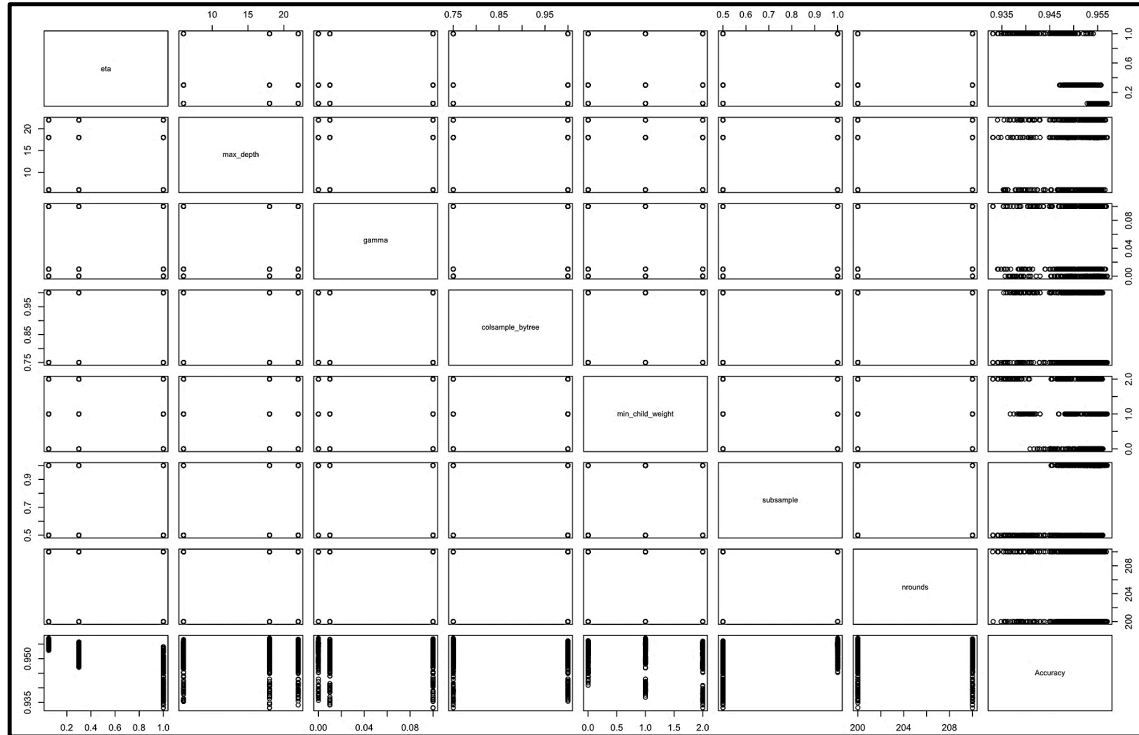


Figure 3.63: Hyperparameter tuning (XG-Boost)

The Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) were computed for validation purposes, as illustrated in Figure 3.64. Additionally, to further evaluate and compare the predictive performance of different boosting techniques, alternative boosting algorithms, including the AdaBoost Classifier and Gradient Boost Classifier, were also implemented. These models were assessed based on their classification accuracy and overall effectiveness in predicting landslide susceptibility. A comparative analysis of these boosting-based models alongside XG-Boost is summarized in Table 3.10.

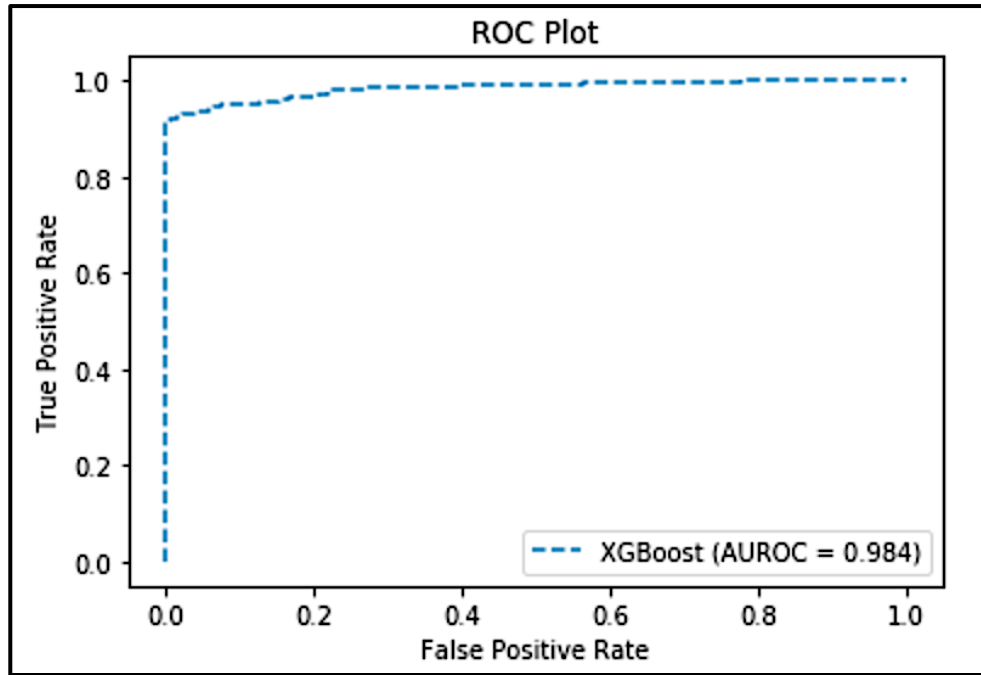


Figure 3.64: ROC Plot for XG Boost

Table 3.10 : Performance comparison for Boosting Models

Model	Accuracy	Recall	Precision
XGB Classifier	0.95510	0.91860	0.99080
AdaBoost Classifier	0.95090	0.91430	0.98610
Gradient Boost	0.95410	0.92720	0.97960

Additionally, LSM was generated using Stacking and Soft Voting ensemble techniques. Stacking represents a more advanced ensemble strategy, combining the predictive power of multiple base classifiers through a meta-learning framework. In this method, several base models (also known as level-0 learners)—such as Decision Trees (DT), Support Vector Machines (SVM), and Logistic Regression (LR) are first trained on the same dataset. Each of these models makes independent predictions, which are then used as input features for a higher-level meta-model (level-1 learner). This meta-model learns how to best combine the base models' outputs to produce a final, more accurate prediction.

The Soft Voting followed by Stacking, combines the predictions of multiple classifiers to enhance overall model accuracy. In this approach, several base models are trained

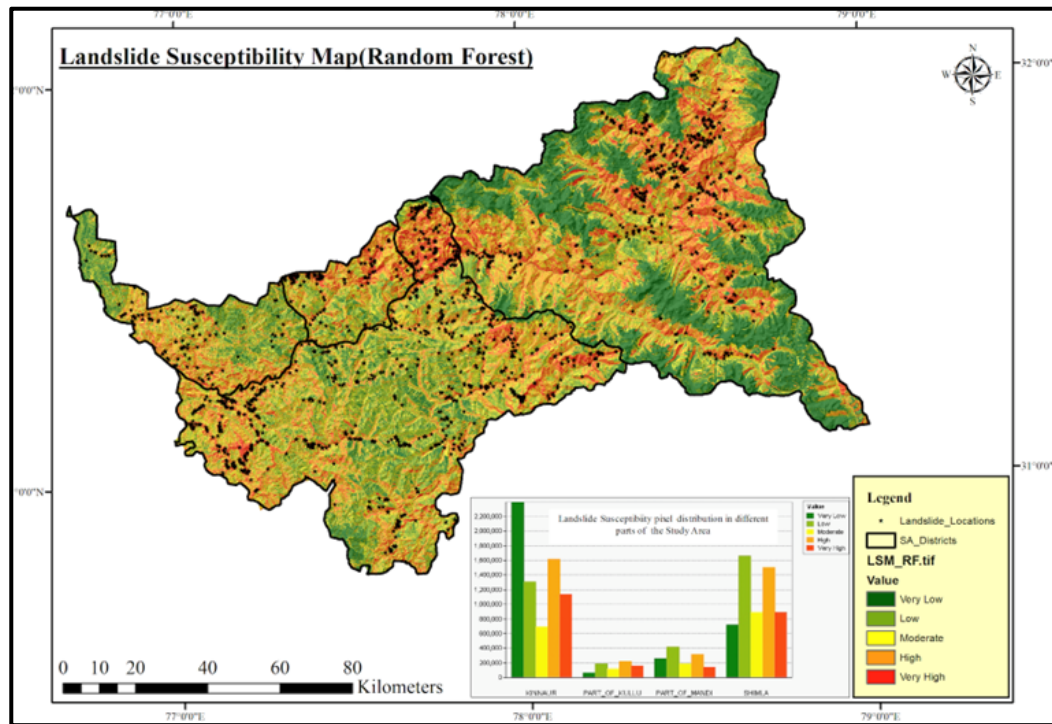
independently on the same dataset, and their predictions are then aggregated to form a final decision.

Soft Voting mechanism, wherein instead of choosing the majority class label directly (as done in hard voting), the ensemble computes the average predicted probability for each class from all participating models. The final classification is based on the class with the highest aggregated probability. This method allows for a more nuanced and balanced decision, particularly when the individual classifiers vary in strength and confidence. By integrating outputs from multiple diverse learners, soft voting reduces the risk of overfitting or bias introduced by any single model. For evaluation, their predictive performance was compared against Random Forest and XG-Boost models. The comparative performance analysis of these methods is summarized in Table 3.11.

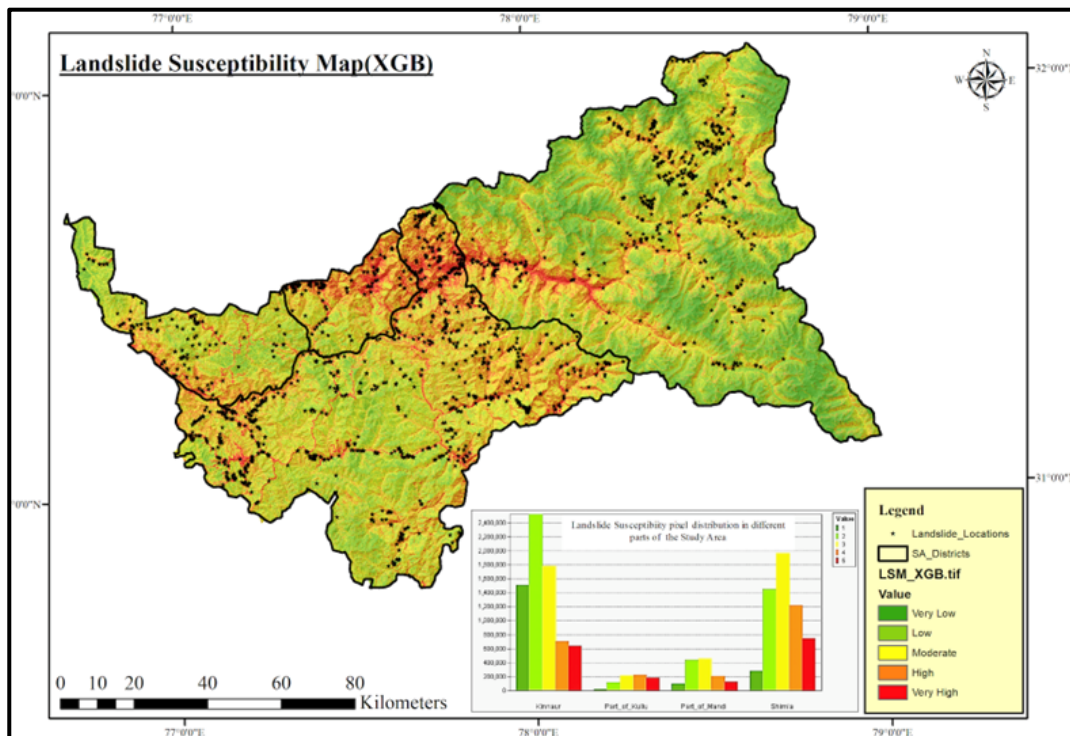
Table 3.11 : Performance Comparison for Voting, Stacking, Boosting and Bagging algorithms

Model	Accuracy	Precision	Recall
Soft Voting	0.9541	0.9796	0.9272
Stacked	0.9573	0.9908	0.9229
XG Boost	0.9551	0.9908	0.9186
Random Forest	0.9594	0.9954	0.9229

The landslide susceptibility maps (LSMs) for Random Forest and XGBoost were developed using probability-based classification at each pixel. The resulting raster datasets were subsequently reclassified into five distinct susceptibility categories (ranging from very low to very high) using an appropriate classification scheme. These maps are illustrated in Figure 3.65.



a.



b.

Figure 3.65: Landslide Susceptibility Maps (a)Random forest and (b) XG-Boost

A comprehensive statistical evaluation of all six standalone ML models and ensemble models, including Random Forest and XG-Boost, was performed to assess their effectiveness in landslide prediction. The results of this analysis are visually represented in Figure 3.66, while the corresponding statistical findings are detailed in Table 3.12.

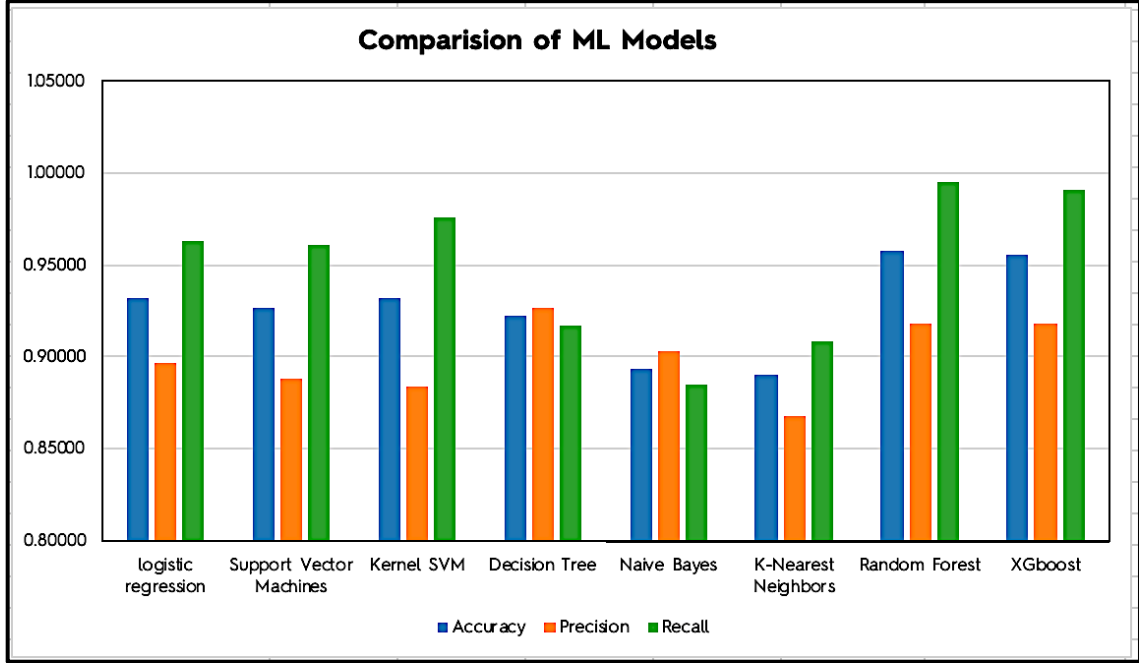


Figure 3.66: Comparison for different ML Models

Table 3.12 : Performance comparison for standalone and Ensemble ML Models

Model	Accuracy	Recall	Precision
Logistic Regression	0.93162	0.89722	0.96322
Support Vector Machine	0.92628	0.8865	0.96065
Kernel SVM	0.93162	0.88437	0.97636
Decision Tree	0.92201	0.92729	0.91737
Naïve Bayes	0.89316	0.90364	0.88470
K-Nearest Neighbors	0.88996	0.86724	0.90807
Random Forest	0.95727	0.91863	0.99536
XG Boost	0.95513	0.91863	0.99076

3.10 Artificial Neural Network Approach for LSM

An attempt was made to generate Landslide Susceptibility Maps (LSM) using Artificial Neural Networks (ANN) to evaluate whether they outperform ensemble models in terms of predictive accuracy.

Artificial Neural Networks (ANNs) are computational frameworks inspired by biological neural networks, designed to detect patterns and learn from complex, non-linear relationships among input variables. The criteria and hyperparameters selected for ANN modelling are detailed in Table 3.13. Graphical representation for the model is also shown in the Figure 3.67.

Table 3.13 : Criteria Details for ANN Modelling

Criteria	Details
Number of Hidden Layer Selected	1
Number of Perceptron Selected	5
Activation Function used	Sigmoid Function (As probability required for the susceptibility modelling)
Method of updating weights	Stochastic Gradient Descent (Reinforcement Learning – updating weights at each observation)
Number of Epoch	100

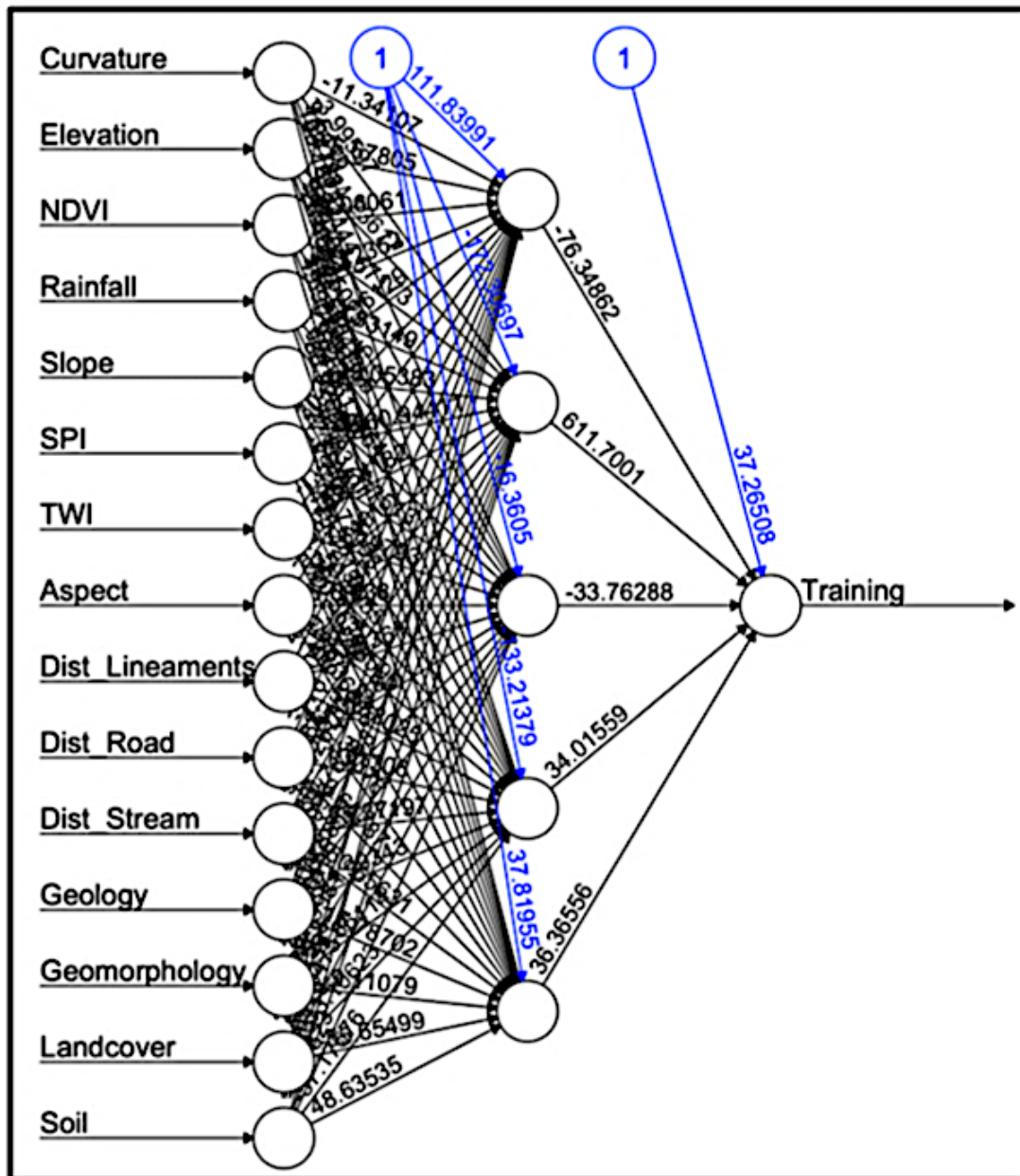


Figure 3.67: Graphical Representation for the Artificial Neural Network Approach

To assess model's performance, accuracy and loss metrics were monitored over multiple training epochs. As depicted in Figure 3.68 and Figure 3.69, accuracy progressively increases while loss decreases as the number of epochs increases for both the training and validation datasets, indicating effective learning and model convergence.

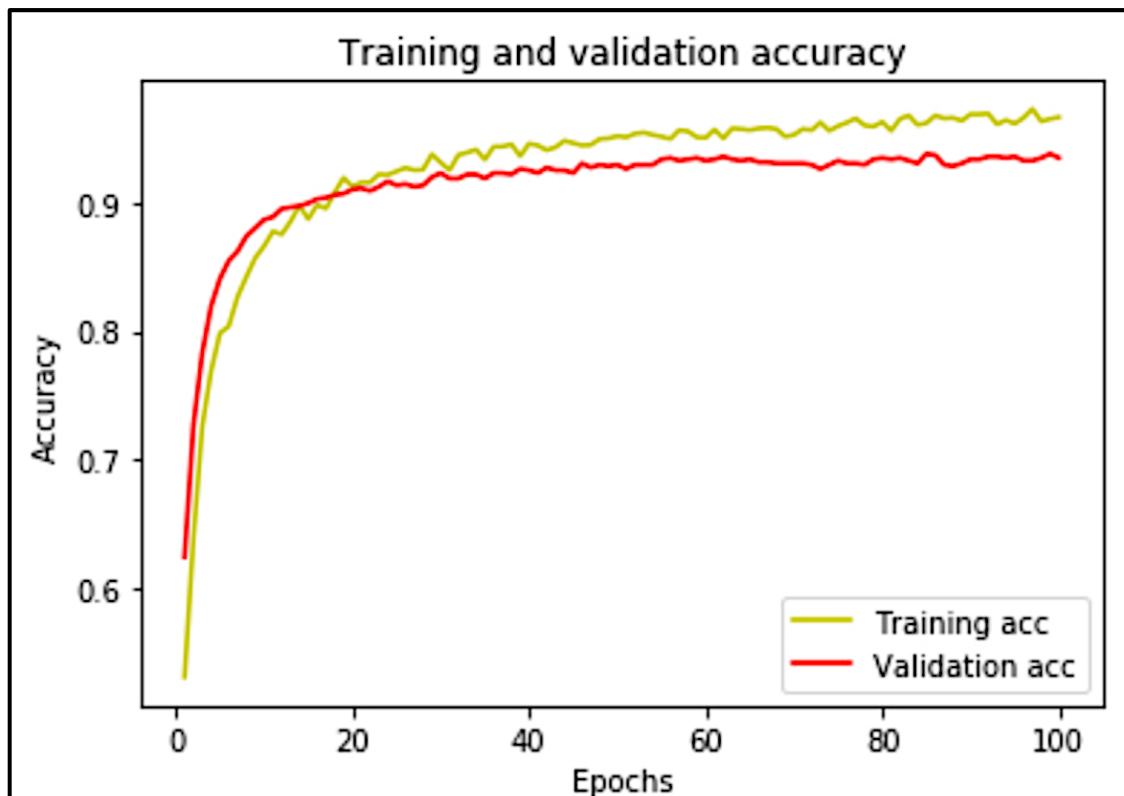


Figure 3.68: Training Accuracy vs. ANN Epoch

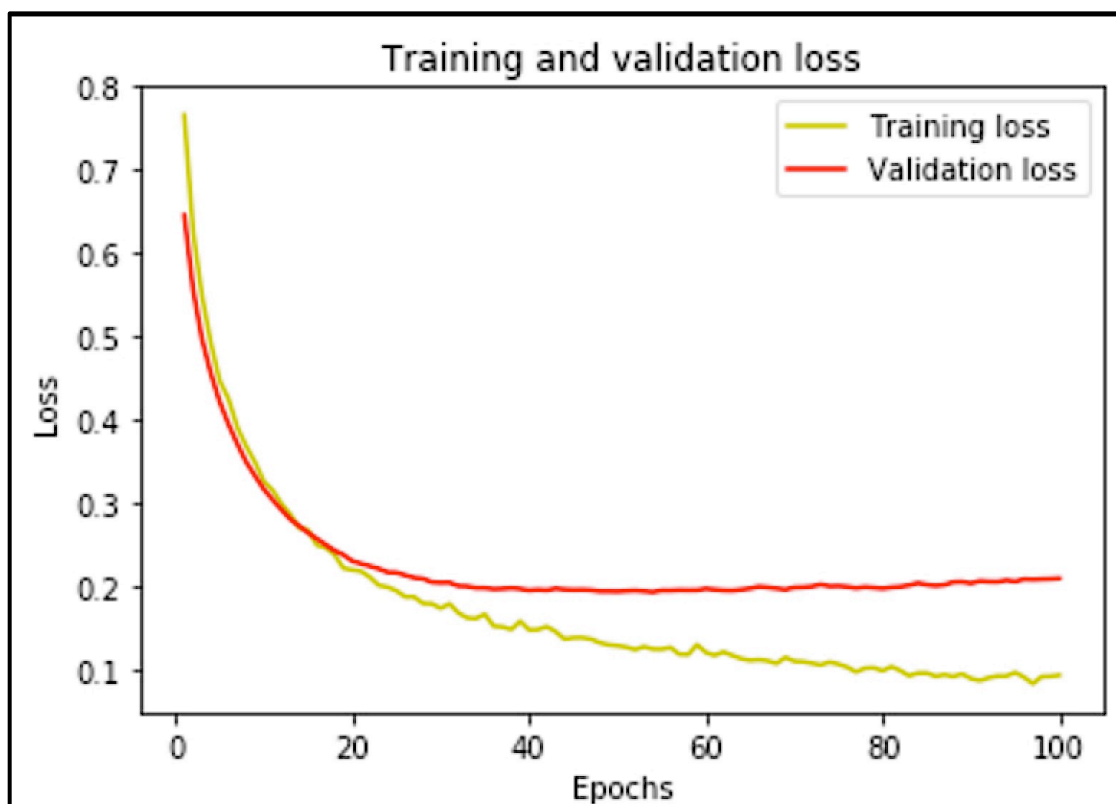


Figure 3.69: Training Loss vs. ANN Epochs

At the conclusion of ANN modelling, the confusion matrix indicated an accuracy of 93.48%, a recall value of 89.50%, and a precision of 97.44%, demonstrating the model's effectiveness in classifying landslide-prone and non-landslide areas. The Receiver Operating Characteristic (ROC) curve and the corresponding landslide susceptibility map (LSM) generated using ANN are presented in Figure 3.70 and Figure 3.71 respectively, offering a visual representation of the model's predictive capability.

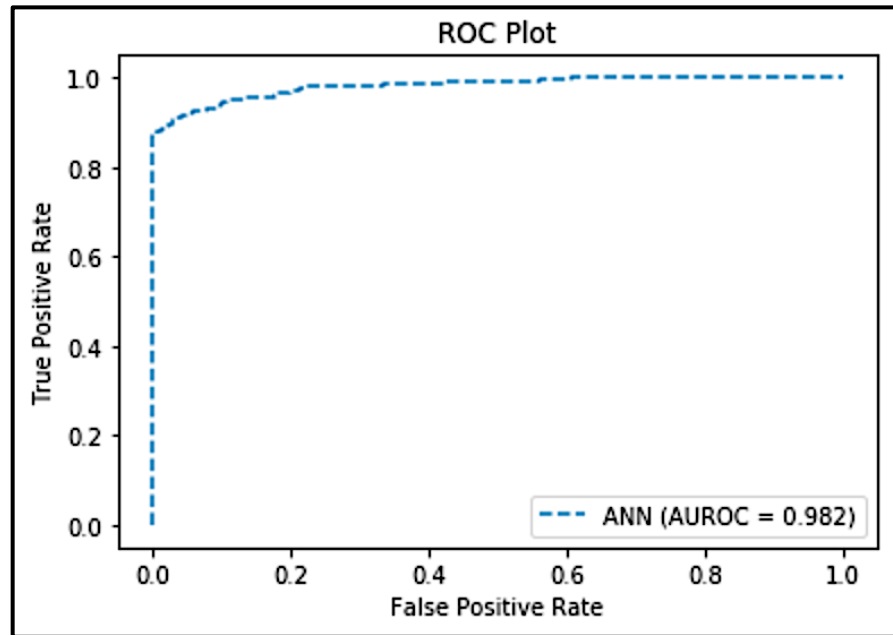


Figure 3.70: ROC Plot for ANN

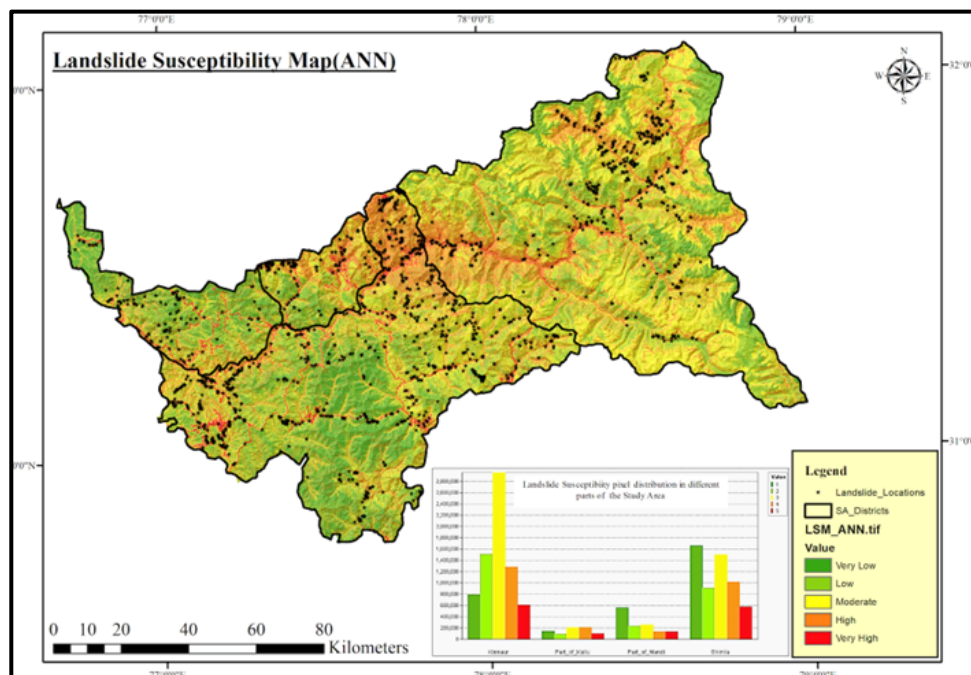


Figure 3.71: Landslide susceptibility map generated with ANN

CHAPTER 4

DISCUSSION ON MODEL OUTPUTS

4.1 Performance Analysis for GIS based Model: AHP and FR

The higher performance of FR can be attributed to its ability to directly utilize statistical relationships between landslide occurrences and causative factors. The FR method assigns weights based on observed landslide distributions, ensuring that conditioning factors with higher landslide occurrences receive greater influence in the model. This data-driven approach enhances its predictive accuracy. In contrast, the AHP method relies on expert judgment for weight assignment, making it more subjective. The lower AUC and LD values in AHP suggest that despite being a structured multi-criteria decision-making approach, it lacks adaptability in capturing complex, nonlinear relationships between conditioning factors and landslides. Additionally, the discretization of numerical variables in AHP may introduce information loss, reducing model precision.

Further analysis was conducted by comparing the percentage of class pixels and percentage of landslide pixels for both FR and AHP models. The plotted values shown in **Figure 4.1** for different susceptibility classes reveal distinct patterns in classification behaviour. The shape of both curves indicates that the FR model exhibits less non-linearity in classifying the index map, leading to a more structured and consistent susceptibility distribution. In contrast, the AHP model demonstrates a higher degree of non-linearity, suggesting greater variability in susceptibility classification. This observation aligns with the inherent methodological differences between the two models. FR follows a statistical approach that maintains proportionality in class distribution, whereas AHP relies on hierarchical weightings that may introduce inconsistencies, leading to greater fluctuations in classification outputs.

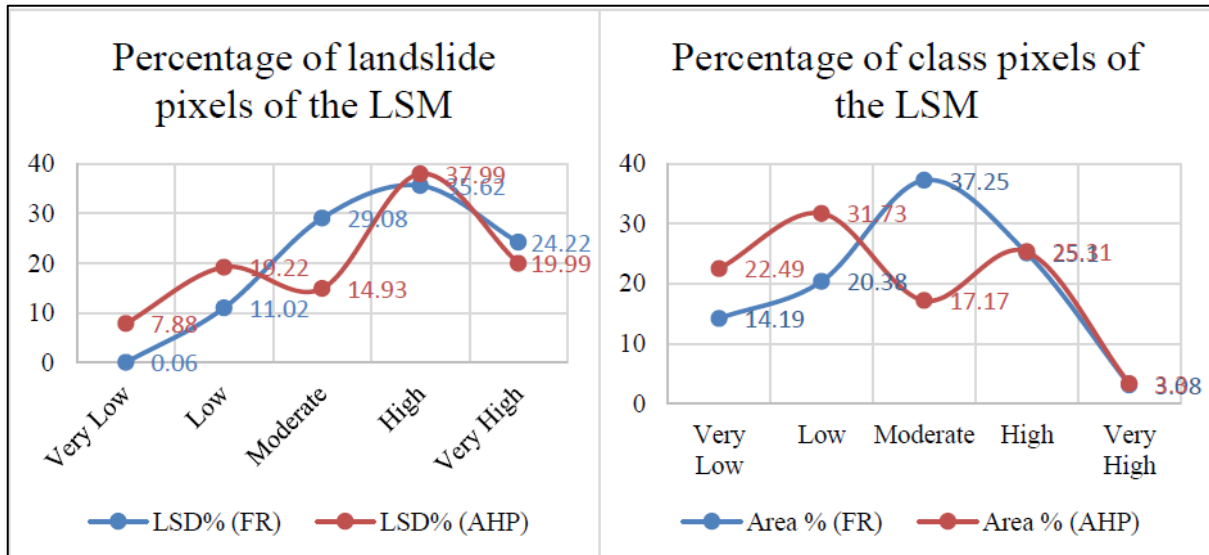


Figure 4.1: Percentage of Landslide Pixels and Class Pixels for AHP and FR

The comparative performance of these GIS-based models highlights a fundamental limitation; they rely on predefined assumptions rather than learning from data patterns. This limitation becomes particularly evident in heterogeneous terrains where landslide susceptibility is governed by complex interactions among geological, hydrological, and geomorphological variables. The observed discrepancies between FR and AHP reinforce the need for data-driven machine learning models, which can dynamically learn from the dataset and improve prediction accuracy.

4.2 Performance Evaluation and Statistical Validation of Machine Learning Models

To ensure a robust evaluation of the models used in this research, various performance metrics were analysed, including accuracy, precision, recall, Cohen's Kappa, and AUC-ROC. This study employs AUROC (Area Under the Receiver Operating Characteristic Curve) and the Kappa (K) score to assess the predictive capabilities of various standalone as well as ensemble machine learning models as shown in Table 4.1 and Figure 4.2.

Random Forest emerges as the top-performing model with the highest accuracy (95.94%), excellent specificity (99.57%), and precision (99.54%), as well as the highest Kappa coefficient (0.91879). This indicates exceptional agreement between predictions and actual landslide occurrences, with almost perfect identification of non-landslide areas. XGBoost follows

closely with 95.51% accuracy and the highest AUC (0.984), suggesting superior discriminative capability. Its performance metrics are very similar to Random Forest, with slightly lower sensitivity (91.86% vs. 92.29%) but still excellent specificity (99.15%) and precision (99.08%).

Amongst all standalone models, Decision Tree shows the highest overall performance with an accuracy of 93.27%, the highest sensitivity (94.22%), and the highest Kappa coefficient (0.86539), indicating excellent agreement between predictions and actual values. Despite having a slightly lower AUC (0.933) compared to other models, its balanced performance across metrics makes it particularly effective for landslide susceptibility modelling. However, its specificity (92.32%) is lower than ensemble methods. Kernel SVM and Logistic Regression both achieved the same accuracy (93.16%) and comparable Kappa values (0.86322). However, Kernel SVM has the highest AUC value (0.98) and specificity (97.87%), suggesting superior ability to distinguish between classes and particularly strong performance in correctly identifying non-landslide areas. Support Vector Machine (standard) shows good performance with 91.99% accuracy and high specificity (97.01%), but its sensitivity (86.94%) is somewhat lower than other top models, indicating it misses slightly more actual landslide cases. Naive Bayes has a moderate accuracy (89.32%) but demonstrates the second-highest sensitivity (90.36%), making it potentially valuable when detection of positive cases (landslides) is prioritized over overall accuracy. K-Nearest Neighbour shows the lowest overall performance with 89.00% accuracy and lowest Kappa (0.77989), suggesting it is less suitable for this landslide susceptibility modelling task compared to other algorithms.

As far as cross validation is concern, XG-Boost shows excellent stability with the lowest standard deviation in cross-validation (1.26%), indicating consistent performance across different subsets of data. Random Forest also demonstrates good stability (1.36% standard deviation). Logistic Regression has similar stability (1.17%). Naive Bayes has the highest standard deviation (3.14%), suggesting less stable performance across different data partitions.

Table 4.1 : Statistical Evaluation Parameters of ML Models

No	Metric	LR	KNN	SVM	Kernel SVM	Naïve Bayes	Decision Tree	Random Forest	XG-Boost
1	TN	453	428	455	459	414	433	467	465
2	TP	419	405	406	413	422	440	431	429
3	FN	48	62	61	54	45	27	36	38
4	FP	16	41	14	10	55	36	2	4
5	Accuracy Rate (AR)	93.16%	89.00%	91.99%	93.16%	89.32%	93.27%	95.94%	95.51%
6	Error Rate (ER)	6.84%	11.00%	8.01%	6.84%	10.68%	6.73%	4.06%	4.49%
7	Accuracy - K-fold cross validation	92.80%	89.00%	92.90%	92.80%	89.32%	91.66%	94.64%	94.87%
8	Sd - K-fold cross validation	1.17%	2.37%	1.76%	1.87%	3.14%	1.52%	1.36%	1.26%
9	Sensitivity (SST)/Recall	89.72%	86.72%	86.94%	88.44%	90.36%	94.22%	92.29%	91.86%
10	Specificity (SPF)	96.59%	91.26%	97.01%	97.87%	88.27%	92.32%	99.57%	99.15%
11	Positive Predictive Value (PPV)/Precision	96.32%	90.81%	96.67%	97.64%	88.47%	92.44%	99.54%	99.08%
12	Negative Predictive Value (NPV)	90.42%	87.35%	88.18%	89.47%	90.20%	94.13%	92.84%	92.45%
13	Kappa (K)	0.863	0.780	0.840	0.863	0.786	0.865	0.919	0.910
14	AUC	0.975	0.949	0.974	0.980	0.960	0.933	0.982	0.984

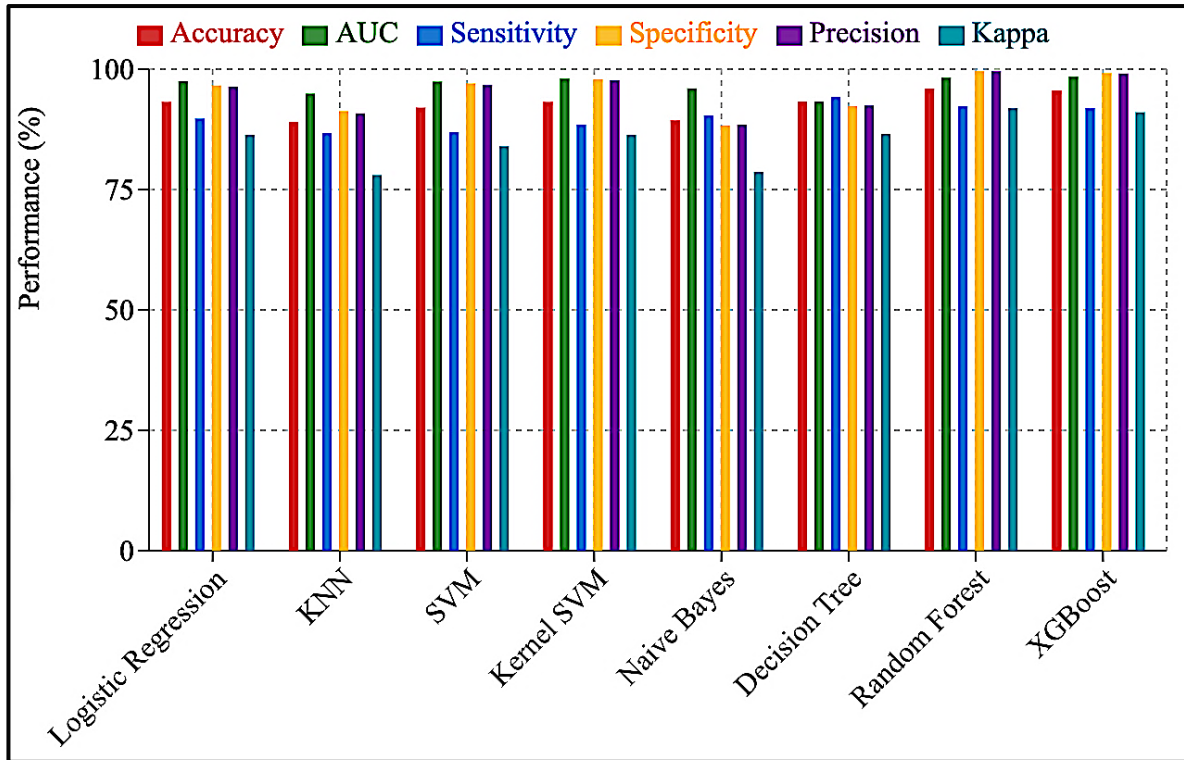


Figure 4.2: Statistical evaluation of the confusion matrix parameter for ML Models

From a theoretical perspective, model performance assessment in classification tasks relies on evaluating bias-variance trade-offs. Decision Tree models, while simple and interpretable, often exhibit high variance, making them prone to overfitting. However, ensemble techniques such as Bagging (Random Forest) mitigate this issue by aggregating predictions from multiple decision trees, thereby reducing variance and improving stability. This explains why Random Forest outperformed individual decision trees in this study. Boosting algorithms like XGBoost, on the other hand, iteratively improve weak learners by adjusting weights, making them highly effective in capturing complex patterns. The superior performance of ensemble models compared to standalone models aligns with theoretical expectations, as ensemble learning leverages multiple weak models to produce a more robust predictive model.

Furthermore, AUC-ROC analysis provided additional insights into model efficacy. A higher AUC value indicates better discriminatory power, meaning the model is more capable of correctly distinguishing between landslide and non-landslide zones. The results from this study confirm that ensemble models, particularly Random Forest and XGBoost, achieved higher AUC values, underscoring their ability to generate reliable landslide susceptibility maps. The confusion matrix analysis also demonstrated lower false positive and false negative rates for ensemble models, reinforcing their effectiveness in real-world applications.

The performance evaluation of the Random Forest model demonstrates its robustness in classification tasks. The model achieved a high accuracy of approximately 95.94%, indicating its strong predictive capability. Furthermore, the specificity of 99.57% suggests that the model effectively identifies negative instances, minimizing false positives. The sensitivity, or recall, is recorded at 92.29%, highlighting the model's effectiveness in correctly detecting positive cases. Additionally, Cohen's Kappa score of 0.9187 reflects a strong agreement between predicted and actual classifications, reinforcing the model's reliability beyond mere chance. The application of k-fold cross-validation further supports these findings, with an average accuracy of 94.64% and a standard deviation of 1.36%, ensuring consistent performance across different subsets of the dataset. These metrics collectively affirm the Random Forest model's suitability for classification tasks requiring high accuracy and reliability. In this study, K-fold cross-validation was employed to ensure the robustness and reliability of the machine learning models used for landslide susceptibility mapping. Given the challenges associated with imbalanced datasets in landslide modelling, relying on a single train-test split could lead to biased performance evaluation. To mitigate this, K-fold cross-validation was applied, where the dataset was systematically divided into multiple subsets, allowing each model to be trained and tested multiple times on different data partitions. This process helped in avoiding overfitting by ensuring that the models did not memorize patterns specific to a single dataset split but instead generalized well across different data subsets.

Additionally, K-fold cross-validation provided a more stable performance evaluation by computing the average accuracy, precision, recall, and standard deviation across multiple iterations. This was particularly useful in comparing the effectiveness of different models, including Random Forest, XG-Boost, Logistic Regression, and Artificial Neural Networks (ANN). The results indicated that ensemble models such as Random Forest and XG-Boost consistently outperformed standalone models, with lower variance across folds and higher predictive accuracy. Furthermore, this approach helped in addressing data imbalance by ensuring that each fold contained a representative mix of landslide-prone and stable regions, leading to unbiased model evaluation. By implementing K-fold cross-validation, the study strengthened the credibility of its findings and confirmed that the observed model performance was not a result of dataset-specific biases. The comparative analysis further reinforced that tree-based ensemble models exhibited higher stability and reliability, making them more suitable for landslide susceptibility mapping.

4.3 Comparative Analysis of LSM in Uttarakhand Region

A study [162] conducted in the Uttarkashi district employed the Random Forest (RF) algorithm, achieving an Area Under the Curve (AUC) of 0.89, with training and testing accuracies of 96% and 86%, respectively. The final susceptibility map classified the region into Low, Moderate, and High susceptibility zones, demonstrating the model's strong predictive capabilities. When compared to the present study, RF also emerged as one of the best-performing models, exhibiting high accuracy, precision, and AUC values. The consistency in the RF model's performance across different studies suggests its robustness and adaptability to varying terrains. However, minor differences in classification outcomes can be attributed to variations in geological features, topographic conditions, and dataset resolutions.

Further comparative analysis in Uttarakhand assessed landslide susceptibility using Statistical Evaluation (SE), Fuzzy-Analytic Hierarchy Process (Fuzzy-AHP), Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XG-Boost). The study [163] found that XG-Boost classified approximately 73.84% of the area as having low to very low susceptibility, with RF and LR following closely in performance. Similar findings were observed in the present study, where XG-Boost and RF outperformed traditional statistical models. However, while the Uttarakhand study emphasized the strength of XG-Boost in classifying large portions of the region under low susceptibility, the present research highlighted the ability of RF and XG-Boost to provide a more balanced classification across susceptibility zones. The variations in susceptibility zoning further indicate that model performance is highly dependent on feature selection, input variables, and study area characteristics.

A study [164] conducted in the Mussoorie region applied a bivariate statistical approach to assess landslide susceptibility, emphasizing the role of key terrain factors such as slope gradient, lithology, and land use. The susceptibility map provided a detailed classification of landslide-prone areas based on statistical correlations between conditioning factors and historical landslide occurrences. In comparison, the present study utilized advanced machine learning-based feature selection techniques, ensuring that only statistically significant variables were retained for modelling. While slope, geology, and land use remained critical contributors in both studies, additional parameters such as distance to streams, soil type, and rainfall variability were incorporated in the present research, leading to improved predictive accuracy. This demonstrates the advantage of integrating data-driven machine learning approaches over traditional statistical methods, allowing for a more refined and adaptive susceptibility assessment.

Another study [165] in the Himalayan region of Uttarakhand developed spatial predictive models considering both rainfall-induced and earthquake-triggered landslides. The research employed relative frequency ratio methods to evaluate susceptibility, emphasizing the compounded effects of multiple triggering mechanisms. The findings underscored the importance of incorporating dynamic environmental factors for more comprehensive landslide risk assessments. The present study aligns with this perspective by integrating multiple conditioning factors, including historical landslide events, rainfall, and geomorphological variations. However, unlike the Uttarakhand study, which primarily focused on natural triggers such as rainfall and earthquakes, the present research extends the scope by considering anthropogenic influences, such as distance to roads and landcover. This broader approach enhances model adaptability, allowing it to account for evolving environmental and human-induced conditions.

The comparison between previous studies in Uttarakhand and the present research highlights key trends in landslide susceptibility modelling. The superior performance of machine learning models, particularly RF and XG-Boost, is evident across different studies, demonstrating their ability to capture complex, nonlinear relationships between conditioning factors and landslide occurrences. However, variations in model accuracy and susceptibility classification emphasize the importance of region-specific geological and environmental conditions. Feature selection plays a crucial role in refining model performance, as studies incorporating additional terrain attributes and climate-related parameters tend to yield more precise results. Moreover, the integration of multi-trigger factors, including both natural and human-induced influences, is becoming increasingly essential for accurate landslide prediction. By aligning with best practices observed in previous research while incorporating new methodological enhancements, the present study contributes to the ongoing advancements in landslide susceptibility assessment, ensuring a more comprehensive and reliable predictive framework.

4.4 Performance Evaluation of Artificial Neural Networks (ANN) in Landslide Susceptibility Modelling

Artificial Neural Networks (ANN) were also employed in this study to assess their effectiveness in landslide susceptibility mapping. The results indicate that while ANN achieved better accuracy and precision than standalone ML models, it did not surpass the performance of ensemble techniques such as Random Forest and XG-Boost.

From a theoretical standpoint, ANN models are inspired by biological neural networks and are highly effective in modelling complex, nonlinear relationships. Unlike traditional ML algorithms, ANNs rely on multiple layers of interconnected neurons, which enable them to learn high-level feature representations. However, the effectiveness of ANNs is heavily dependent on factors such as network architecture, hyperparameter tuning, and dataset size. The accuracy and loss curves from this study show that as the number of epochs increased, training and validation accuracy improved while loss values decreased, indicating successful learning.

The Artificial Neural Network (ANN) model demonstrated strong classification performance in landslide susceptibility modelling, as reflected in various evaluation metrics. The model achieved an overall accuracy of 93.48%, indicating a high degree of reliability in correctly identifying both landslide-prone and non-landslide-prone areas. The specificity of 97.44% suggests that the model effectively minimizes false positives, making it highly reliable in correctly classifying negative instances. Additionally, the sensitivity of 89.50% highlights its capability to capture true positive cases, ensuring that most landslide-prone areas are correctly identified.

To further assess the model's consistency, Cohen's Kappa score was calculated, yielding a value of 0.8696, which signifies a strong agreement between predicted and actual classifications beyond mere chance. These results indicate that the ANN model offers a balanced trade-off between specificity and sensitivity, making it a suitable choice for landslide susceptibility assessment. Given its ability to minimize false positives while maintaining a high recall rate, ANN proves to be a robust method for predictive modelling in geospatial hazard analysis.

However, despite its capability to model intricate relationships, ANN did not outperform ensemble ML models in this research. This can be attributed to several factors. First, ensemble models such as Random Forest and XG-Boost inherently reduce overfitting by combining multiple weak learners, whereas ANN models require careful regularization techniques (such as dropout and batch normalization) to prevent overfitting. Second, ANN models generally require larger datasets to achieve optimal performance. While ML models efficiently learn from moderate-sized datasets, deep learning models often require extensive training data to fully leverage their potential. Third, the computational complexity of ANN models is higher compared to ML models, making them less practical for real-time applications in landslide susceptibility mapping. Despite these limitations, ANN remains a promising approach for

future studies, particularly when integrated with hybrid modelling techniques. Combining deep learning with ensemble ML models may yield superior results by leveraging the strengths of both methodologies. For instance, feature extraction using deep learning followed by classification using ensemble methods could enhance prediction accuracy while maintaining computational efficiency.

4.5 Landslide Density Index and Percentage Pixel Distribution

The analysis of Landslide Density Index (LDI) as shown in Figure 4.3 across different machine learning models reveals consistent and meaningful patterns in how susceptibility classes are distinguished. A general trend observed in all models is a progressive increase in LDI values from *Very Low* to *Very High* susceptibility classes, indicating that all models are successfully capturing the expected gradient of landslide risk. This consistency strengthens the reliability of the modelling framework. Notably, Random Forest (RF) exhibited the highest LDI in the *Very High*-risk category (3.79), followed closely by XG-Boost (3.66) and Artificial Neural Network (ANN) (3.56), establishing these as the top-performing models for identifying severe risk zones. Decision Tree (DT) also showed strong results with high LDI values in both the *High* (2.07) and *Very High* (3.53) categories, reflecting its utility in moderate to critical risk ranges. Models like K-SVM, SVM, and KNN displayed moderate performance with *Very High* indices ranging between 2.93 and 3.25, while Logistic Regression (LR) and Naïve Bayes (NB) recorded the lowest values in this category (2.55 and 2.21, respectively), indicating relatively weaker differentiation in high-risk areas. In contrast, when evaluating false positives in *Very Low* categories, RF and KNN again performed best, showing near-zero LDI values (0.00 and 0.02), whereas K-SVM showed a notably high value (0.64), suggesting greater misclassification of low-risk zones.

The LDI trend lines further emphasize these distinctions. Models like RF, XGB, and ANN display exponential growth in LDI values, particularly between *High* and *Very High*-risk classes, highlighting their sensitivity to severe landslide indicators and superior discriminative power. These models demonstrate steep slopes, while K-SVM, SVM, and KNN exhibit moderate increases, and NB shows a comparatively flat curve—especially across higher risk levels—indicating limited resolution between critical categories. Importantly, crossover points were observed in trend behaviour: DT outperformed others in the *Moderate* to *High*-risk range but was surpassed by RF, XGB, and ANN in the *Very High* category. Surprisingly, NB showed competitive performance in *Low* to *Moderate* categories before dropping off in the higher

ranges. The starting points of these curves also provide insight into low-risk classification behaviour. RF and KNN began closest to zero, while K-SVM started significantly higher, again suggesting false positives in stable zones.

These patterns reveal distinct model groupings based on LDI behaviour. The top-performing tier consists of RF, XGB, ANN, and DT, which show strong contrast across susceptibility levels and superior risk classification. The middle tier—K-SVM, SVM, and KNN—offers balanced but less pronounced differentiation. The lower tier, comprising LR and NB, struggles to sharply separate high-risk from low-risk zones. From a risk management perspective, these findings are pivotal. RF, XGB, and ANN are ideal for scenarios prioritizing the accurate identification of high-risk zones, while DT may be preferable for applications requiring balanced performance across all classes. The sharp contrast in LDI values, particularly in high-risk zones, also makes tree-based models (RF, XGB, DT) more suitable for binary decision-making in emergency planning. Detailed results are shown in the Appendix B. Ultimately, the consistent upward trend across models confirms the overall robustness of the susceptibility modelling approach, while the variations between models stress the need for thoughtful model selection and validation, aligned with specific environmental risk assessment goals.

Table 4.2 : Landslide Density Index for different models

LABEL	RF	XGB	ANN	DT	K-SVM	SVM	KNN	LR	NB
Very Low	0.00	0.03	0.37	0.45	0.64	0.31	0.02	0.33	0.05
Low	0.21	0.27	0.49	0.90	0.56	0.43	0.22	0.52	0.87
Moderate	0.53	0.78	0.66	1.33	0.58	0.66	0.50	0.83	0.91
High	1.17	1.69	1.53	2.07	1.14	1.36	0.94	1.13	0.86
Very High	3.79	3.66	3.56	3.53	3.25	3.08	2.93	2.55	2.21

To establish a consistent landslide susceptibility classification, LDI thresholds were derived by calculating the midpoints between adjacent risk classes for each model. The average of these thresholds across all models was then taken to define common class boundaries. The resulting LDI ranges are: Very Low (0.00–0.37), Low (0.37–0.63), Moderate (0.63–1.04), High (1.04–2.31), and Very High (>2.31). These standardized thresholds ensure uniform interpretation across different models.

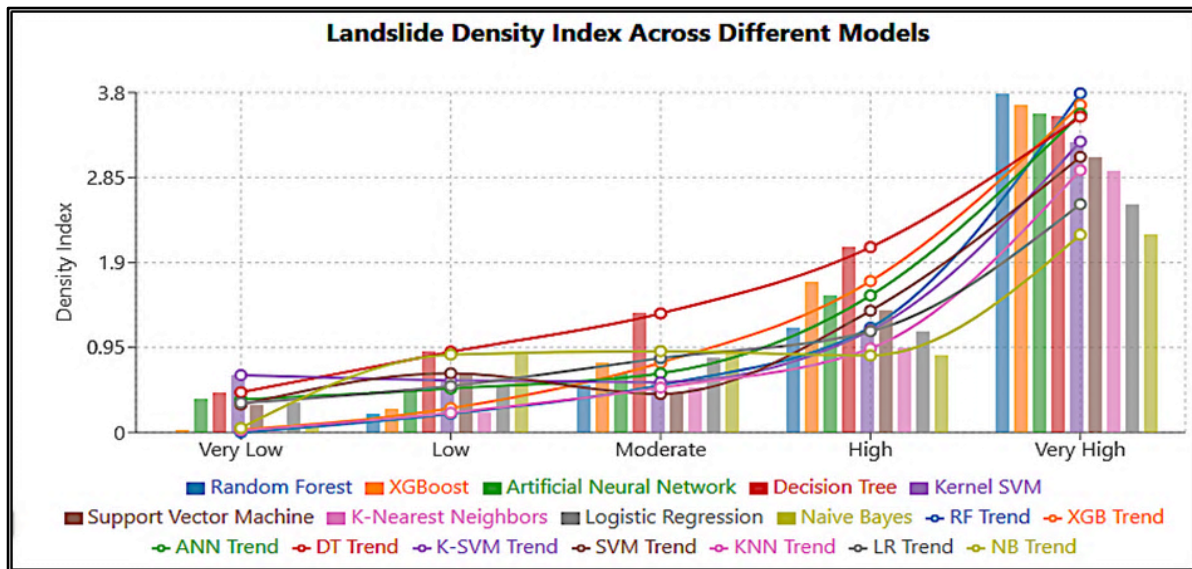


Figure 4.3: Chart for Landslide Density Index for Different Models

To understand the model specific distribution pattern, percentage pixel distribution was studied and chart for the same is shown in Figure 4.4.

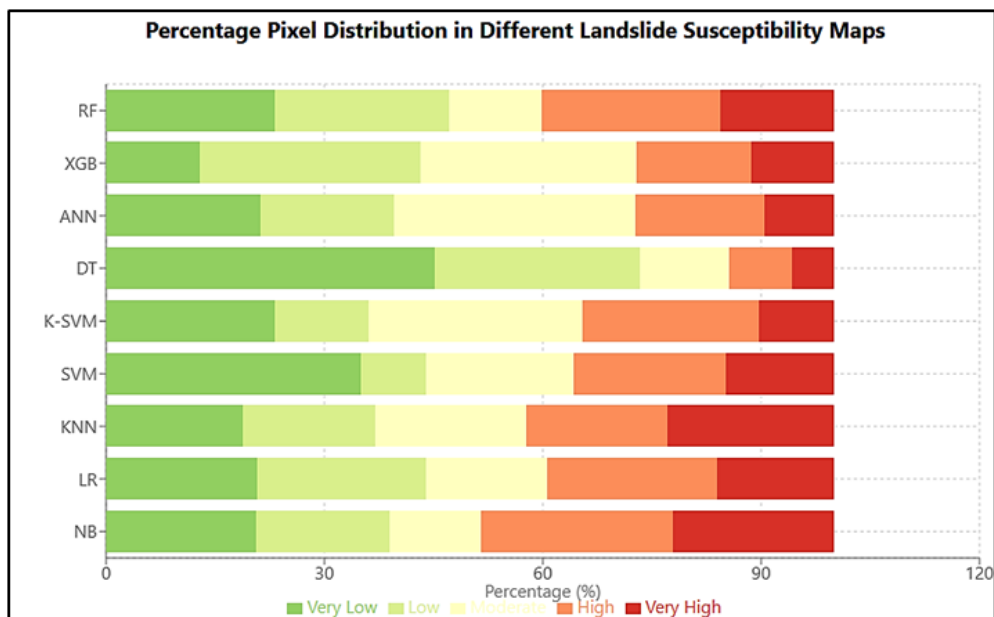


Figure 4.4: Percentage Pixel distribution for Landslide Susceptibility Models

Table 4.3 : Percentage Pixel Distribution for Landslide Susceptibility Models

LABEL	RF	XGB	ANN	DT	K-SVM	SVM	KNN	LR	NB
Very Low	23.12	12.82	21.15	45.16	23.17	34.98	18.82	20.70	20.67
Low	24.02	30.41	18.44	28.16	12.84	8.99	18.13	23.23	18.21
Moderate	12.70	29.67	33.16	12.31	29.44	20.22	20.81	16.65	12.71
High	24.55	15.77	17.76	8.66	24.21	20.92	19.29	23.43	26.25
Very High	15.61	11.33	9.49	5.72	10.34	14.89	22.96	15.98	22.16

The landslide susceptibility classification patterns across different machine learning models revealed significant variability in how each model interprets spatial risk. Decision Tree (DT) showed a highly conservative approach with 45.16% of pixels classified as "Very Low" risk and only 5.72% as "Very High" risk. This suggests DT may be under-predicting high-risk areas compared to other models. K-Nearest Neighbours (KNN) offered the most balanced distribution across all categories and has the second-highest allocation to "Very High" risk areas (22.96%). This suggests a less polarized risk assessment. Naive Bayes (NB) allocated the highest percentage to "High" risk (26.25%) and second-highest to "Very High" risk (22.16%), making it the most conservative model in terms of identifying potential danger zones. Support Vector Machine (SVM) has the second-highest allocation to "Very Low" risk areas (34.98%), indicating a less conservative approach like DT but not as extreme. Artificial Neural Network (ANN) heavily favoured the "Moderate" risk category (33.16%) while assigning the lowest percentage to "Very High" risk (9.49%), suggesting a tendency toward middle-ground classifications.

As far as risk distribution is concerned, NB and KNN are the most conservative models exhibiting highest percentage in high + very high categories. RF, K-SVM, LR, and SVM can be termed as the middle-ground models showcasing more balanced distribution across the susceptibility classes. DT, XGB, and ANN are the least conservative models demonstrating lower percentages in High + Very High categories.

With reference to class imbalance, DT shows the most imbalanced distribution, heavily weighted toward low-risk categories. KNN shows the most balanced distribution across all five categories. XGB and ANN concentrate a large portion of pixels in one category ("Low" and "Moderate" respectively).

From practical implications and risk management strategy point of view, NB and KNN look to be preferable for conservative landslide risk management (where missing high-risk areas can prove to be costly). For balanced assessments, RF and LR provide more evenly distributed classifications. For development-focused applications (where over-predicting risk areas limits land use), DT might be preferred. The significant variation between models suggests potential benefits from ensemble approaches. Areas consistently classified as "Very High" risk across models would warrant particular attention.

This shows that the choice of model significantly impacts the resulting susceptibility map and should be guided by the specific risk management objectives and local landslide characteristics. The cost of false negatives (missed landslide risks) versus false positives (unnecessarily restricted areas) should inform model selection. This analysis demonstrates that different machine learning models produce significantly different landslide susceptibility maps even when trained on the same data, highlighting the importance of model selection and validation in landslide hazard assessment.

4.6 Zonal Distribution of Landslide Susceptibility

A district-wise analysis of landslide susceptibility distribution revealed significant spatial disparities in risk concentration across the four administrative units Kullu (Part), Shimla, Mandi (Part), and Kinnaur. The evaluation was based on multiple indicators, including the landslide and area proportion as shown in Figure 4.5 and Figure 4.6, the percentage of area classified under High and Very High susceptibility. Detailed results are shown in Appendix B.

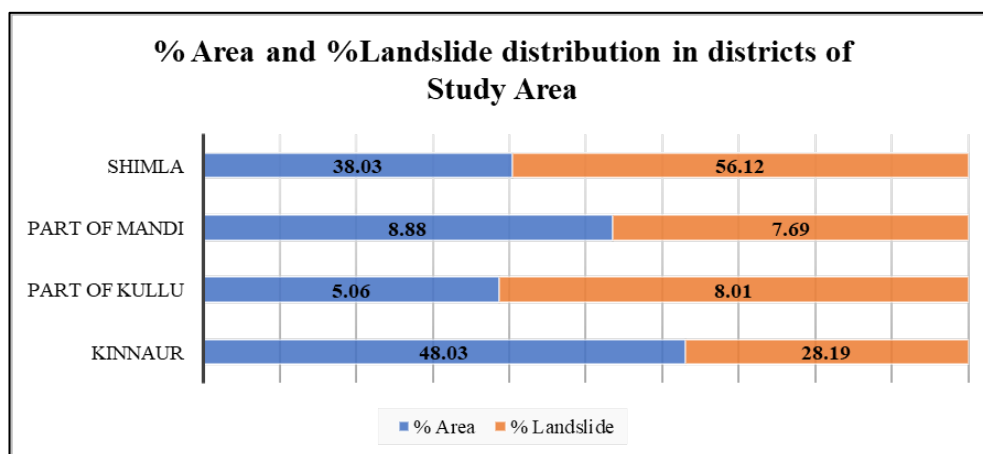


Figure 4.5: Area % and Landslide % Distribution in Study Area

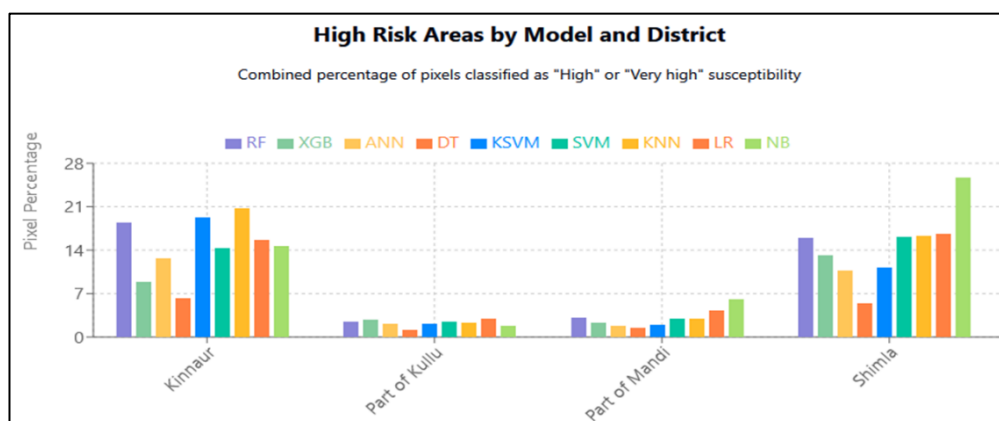


Figure 4.6: High Risk Areas with respect to Districts and Models

Figure 4.7 represents cumulative pixel percentages across 9 machine learning models (RF, XGB, ANN, DT, KSVM, SVM, KNN, LR, NB) for each susceptibility class and a calculated risk index (on a scale of 1–5), which provided a weighted perspective on overall vulnerability. The risk index of different parts of the study area along with zone area percentage, corresponding landslide percentage as well as landslide to area ratio has been graphically represented in Table 4.4 and Figure 4.8.

Table 4.4 : Zone wise landslide-area ratio and risk index

District	Area %	Landslide %	Landslide/Area Ratio	Risk Index
Kinnaur	48.03	28.19	0.58	2.66
Part of Kullu	5.06	8.01	1.58	3.14
Part of Mandi	8.88	7.69	0.86	2.76
Shimla	38.03	56.12	1.47	2.93

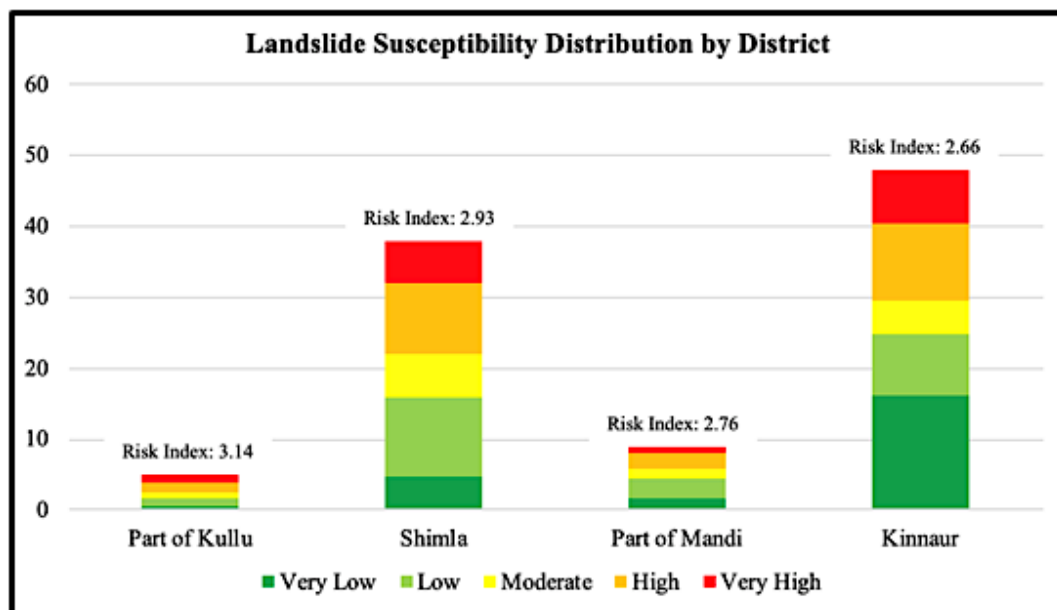


Figure 4.7: Landslide Susceptibility distribution by District along with risk index

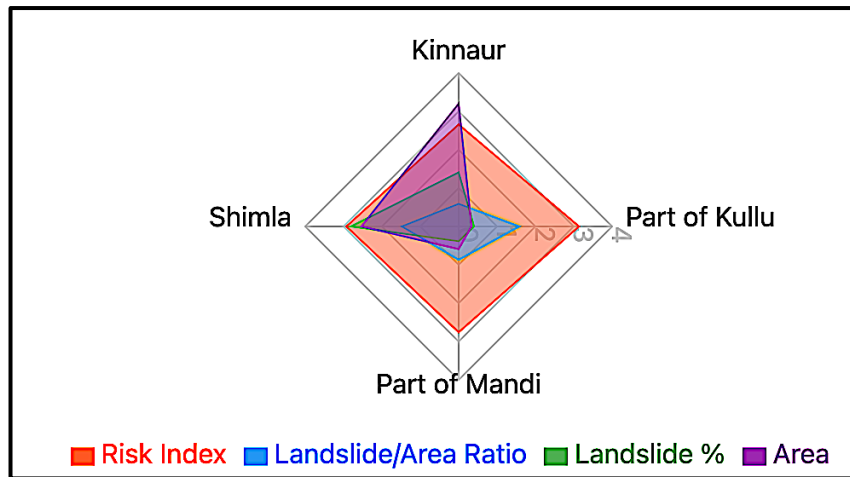


Figure 4.8: Radar Chart showing Risk Index, Landslide/Area Ratio, %Landslide and %Area

Among the districts, Kullu (Part) exhibited the highest risk index (3.15) despite comprising only 5.06% of the study area, highlighting a disproportionately high concentration of landslide events. It also recorded the highest landslide-to-area ratio (1.58) and accounted for 8.01% of total landslides, underscoring the critical need for targeted intervention. Shimla, occupying 38.03% of the total area, ranked second with a risk index of 3.09, hosting the largest number of landslides (56.12%) and the highest absolute extent of High and Very High-risk zones (14.49%). The Mandi (Part) region demonstrated a more balanced profile with a moderate risk index of 2.96 and a proportionate landslide/area ratio of 0.87, aligning well with its moderate share of high-risk zones (2.96%). In contrast, Kinnaur, which spans nearly half of the study area (48.03%), emerged as the least vulnerable district with the lowest risk index (2.86) and a landslide/area ratio of 0.59, suggesting a relatively safer terrain. Despite containing 28.19% of total landslides, Kinnaur's expansive area dilutes its relative hazard concentration, making it more favourable for development activities, especially in zones classified as Very Low to Moderate risk considering other favourable atmospheric conditions as well as elevation limitations.

These findings carry substantial practical implications. From a resource allocation perspective, Kullu and Shimla should be prioritized for landslide mitigation efforts, including intensive monitoring, stabilization measures, and the deployment of early warning systems due to their high-risk profiles. Development planning in Kinnaur can be approached with more flexibility, given its relatively low landslide density, while Kullu requires rigorous geotechnical assessments prior to any infrastructural expansion due to the spatial intensity of its risk. In Shimla, efforts should be geared toward managing the extensive distribution of high-risk zones,

whereas Kullu demands a more focused, localized approach to monitor a smaller but highly susceptible terrain. Overall, the landslide-to-area ratio emerged as a particularly valuable metric, effectively identifying districts like Kullu and Shimla where landslide risks are significantly concentrated relative to land extent, guiding more informed, location-specific risk reduction strategies.

CHAPTER 5

Conclusion

5.1 Objectives Achieved

Table 5.1 : Summary of achieved Objectives

Sr. No.	Objectives	Achievements
1.	Development of Landslide Susceptibility Maps using Machine Learning and GIS	Landslide susceptibility maps were generated utilizing GIS-based methods such as Frequency Ratio and AHP, alongside standalone and ensemble machine learning models, including Artificial Neural Networks (ANN).
2.	Identification and Selection of Landslide Influencing Factors	Both unsupervised and supervised feature selection techniques were applied to evaluate feature importance after extracting relevant attributes from landslide and non-landslide locations.
3.	Validation of Model Performance using Statistical Metrics	The performance of all models was assessed using standard statistical measures, including confusion matrix, accuracy, precision, recall, ROC curve, AUC, and Kappa coefficient. Additionally, k-fold cross-validation was conducted to ensure reliability and prevent overfitting.
4.	Identification of Landslide-Prone Areas for Planning and Mitigation Strategies	Landslide susceptibility maps (LSMs) were developed by probabilistic analysis of each pixel, followed by reclassification into five risk categories—ranging from very low to very high—facilitating future planning and risk mitigation efforts.

As shown in Table 5.1, the research objectives were successfully achieved through the integration of machine learning, GIS, and statistical validation techniques, resulting in reliable landslide susceptibility assessments to support planning and mitigation efforts..

5.2 Concluding Remarks

Machine learning techniques have significantly improved the accuracy and efficiency of landslide susceptibility modelling, outperforming traditional GIS-based methods such as Frequency Ratio (FR) and Analytical Hierarchy Process (AHP) in capturing complex, non-linear relationships. The findings reveal that traditional methods like FR and AHP have inherent limitations due to their reliance on categorical variable conversion and restricted validation approaches.

A rigorous feature selection process, incorporating both unsupervised and supervised techniques, was employed to determine the most statistically significant factors influencing landslides. This ensured that only the most relevant variables were included in the modelling process. The study confirms that slope is the most critical conditioning factor across all feature selection techniques.

Various machine learning algorithms, including Random Forest (RF) and XG-Boost, alongside standalone models like Logistic Regression (LGR), Support Vector Machines (SVM), Kernel SVM (K-SVM), K-Nearest Neighbours (K-NN), Naïve Bayes (NB), and Decision Tree Classification (DT), have been effectively utilized for landslide susceptibility mapping. Among these, ensemble models such as Random Forest demonstrated superior predictive performance, yielding higher accuracy and reliability based on statistical evaluation metrics such as recall, precision, and Cohen's Kappa coefficient. Furthermore, ensemble models exhibited higher Area Under the Receiver Operating Characteristic (ROC) curve values compared to FR, AHP, and standalone ML models, reinforcing their robustness in predictive modelling. The developed ensemble models shows promising potential for application in similar Himalayan terrains, providing a robust framework for landslide risk assessment and disaster management planning.

Overall, tree-based models—whether used individually or in ensemble configurations—have consistently delivered superior accuracy compared to other approaches in the same category. However, certain challenges persist, including the scarcity of high-resolution data and the increasing impact of climate change, both of which present hurdles to the advancement of

landslide susceptibility modelling. Future research in this domain is expected to focus on hybrid and deep learning-based approaches, enhanced data acquisition techniques for improved resolution, and the integration of climate change considerations to refine Landslide Susceptibility Mapping (LSM).

Further, landslide density index and percentage pixel distribution analysis across the different susceptibility classes for different models revealed that Naive Bayes and K-Nearest Neighbours appear more suitable for conservative landslide risk management strategies, where failing to identify high-risk areas could have significant costs. Random Forest and Logistic Regression offer more balanced classification distributions, making them appropriate for general assessments. When prioritizing development opportunities and minimizing unnecessary land use restrictions, Decision Trees may be the preferred approach as they reduce over-prediction of risk areas.

Also, zonal statistics of landslide susceptibility maps of various models suggested to offer priority for landslide mitigation to Kullu and Shimla through enhanced monitoring, stabilization measures, and early warning systems due to their significant risk profiles. Kinnaur can support more flexible development planning given its lower landslide density. Infrastructure expansion in Kullu necessitates stringent geotechnical assessments because of its intense localized risk patterns.

The successful implementation of five-category susceptibility maps (very low to very high) using Jenks Natural Break classification offers practical utility for stakeholders and policymakers in identifying high-risk zones and implementing targeted mitigation strategies.

5.3 Further Research Potential

The continuous advancement of machine learning (ML) techniques and the increasing availability of high-resolution spatial data present several promising directions for future research in landslide susceptibility modelling (LSM). While existing models, particularly tree-based and ensemble approaches, have demonstrated superior accuracy, several aspects require further improvement to enhance their reliability, adaptability, and real-world applicability. Some potential areas that can be further explored are listed below;

1. **Data Availability and Quality:** Reliable, high-quality data is often scarce in remote or underdeveloped areas, which can limit the effectiveness of machine learning models.

2. **Interpretability of ML Models:** Many machine learning models, particularly deep learning techniques, are complex and difficult to interpret. This lack of transparency poses challenges in decision-making, particularly in risk management and disaster response.
3. **Climate and Landcover Change Impacts:** The role of climate change in altering rainfall patterns and triggering landslides is an emerging research area. Future models need to incorporate climate change scenarios to predict how shifting environmental conditions might influence landslide risk.
4. **Transferability of Models:** Models trained in one region may not perform well in another due to the differences in environmental conditions, data quality, and spatial characteristics. This challenge highlights the need for models that are adaptable and can be transferred across different geographic areas.
5. **Zonal statistics at micro-level:** Percentage pixel distribution of different ML models for different sub-divisions of districts as well as villages and towns of sub-division can be carried out to determine risk indexes at micro levels. This can efficiently be used for detailed hazard analysis for the area under consideration.

REFERENCES

1. J. J. Van Dijke and C. J. van Westen, “Rockfall hazard, a geomorphological application of neighbourhood analysis with ILWIS,” *ITC Journal*, vol. 1, pp. 40–44, 1990.
2. L. Highland, *Landslide Types and Processes*, U.S. Geological Survey Fact Sheet FS 2004-3072, 2004.
3. D. J. Varnes, “Slope movement types and processes,” *Transportation Research Board Special Report*, 1978.
4. D. J. Varnes, *Landslide Hazard Zonation: A Review of Principles and Practice*, vol. 3. UNESCO, Paris, 1984.
5. A. K. Turner and R. L. Schuster, *Landslides: Investigation and Mitigation*. Washington, D.C.: Transportation Research Board, 1996.
6. US Geological Survey, “Landslide Types and Processes.” [Online]. Available: <https://pubs.usgs.gov/fs/2004/3072/fs-2004-3072.html>
7. D. P. Kanungo, M. Arora, and S. Sarkar, “Landslide Susceptibility Zonation (LSZ) Mapping—A Review,” *www.em-dat.net*, 2009. [Online]. Available: <https://www.emdat.be/>
8. NASA Earth Data, “A Global View of Landslide Susceptibility.” [Online]. Available: <https://visibleearth.nasa.gov/images/89937/a-global-view-of-landslide-susceptibility>
9. S. Lacasse and F. Nadim, “Landslide risk assessment and mitigation strategy,” in *First World Landslide Forum, Global Landslide Risk Reduction*, International Consortium of Landslides, Tokyo, 2009, pp. 31–61.
10. D. Petley, “Global Patterns of Loss of Life from Landslides,” *Geology*, vol. 40, pp. 927–930, 2012. doi:10.1130/G33217.1
11. EM-DAT, “EM-DAT - the International Disasters Database.” [Online]. Available: <https://www.emdat.be/>, 2024.
12. K. Nagamani, A. K. Mishra, M. S. Meer, *et al.*, “Understanding flash flooding in the Himalayan Region: a case study,” *Scientific Reports*, vol. 14, 7060, 2024. doi:10.1038/s41598-024-53535-w
13. A. Kumar, *et al.*, “Himalayan Communities Are under Siege from Landslides – and Climate Change Is Worsening the Crisis,” *The Conversation*, Oct. 24, 2023. [Online]. Available: <https://theconversation.com/himalayan-communities-are-under-siege-from-landslides-and-climate-change-is-worsening-the-crisis-215202>
14. E. Mitra, *et al.*, “Himalayan Region Landslides Kill 10, Leave 2,400 Tourists Stranded in Northeastern India,” *CNN*, Jun. 14, 2024. [Online]. Available:

- <https://edition.cnn.com/2024/06/14/india/india-sikkim-landslide-deaths-tourists-stranded-intl-hnk/index.html>
15. S. Dubey, *et al.*, “Mass Movement Hazard and Exposure in the Himalaya,” *Earth’s Future*, vol. 11, no. 9, Sep. 2023. doi:10.1029/2022ef003253
 16. M. Ercanoglu, C. Gokceoglu, and T. W. J. Van Asch, “Landslide Susceptibility Zoning North of Yenice (NW Turkey) by Multivariate Statistical Techniques,” *Natural Hazards*, vol. 32, 2004.
 17. D. M. Cruden, “A Simple Definition of Landslide,” 1991. doi:10.1007/BF02590167
 18. G. F. Wieczorek, “Preparing a Detailed Landslide-Inventory Map for Hazard Evaluation and Reduction,” *Bulletin of the Association of Engineering Geologists*, vol. 3, pp. 337–342, 1984.
 19. F. Guzzetti, P. Reichenbach, M. Cardinali, M. Galli, and F. Ardizzone, “Probabilistic landslide hazard assessment at the basin scale,” *Geomorphology*, vol. 72, no. 1–4, pp. 272–299, 2005. doi:10.1016/j.geomorph.2005.06.002
 20. F. Guzzetti, “Landslide Hazard Assessment and Risk Evaluation: Limits and Prospectives,” 2003.
 21. A. Colombo, L. Lanteri, M. Ramasco, and C. Troisi, “Systematic GIS-based landslide inventory as the first step for effective landslide-hazard management,” *Landslides*, vol. 2, no. 4, pp. 291–301, 2005, doi: 10.1007/s10346-005-0025-9.
 22. P. Aleotti and R. Chowdhury, “Landslide hazard assessment: summary review and new perspectives,” *Bull. Eng. Geol. Env.*, vol. 58, 1999.
 23. F. Guzzetti, A. Carrara, M. Cardinali, and P. Reichenbach, “Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy,” *Geomorphology*, vol. 31, 1999.
 24. A. Carrara, M. Sorriso-Valvo, and C. Reali, *Analysis of Landslide Form and Incidence by Statistical Techniques, Southern Italy*, vol. 9, 1982.
 25. J. Chacón, C. Irigaray, T. Fernández, and R. El Hamdouni, “Engineering geology maps: Landslides and geographical information systems,” *Bull. Eng. Geol. Environ.*, vol. 65, no. 4, pp. 341–411, 2006, doi: 10.1007/s10064-006-0064-z.
 26. W. Huabin, L. Gangjun, X. Weiya, and W. Gonghui, “GIS-based landslide hazard assessment: An overview,” *Prog. Phys. Geogr.*, vol. 29, no. 4, pp. 548–567, 2005, doi: 10.1191/0309133305pp462ra.
 27. C. J. Van Westen, E. Castellanos, and S. L. Kuriakose, “Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview,” *Eng. Geol.*, vol. 102, no. 3–4, pp. 112–131, 2008, doi: 10.1016/j.enggeo.2008.03.010.
 28. S. Sarkar, D. P. Kanungo, and G. S. Mehrotra, “Landslide Hazard Zonation: A Case Study in Garhwal Himalaya, India,” *Mountain Res. Dev.*, vol. 15, no. 4, pp. 301–309, 1995.

29. M. Cardinali et al., “A geomorphological approach to the estimation of landslide hazards and risks in Umbria, Central Italy,” *Nat. Hazards Earth Syst. Sci.*, vol. 2, 2002.
30. M. Parise, “Landslide hazard zonation of slopes susceptible to rock falls and topples,” *Nat. Hazards Earth Syst. Sci.*, vol. 2, 2002.
31. T. Preuth, T. Glade, and A. Demoulin, “Stability analysis of a human-influenced landslide in eastern Belgium,” *Geomorphology*, vol. 120, no. 1–2, pp. 38–47, 2010, doi: 10.1016/j.geomorph.2009.09.013.
32. S. Lee, “Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data,” *Int. J. Remote Sens.*, vol. 26, no. 7, pp. 1477–1491, 2005, doi: 10.1080/01431160412331331012.
33. T. Kavzoglu, E. K. Sahin, and I. Colkesen, “Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression,” *Landslides*, vol. 11, no. 3, pp. 425–439, 2014, doi: 10.1007/s10346-013-0391-7.
34. F. Ma, J. Wang, R. Yuan, H. Zhao, and J. Guo, “Application of analytical hierarchy process and least-squares method for landslide susceptibility assessment along the Zhong-Wu natural gas pipeline, China,” *Landslides*, vol. 10, no. 4, pp. 481–492, 2013, doi: 10.1007/s10346-013-0402-8.
35. S. Mondal and R. Maiti, “Landslide Susceptibility Analysis of Shiv-Khola Watershed, Darjeeling: A Remote Sensing & GIS Based Analytical Hierarchy Process (AHP),” *J. Indian Soc. Remote Sens.*, vol. 40, no. 3, pp. 483–496, 2012, doi: 10.1007/s12524-011-0160-9.
36. V. Tofani, S. Segoni, A. Agostini, F. Catani, and N. Casagli, “Technical note: Use of remote sensing for landslide studies in Europe,” *Nat. Hazards Earth Syst. Sci.*, vol. 13, no. 2, pp. 299–309, 2013, doi: 10.5194/nhess-13-299-2013.
37. C. Van Westen, N. Rengers, and R. Soeters, “Use of geomorphological information in indirect landslide susceptibility assessment,” *Nat. Hazards*, vol. 30, pp. 399–419, 2003.
38. C. Zhou, C. Lee, J. Li, and Z. Xu, “On the spatial relationship between landslides and causative factors on Lantau Island, Hong Kong,” *Geomorphology*, vol. 43, pp. 197–207, 2002.
39. S. Lee and K. Min, “Statistical analysis of landslide susceptibility at Yongin, Korea,” *Environ. Geol.*, vol. 40, pp. 1095–1113, 2001.
40. R. Anbalagan, “Landslide hazard evaluation and zonation mapping in mountainous terrain,” *Eng. Geol.*, vol. 32, pp. 269–277, 1992.
41. A. K. Saha, R. P. Gupta, I. Sarkar, M. K. Arora, and E. Csaplovics, “An approach for GIS based statistical landslide susceptibility zonation—with a case study in the Himalayas,” *Landslides*, vol. 2, pp. 61–69, 2005.

42. C. Van Westen, T. Asch, and R. Soeters, "Landslide hazard and risk zonation—why is it still so difficult," *Bull. Eng. Geol. Environ.*, vol. 65, pp. 67–184, 2006.
43. R. C. Sidle and H. Ochiai, *Landslides: Processes, Prediction, and Land Use*, vol. 18. Washington, DC, USA: American Geophysical Union, 2006.
44. T. Kavzoglu, E. K. Sahin, and I. Colkesen, "Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression," *Landslides*, vol. 11, pp. 425–439, 2014.
45. D. T. Bui, B. Pradhan, O. Lofman, and I. Revhaug, "Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes models," *Math. Probl. Eng.*, vol. 2012, 2012.
46. L. Ayalew and H. Yamagishi, "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan," *Geomorphology*, vol. 65, pp. 15–31, 2005.
47. F. Dai and C. Lee, "Landslide characteristics and slope instability modelling using GIS, Lantau Island, Hong Kong," *Geomorphology*, vol. 42, pp. 213–228, 2002.
48. S. Lee, J.-H. Ryu, J.-S. Won, and H.-J. Park, "Determination and application of the weights for landslide susceptibility mapping using an artificial neural network," *Eng. Geol.*, vol. 71, pp. 289–302, 2004.
49. S. Lee and M.-J. Lee, "Detecting landslide location using KOMPSAT 1 and its application to landslide-susceptibility mapping at the Gangneung area, Korea," *Adv. Space Res.*, vol. 38, pp. 2261–2271, 2006.
50. E. Semenza and M. Ghirotti, "History of the 1963 Vaiont slide: The importance of geological factors," *Bull. Eng. Geol. Environ.*, vol. 59, pp. 87–97, 2000.
51. M. Van Den Eeckhaut et al., "Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium)," *Geomorphology*, vol. 76, pp. 392–410, 2006.
52. M. P. Sadr, A. Maghsoudi, and B. S. Saljoughi, "Landslide susceptibility mapping of Komroud Sub-basin using fuzzy logic approach," *Geodyn. Res. Int. Bull.*, vol. 2, 2014.
53. H. Lan, C. Zhou, L. Wang, H. Zhang, and R. Li, "Landslide hazard spatial analysis and prediction using GIS in the Xiaojiang watershed, Yunnan, China," *Eng. Geol.*, vol. 76, pp. 109–128, 2004.
54. K. Terzaghi, R. B. Peck, and G. Mesri, *Soil Mechanics in Engineering Practice*, 3rd ed. New York, NY, USA: John Wiley & Sons, 1996.
55. D. Greenway, "Vegetation and slope stability," in *Slope Stability: Geotechnical Engineering and Geomorphology*, M. G. Anderson and K. S. Richards, Eds. Chichester, UK: Wiley, 1987.

56. F. Catani, N. Casagli, L. Ermini, G. Righini, and G. Menduni, “Landslide hazard and risk mapping at catchment scale in the Arno River basin,” *Landslides*, vol. 2, pp. 329–342, 2005.
57. J. Liu et al., “Landslide hazard assessment in the Three Gorges area of the Yangtze river using ASTER imagery: Zigui–Badong,” *Geomorphology*, vol. 61, pp. 171–187, 2004.
58. J. Corominas and J. Moya, “Reconstructing recent landslide activity in relation to rainfall in the Llobregat River basin, Eastern Pyrenees, Spain,” *Geomorphology*, vol. 30, pp. 79–93, 1999.
59. R. K. Dahal and S. Hasegawa, “Representative rainfall thresholds for landslides in the Nepal Himalaya,” *Geomorphology*, vol. 100, pp. 429–443, 2008.
60. P. Jaiswal and C. J. van Westen, “Estimating temporal probability for landslide initiation along transportation routes based on rainfall thresholds,” *Geomorphology*, vol. 112, pp. 96–105, 2009.
61. A. A. Sarker and A. M. Rashid, “Landslide and flashflood in Bangladesh,” in *Disaster Risk Reduction Approaches in Bangladesh*, Springer, 2013, pp. 165–189.
62. E. Yesilnacar and T. Topal, “Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey),” *Eng. Geol.*, vol. 79, pp. 251–266, 2005.
63. H. R. Pourghasemi et al., “Analysis and evaluation of landslide susceptibility: A review on articles published during 2005–2016 (Periods of 2005–2012 and 2013–2016),” *Arab. J. Geosci.*, vol. 11, no. 9, Apr. 2018. [Online]. Available: <https://doi.org/10.1007/s12517-018-3531-5>
64. J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, pp. 81–106, 1986.
65. J. C. Kim, S. Lee, H. S. Jung, and S. Lee, “Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea,” *Geocarto Int.*, vol. 33, no. 9, pp. 1000–1015, 2018. [Online]. Available: <https://doi.org/10.1080/10106049.2017.1323964>
66. Y. Liu, Y. Wang, and J. Zhang, “LNCS 7473 – New Machine Learning Algorithm: Random Forest,” in *LNCS*, vol. 7473, 2012.
67. L.-L. Liu, C. Yang, and X.-M. Wang, “Landslide susceptibility assessment using feature selection-based machine learning models,” in *GECE*, 2020.
68. A. Naithani, “Macro landslide hazard zonation mapping using univariate statistical analysis in a part of Garhwal Himalaya,” *J. Geol. Soc. India*, vol. 70, pp. 353–368, 2007.
69. H. Akinci and M. Zeybek, “Comparing classical statistic and machine learning models in landslide susceptibility mapping in Ardanuc (Artvin), Turkey,” *Nat. Hazards*, vol. 108, no. 2, pp. 1515–1543, 2021, doi: 10.1007/s11069-021-04743-4.

70. K. Taalab, T. Cheng, and Y. Zhang, “Mapping landslide susceptibility and types using Random Forest,” *Big Earth Data*, vol. 2, no. 2, pp. 159–178, 2018, doi: 10.1080/20964471.2018.1472392.
71. Q. B. Pham *et al.*, “A comparison among fuzzy multi-criteria decision making, bivariate, multivariate and machine learning models in landslide susceptibility mapping,” *Geomatics, Nat. Hazards Risk*, vol. 12, no. 1, pp. 1741–1777, 2021, doi: 10.1080/19475705.2021.1944330.
72. Z. Zhao, Z. Liu, and C. Xu, “Slope Unit-Based Landslide Susceptibility Mapping Using Certainty Factor, Support Vector Machine, Random Forest, CF-SVM and CF-RF Models,” *Front. Earth Sci.*, vol. 9, 2021, doi: 10.3389/feart.2021.589630.
73. N. Micheletti *et al.*, “Machine Learning Feature Selection Methods for Landslide Susceptibility Mapping,” *Math. Geosci.*, vol. 46, no. 1, pp. 33–57, 2014, doi: 10.1007/s11004-013-9511-0.
74. W. S. Noble, “What is a support vector machine?” *Nat. Biotechnol.*, vol. 24, 2006.
75. W. Chen, H. Chai, Z. Zhao, Q. Wang, and H. Hong, “Landslide susceptibility mapping based on GIS and support vector machine models for the Qianyang County, China,” *Environ. Earth Sci.*, vol. 75, no. 6, 2016, doi: 10.1007/s12665-015-5093-0.
76. M. Marjanović, B. Bajat, and M. Kovačević, “Landslide susceptibility assessment with machine learning algorithms,” in *Int. Conf. Intell. Netw. Collab. Syst. (INCoS)*, 2009, pp. 273–278, doi: 10.1109/INCOS.2009.25.
77. H. J. Oh, P. R. Kadavi, C. W. Lee, and S. Lee, “Evaluation of landslide susceptibility mapping by evidential belief function, logistic regression, and support vector machine models,” *Geomatics, Nat. Hazards Risk*, vol. 9, no. 1, pp. 1053–1070, 2018, doi: 10.1080/19475705.2018.1481147.
78. X. Wang, Y. Yu, R. Niu, and Y. Hu, “A combination of geographically weighted regression, particle swarm optimization and support vector machine for landslide susceptibility mapping: A case study at Wanzhou in the Three Gorges area, China,” *Int. J. Environ. Res. Public Health*, vol. 13, no. 5, 2016, doi: 10.3390/ijerph13050487.
79. A. Miner *et al.*, “A comparative study of various data mining techniques as applied to the Modelling of landslide susceptibility on the Bellarine Peninsula,” *Geol. Act* (Victoria), 2010.
80. J. Bröcker and L. A. Smith, “Increasing the reliability of reliability diagrams,” *Weather Forecast.*, vol. 22, pp. 651–661, 2007, doi: 10.1175/WAF993.1.
81. J. S. Chen, H. Y. Huang, and C. Y. Hsu, “A kNN based position prediction method for SNS places,” in *Lect. Notes Comput. Sci.*, 2020.
82. J. Lopez Saez, C. Corona, M. Stoffel, P. Schoeneich, and F. Berger, “Probability maps of landslide reactivation derived from tree-ring records: Pra Bellon landslide,

- southern French Alps,” *Geomorphology*, vol. 138, no. 1, pp. 189–202, 2012, doi: 10.1016/j.geomorph.2011.08.034.
83. S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: A methodology review,” *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002, doi: 10.1016/S1532-0464(03)00034-0.
 84. A. Subasi, “Machine learning techniques,” in *Practical Machine Learning for Data Analysis Using Python*, pp. 91–202, Elsevier, 2020, doi: 10.1016/B978-0-12-821379-7.00003-5.
 85. H. Belyadi and A. Haghighat, “Supervised learning,” in *Machine Learning Guide for Oil and Gas Using Python*, pp. 169–295, Gulf Prof. Publ., 2021, doi: 10.1016/B978-0-12-821929-4.00004-4.
 86. A. Bartosik and H. Whittingham, “Evaluating safety and toxicity,” in *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, S. K. Ashenden, Ed. Academic Press, 2021, pp. 119–137. doi: 10.1016/B978-0-12-820045-2.00008-8.
 87. T. Kavzoglu, I. Colkesen, and E. K. Sahin, “Machine learning techniques in landslide susceptibility mapping: A survey and a case study,” in *Advances in Natural and Technological Hazards Research*, vol. 50, Springer Netherlands, 2019, pp. 283–301. doi: 10.1007/978-3-319-77377-3_13.
 88. Z. Fang, Y. Wang, L. Peng, and H. Hong, “A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping,” *Int. J. Geogr. Inf. Sci.*, vol. 35, no. 2, pp. 321–347, 2021. doi: 10.1080/13658816.2020.1808897.
 89. I. Yilmaz, “Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: Conditional probability, logistic regression, artificial neural networks, and support vector machine,” *Environ. Earth Sci.*, vol. 61, no. 4, pp. 821–836, 2010. doi: 10.1007/s12665-009-0394-9.
 90. Y. Li and W. Chen, “Landslide susceptibility evaluation using hybrid integration of evidential belief function and machine learning techniques,” *Water*, vol. 12, no. 1, 2020. doi: 10.3390/w12010113.
 91. E. Azeraf, E. Monfrini, and W. Pieczynski, “Using the Naive Bayes as a discriminative classifier,” *Unpublished*, 2020.
 92. S. Lee, M. J. Lee, H. S. Jung, and S. Lee, “Landslide susceptibility mapping using Naïve Bayes and Bayesian network models in Umyeonsan, Korea,” *Geocarto Int.*, 2019. doi: 10.1080/10106049.2019.1585482.
 93. I. Imtiaz, M. Umar, M. Latif, R. Ahmed, and M. Azam, “Landslide susceptibility mapping: Improvements in variable weights estimation through machine learning algorithms—a case study of Upper Indus River Basin, Pakistan,” *Environ. Earth Sci.*, vol. 81, no. 4, p. 112, 2022. doi: 10.1007/s12665-022-10233-y.

94. A. M. Youssef and H. R. Pourghasemi, "Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia," *Geosci. Front.*, vol. 12, no. 2, pp. 639–655, 2021. doi: 10.1016/j.gsf.2020.05.010.
95. S. Wang, "Artificial Neural Network," in *Interdisciplinary Computing in Java Programming*, vol. 743, *The Springer International Series in Engineering and Computer Science*, Springer, Boston, MA, 2003. doi: 10.1007/978-1-4615-0377-4_5.
96. J. Choi, H. J. Oh, J. S. Won, and S. Lee, "Validation of an artificial neural network model for landslide susceptibility mapping," *Environ. Earth Sci.*, vol. 60, no. 3, pp. 473–483, 2010. doi: 10.1007/s12665-009-0188-0.
97. L. V. Lucchese, G. G. de Oliveira, and O. C. Pedrollo, "Mamdani fuzzy inference systems and artificial neural networks for landslide susceptibility mapping," *Nat. Hazards*, vol. 106, no. 3, pp. 2381–2405, 2021. doi: 10.1007/s11069-021-04547-6.
98. I. Yilmaz, "A case study from Koyulhisar (Sivas-Turkey) for landslide susceptibility mapping by artificial neural networks," *Bull. Eng. Geol. Environ.*, vol. 68, no. 3, pp. 297–306, 2009. doi: 10.1007/s10064-009-0185-2.
99. H. C. Quan and B. G. Lee, "GIS-based landslide susceptibility mapping using analytic hierarchy process and artificial neural network in Jeju (Korea)," *KSCE J. Civ. Eng.*, vol. 16, no. 7, pp. 1258–1266, 2012. doi: 10.1007/s12205-012-1242-0.
100. Q. Wang et al., "Landslide susceptibility mapping at Gongliu county, China using artificial neural network and weight of evidence models," *Geosci. J.*, vol. 20, no. 5, pp. 705–718, 2016. doi: 10.1007/s12303-016-0003-3.
101. M. Zare, H. R. Pourghasemi, M. Vafakhah, and B. Pradhan, "Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: A comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms," *Arab. J. Geosci.*, vol. 6, no. 8, pp. 2873–2888, 2013. doi: 10.1007/s12517-012-0610-x.
102. A. Can, G. Dagdelenler, M. Ercanoglu, and H. Sonmez, "Landslide susceptibility mapping at Ovacık-Karabük (Turkey) using different artificial neural network models: Comparison of training algorithms," *Bull. Eng. Geol. Environ.*, vol. 78, no. 1, pp. 89–102, 2019. doi: 10.1007/s10064-017-1034-3.
103. H. Harmouzi et al., "Landslide susceptibility mapping of the Mediterranean coastal zone of Morocco between Oued Laou and El Jebha using artificial neural networks (ANN)," *Arab. J. Geosci.*, vol. 12, no. 22, 2019. doi: 10.1007/s12517-019-4892-0.
104. E. K. Kutlug Sahin and I. Colkesen, "Performance analysis of advanced decision tree-based ensemble learning algorithms for landslide susceptibility mapping," *Geocarto Int.*, vol. 36, no. 11, pp. 1253–1275, 2021. doi: 10.1080/10106049.2019.1641560.

105. R. Li and N. Wang, “Landslide susceptibility mapping for the Muchuan County (China): A comparison between bivariate statistical models (WOE, EBF, and IOE) and their ensembles with logistic regression,” *Symmetry*, vol. 11, no. 6, 2019. doi: 10.3390/sym11060762.
106. T. Y. Zhang et al., “GIS-based landslide susceptibility mapping using hybrid integration approaches of fractal dimension with index of entropy and support vector machine,” *J. Mt. Sci.*, vol. 16, no. 6, pp. 1275–1288, 2019. doi: 10.1007/s11629-018-5337-z.
107. X. Hu et al., “Landslide susceptibility mapping using the stacking ensemble machine learning method in Lushui, Southwest China,” *Appl. Sci.*, vol. 10, no. 11, 2020. doi: 10.3390/app10114016.
108. B. T. Pham et al., “Landslide susceptibility mapping using state-of-the-art machine learning ensembles,” *Geocarto Int.*, vol. 37, no. 18, pp. 5175–5200, 2022. doi: 10.1080/10106049.2021.1914746.
109. G. Wang, X. Lei, W. Chen, H. Shahabi, and A. Shirzadi, “Hybrid computational intelligence methods for landslide susceptibility mapping,” *Symmetry*, vol. 12, no. 3, 2020. doi: 10.3390/sym12030325.
110. J. Song, Y. Wang, Z. Fang, L. Peng, and H. Hong, “Potential of ensemble learning to improve tree-based classifiers for landslide susceptibility mapping,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 4642–4662, 2020. doi: 10.1109/JSTARS.2020.3014143.
111. L. Peng, R. Niu, B. Huang, X. Wu, Y. Zhao, and R. Ye, “Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China,” *Geomorphology*, vol. 204, pp. 287–301, 2014, doi: 10.1016/j.geomorph.2013.08.013.
112. B. T. Pham et al., “A novel intelligence approach of a sequential minimal optimization-based support vector machine for landslide susceptibility mapping,” *Sustainability*, vol. 11, no. 22, 2019, doi: 10.3390/su11226323.
113. M. S. G. Adnan et al., “Improving spatial agreement in machine learning-based landslide susceptibility mapping,” *Remote Sensing*, vol. 12, no. 20, pp. 1–23, 2020, doi: 10.3390/rs12203347.
114. Y. Wang, L. Feng, S. Li, F. Ren, and Q. Du, “A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China,” *Catena*, vol. 188, 2020, doi: 10.1016/j.catena.2019.104425.
115. Q. Hu, Y. Zhou, S. Wang, and F. Wang, “Machine learning and fractal theory models for landslide susceptibility mapping: Case study from the Jinsha River Basin,” *Geomorphology*, vol. 351, 2020, doi: 10.1016/j.geomorph.2019.106975.
116. G. Rong et al., “Rainfall induced landslide susceptibility mapping based on bayesian optimized random forest and gradient boosting decision tree models—a case study

- of Shuicheng County, China,” *Water*, vol. 12, no. 11, pp. 1–22, 2020, doi: 10.3390/w12113066.
117. T. Kavzoglu and A. Teke, “Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (xgboost) and natural gradient boosting (ngboost),” *Arab. J. Sci. Eng.*, vol. 47, no. 6, pp. 7367–7385, 2022, doi: 10.1007/s13369-022-06560-8.
 118. H. Zhang *et al.*, “Combining a class-weighted algorithm and machine learning models in landslide susceptibility mapping: A case study of Wanzhou section of the Three Gorges Reservoir, China,” *Computers & Geosciences*, vol. 158, 2022, doi: 10.1016/j.cageo.2021.104966.
 119. B. Kalantar *et al.*, “Landslide susceptibility mapping: Machine and ensemble learning based on remote sensing big data,” *Remote Sensing*, vol. 12, no. 11, 2020, doi: 10.3390/rs12111737.
 120. Z. Liang *et al.*, “A hybrid model consisting of supervised and unsupervised learning for landslide susceptibility mapping,” *Remote Sensing*, vol. 13, no. 8, 2021, doi: 10.3390/rs13081464.
 121. V. H. Nhu *et al.*, “Shallow landslide susceptibility mapping by Random Forest base classifier and its ensembles in a semi-arid region of Iran,” *Forests*, vol. 11, no. 4, 2020, doi: 10.3390/f11040421.
 122. S. Saha, J. Roy, B. Pradhan, and T. K. Hembram, “Hybrid ensemble machine learning approaches for landslide susceptibility mapping using different sampling ratios at East Sikkim Himalayan, India,” *Adv. Space Res.*, vol. 68, no. 7, pp. 2819–2840, 2021, doi: 10.1016/j.asr.2021.05.018.
 123. E. K. Sahin, I. Colkesen, and T. Kavzoglu, “A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping,” *Geocarto Int.*, vol. 35, no. 4, pp. 341–363, 2020, doi: 10.1080/10106049.2018.1516248.
 124. O. F. Althuwaynee, B. Pradhan, H. J. Park, and J. H. Lee, “A novel ensemble decision tree-based chi-squared automatic interaction detection (CHAID) and multivariate logistic regression models in landslide susceptibility mapping,” *Landslides*, vol. 11, no. 6, pp. 1063–1078, 2014, doi: 10.1007/s10346-014-0466-0.
 125. T. Duan *et al.*, “NG Boost: Natural Gradient Boosting for probabilistic prediction,” *arXiv preprint arXiv:1910.03225*, 2019.
 126. W. Zhang, H. Li, L. Han, L. Chen, and L. Wang, “Slope stability prediction using ensemble learning techniques: A case study in Yunyang County, Chongqing, China,” *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1089–1099, 2022, doi: 10.1016/j.jrmge.2021.12.011.
 127. X. Zhou, H. Wen, Z. Li, H. Zhang, and W. Zhang, “An interpretable model for the susceptibility of rainfall-induced shallow landslides based on SHAP and xgboost,”

- Geocarto Int.*, vol. 37, no. 26, pp. 13419–13450, 2022, doi: 10.1080/10106049.2022.2076928.
128. G. I. Webb, “Multiboosting: A technique for combining boosting and wagging,” *Mach. Learn.*, vol. 40, 2000.
 129. L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, 1996.
 130. D. T. Bui *et al.*, “GIS-based Modelling of rainfall-induced landslides using data mining-based functional trees classifier with adaboost, bagging, and multiboost ensemble frameworks,” *Environ. Earth Sci.*, vol. 75, no. 14, 2016, doi: 10.1007/s12665-016-5919-4.
 131. K. M. Ting and I. H. Witten, “Stacking bagged and dagged models,” *Computer Science Working Papers*, 1997. [Online]. Available: <https://hdl.handle.net/10289/1072>.
 132. X. Zhou, H. Wen, Y. Zhang, J. Xu, and W. Zhang, “Landslide susceptibility mapping using hybrid random forest with geodetector and RFE for factor optimization,” *Geosci. Front.*, vol. 12, no. 5, 2021, doi: 10.1016/j.gsf.2021.101211.
 133. P. Melville and R. J. Mooney, “Creating diversity in ensembles using artificial data,” *Inf. Fusion*, vol. 6, no. 1, pp. 99–111, 2005, doi: 10.1016/j.inffus.2004.04.001.
 134. Z. Fang, Y. Wang, G. Duan, and L. Peng, “Landslide susceptibility mapping using rotation forest ensemble technique with different decision trees in the Three Gorges Reservoir Area, China,” *Remote Sensing*, vol. 13, no. 2, 2021, doi: 10.3390/rs13020238.
 135. X. Hu, C. Huang, H. Mei, and H. Zhang, “Landslide susceptibility mapping using an ensemble model of Bagging scheme and random subspace-based naïve Bayes tree in Zigui County of the Three Gorges Reservoir Area, China,” *Bull. Eng. Geol. Environ.*, vol. 80, no. 7, pp. 5315–5329, 2021, doi: 10.1007/s10064-021-02275-6.
 136. W. Li, Z. Fang, and Y. Wang, “Stacking ensemble of deep learning methods for landslide susceptibility mapping in the Three Gorges Reservoir area, China,” *Stoch. Environ. Res. Risk Assess.*, vol. 36, no. 8, pp. 2207–2228, 2022, doi: 10.1007/s00477-021-02032-x.
 137. A. Töschner, M. Jahrer, and R. M. Bell, “The bigchaos solution to the Netflix Grand Prize,” 2009.
 138. W. Gong *et al.*, “GIS-based landslide susceptibility mapping using ensemble methods for Fengjie County in the Three Gorges Reservoir Region, China,” *Int. J. Environ. Sci. Technol.*, vol. 19, no. 8, pp. 7803–7820, 2022, doi: 10.1007/s13762-021-03572-z.
 139. M. Di Napoli *et al.*, “Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability,” *Landslides*, vol. 17, no. 8, pp. 1897–1914, 2020, doi: 10.1007/s10346-020-01392-9.

140. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
141. T. Kavzoglu, A. Teke, and E. O. Yilmaz, "Shared blocks-based ensemble deep learning for shallow landslide susceptibility mapping," *Remote Sensing*, vol. 13, no. 23, 2021, doi: 10.3390/rs13234776.
142. R. Liu, X. Yang, C. Xu, L. Wei, and X. Zeng, "Comparative study of convolutional neural network and conventional machine learning methods for landslide susceptibility mapping," *Remote Sensing*, vol. 14, no. 2, 2022, doi: 10.3390/rs14020321.
143. J. M. Habumugisha *et al.*, "Landslide susceptibility mapping with deep learning algorithms," *Sustainability*, vol. 14, no. 3, 2022, doi: 10.3390/su14031734.
144. A. Dikshit, R. Sarkar, B. Pradhan, S. Segoni, and A. M. Alamri, "Rainfall induced landslide studies in Indian Himalayan region: A critical review," *Appl. Sci.*, vol. 10, no. 7, p. 2466, 2020, doi: 10.3390/app10072466.
145. B. T. Pham *et al.*, "Landslide susceptibility assessment by novel hybrid machine learning algorithms," *Sustainability*, vol. 11, no. 15, p. 4386, 2019.
146. D. P. Kanungo, M. K. Arora, S. Sarkar, and R. P. Gupta, "A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas," *Engineering Geology*, vol. 85, no. 3–4, pp. 347–366, 2006, doi: 10.1016/j.enggeo.2006.03.004.
147. S. R. Meena *et al.*, "Landslide detection in the Himalayas using machine learning algorithms and U-Net," *Landslides*, vol. 19, no. 5, pp. 1209–1229, 2022, doi: 10.1007/s10346-022-01861-3.
148. D. P. Kanungo and S. Sarkar, "Terrain analysis and spatial assessment of landslide hazard in parts of Sikkim Himalaya," *Journal of the Geological Society of India*, 1996. [Online]. Available: <https://www.researchgate.net/publication/257679254>
149. D. Pathak, "Knowledge based landslide susceptibility mapping in the Himalayas," *Geoenvironmental Disasters*, vol. 3, no. 1, 2016, doi: 10.1186/s40677-016-0042-0.
150. P. K. C. Ray, S. Dimri, R. C. Lakhera, and S. Sati, "Fuzzy-based method for landslide hazard assessment in active seismic zone of Himalaya," *Landslides*, vol. 4, no. 2, pp. 101–111, 2007, doi: 10.1007/s10346-006-0068-6.
151. S. V. Panikkar and V. Subramanyan, "Landslide hazard analysis of the area around Dehra Dun and Mussoorie, Uttar Pradesh," *Current Science*, vol. 73, no. 12, pp. 1117–1123, 1997. [Online]. Available: <http://www.jstor.org/stable/24101273>
152. T. R. Martha, C. J. van Westen, N. Kerle, V. Jetten, and K. Vinod Kumar, "Landslide hazard and risk assessment using semi-automatically created landslide inventories," *Geomorphology*, vol. 184, pp. 139–150, 2013, doi: 10.1016/j.geomorph.2012.12.001.

153. I. Das, A. Stein, N. Kerle, and V. K. Dadhwal, "Landslide susceptibility mapping along road corridors in the Indian Himalayas using Bayesian logistic regression models," *Geomorphology*, vol. 179, pp. 116–125, 2012, doi: 10.1016/j.geomorph.2012.08.004.
154. R. P. Gupta and B. C. Joshi, "Landslide Hazard Zoning Using the GIS Approach: A Case Study from the Ramganga Catchment, Himalayas," *Engineering Geology*, vol. 28, 1990.
155. S. Sarkar, D. P. Kanungo, A. K. Patra, and P. Kumar, "GIS Based Landslide Susceptibility Mapping – A Case Study in Indian Himalaya," *CSIR-12th Five Year Plan Project Report*. [Online]. Available: <https://www.researchgate.net/publication/242073557>
156. R. K. Dahal *et al.*, "GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping," *Environmental Geology*, vol. 54, no. 2, pp. 311–324, 2008, doi: 10.1007/s00254-007-0818-3.
157. I. Das, A. Stein, N. Kerle, and V. K. Dadhwal, "Probabilistic landslide hazard assessment using homogeneous susceptible units (HSU) along a national highway corridor in the northern Himalayas, India," *Landslides*, vol. 8, no. 3, pp. 293–308, 2011, doi: 10.1007/s10346-011-0257-9.
158. K. Balasubramani and K. Kumaraswamy, "Application of Geospatial Technology and Information Value Technique in Landslide Hazard Zonation Mapping: A Case Study of Giri Valley, Himachal Pradesh," *Disaster Advances*, vol. 6, no. 1, 2013. [Online]. Available: <https://www.researchgate.net/publication/234062937>
159. D. Chakraborty, R. Anbalagan, D. Chakraborty, and A. Kohli, "Landslide hazard zonation (LHZ) mapping on meso-scale for systematic town planning in mountainous terrain," *Journal of Scientific & Industrial Research*, vol. 67, 2008. [Online]. Available: <https://www.researchgate.net/publication/292277786>
160. S. Ghosh *et al.*, "A quantitative approach for improving the BIS (Indian) method of medium-scale landslide susceptibility," *Journal of the Geological Society of India*, vol. 74, no. 5, pp. 625–638, 2009, doi: 10.1007/s12594-009-0167-9.
161. L. P. Sharma, N. Patel, and P. K. Debnath, "Geographical Information System Based Landslide Probabilistic Model with tri-variate approach—a case study in Sikkim Himalayas," *IEEE International Conference on Advanced Computational and Communication Paradigms (ICACCP-2017)*. [Online]. Available: <https://www.researchgate.net/publication/261925928>
162. P. Kainthura and N. Sharma, "Machine learning driven landslide susceptibility prediction for the Uttarkashi region of Uttarakhand in India," *Georisk*, vol. 16, no. 3, pp. 570–583, 2021, doi: 10.1080/17499518.2021.1957484.
163. V. Chauhan, L. Gupta, and J. Dixit, "Landslide susceptibility assessment for Uttarakhand, a Himalayan state of India, using multi-criteria decision making,

- bivariate, and machine learning models," *Geoenvironmental Disasters*, vol. 12, no. 2, 2025, doi: 10.1186/s40677-024-00307-3.
164. P. Ram *et al.*, "Landslide susceptibility mapping using bivariate statistical method for the hilly township of Mussoorie and its surrounding areas, Uttarakhand Himalaya," *Journal of Earth System Science*, vol. 129, p. 167, 2020, doi: 10.1007/s12040-020-01428-7.
165. Sangeeta and B. K. Maheshwari, "Spatial predictive modelling of rainfall- and earthquake-induced landslide susceptibility in the Himalaya region of Uttarakhand, India," *Environmental Earth Sciences*, vol. 81, p. 237, 2022, doi: 10.1007/s12665-022-10352-6.

APPENDIX A : Results of GIS based Landslide susceptibility models (FR and AHP)

Determination of Prediction Rate for different features for Frequency Ratio Modelling

Sr. No	Layer	Parameter	Nos. Of LSD (NLi)	LSD %	Area (No. Of Pixel) NCi	Area %	FR (FRi)	RF	RF%	RF(IN T)	Min (RF)	Max(RF)	MaxRF - Min RF	Min (MaxRF - MinRF)	PR
1	Slope	0 - 15	41	3.75	1837691	12.31	0.30	0.05	5.10	5.00					
		15-30	378	34.58	6584863	44.12	0.78	0.13	13.13	13.00					
		30-45	582	53.25	5733217	38.41	1.39	0.23	23.21	23.00					
		45-60	90	8.23	755076	5.06	1.63	0.27	27.25	27.00					
		60 or greater	2	0.18	14606	0.10	1.87	0.31	31.31	32.00					
		NLt=	1093		14925453		5.97	1.00			0.05	0.31	0.26	0.02	10.50
2	NDVI	Dead Plant or Object (-1 to 0.1)	307	28.09	6271189	42.02	0.67	0.21	20.68	21.00					
		Unhealthy Plant (0.1 to 0.25)	347	31.75	3262865	21.86	1.45	0.45	44.92	45.00					
		Healthy plant (0.25 to 1)	439	40.16	5391399	36.12	1.11	0.34	34.40	34.00					
		NLt=	1093		14925453		3.23	1.00			0.21	0.45	0.24	0.02	9.72
3	Elevation	Less than 2000 m	431	39.43	3685906	24.70	1.60	0.35	35.39	35.00					

Sr. No	Layer	Parameter	Nos. Of LSD (NLi)	LSD %	Area (No. Of Pixel) NCi	Area %	FR (FRi)	RF	RF%	RF(IN T)	Min (RF)	Max(RF)	MaxRF - Min RF	Min (MaxRF - MinRF)	PR
		2000 - 3000 m	299	27.36	3710140	24.86	1.10	0.24	24.39	24.00					
		3000 - 4000 m	215	19.67	2451031	16.42	1.20	0.27	26.55	27.00					
		4000 - 5000 m	147	13.45	3298312	22.10	0.61	0.13	13.49	14.00					
		More than 5000 m	1	0.09	1780064	11.93	0.01	0.00	0.17	0.00					
		NLt=	1093		14925453		4.51	1.00			0.00	0.35	0.35	0.02	14.12
4	Curvature	Upwardly Convex (-)	542	49.59	7203367	48.26	1.03	0.35	34.75	35.00					
		Linear (0)	28	2.56	400920	2.69	0.95	0.32	32.26	32.00					
		Upwardly Concave (+)	523	47.85	7321166	49.05	0.98	0.33	32.99	33.00					
		NLt=	1093		14925453		2.96	1.00			0.32	0.35	0.02	0.02	1.00
5	Distance from River	Less than 100 m	85	7.78	737700	4.94	1.57	0.18	18.02	18.00					
		101-200 m	120	10.98	661060	4.43	2.48	0.28	28.39	28.00					
		201-300 m	76	6.95	711094	4.76	1.46	0.17	16.72	17.00					
		301-400 m	69	6.31	602153	4.03	1.56	0.18	17.92	18.00					
		401-500 m	39	3.57	646716	4.33	0.82	0.09	9.43	9.00					
		More than 500 m	704	64.41	11566730	77.50	0.83	0.10	9.52	10.00					
		NLt=	1093		14925453		8.73	1.00			0.09	0.28	0.19	0.02	7.60

Sr. No	Layer	Parameter	Nos. Of LSD (NLi)	LSD %	Area (No. Of Pixel) NCi	Area %	FR (FRi)	RF	RF%	RF(IN T)	Min (RF)	Max(RF)	MaxRF - Min RF	Min (MaxRF - MinRF)	PR
6	Distance from Road	Less than 100 m	248	22.69	454734	3.05	7.45	0.56	56.11	56.00					
		101-200 m	26	2.38	319134	2.14	1.11	0.08	8.38	8.00					
		201-300 m	24	2.20	311578	2.09	1.05	0.08	7.93	8.00					
		301-400 m	26	2.38	251378	1.68	1.41	0.11	10.64	11.00					
		401-500 m	28	2.56	256873	1.72	1.49	0.11	11.22	11.00					
		More than 500 m	741	67.80	13331756	89.32	0.76	0.06	5.72	6.00					
		NLt=	1093		14925453		13.27	1.00			0.06	0.56	0.50	0.02	20.19
7	Geology	Archaean (?) - Proterozoic (undifferentiated)	296	27.08	4263252	28.56	0.95	0.05	5.26	5.00					
		Mesoproterozoic	139	12.72	1602593	10.74	1.18	0.07	6.57	7.00					
		Carboniferous - Permian	18	1.65	276570	1.85	0.89	0.05	4.93	5.00					
		Neoproterozoic (undifferentiated)	105	9.61	1573188	10.54	0.91	0.05	5.06	5.00					
		Neoproterozoic (Upper)	1	0.09	214116	1.43	0.06	0.00	0.35	0.00					
		Permian	30	2.74	221596	1.48	1.85	0.10	10.26	10.00					
		Palaeozoic	34	3.11	1475214	9.88	0.31	0.02	1.75	2.00					
		Proterozoic (undifferentiated)	120	10.98	1031766	6.91	1.59	0.09	8.81	9.00					
		Ordovician - Carboniferous	22	2.01	782434	5.24	0.38	0.02	2.13	2.00					

Sr. No	Layer	Parameter	Nos. Of LSD (NLi)	LSD %	Area (No. Of Pixel) NCi	Area %	FR (FRi)	RF	RF%	RF(IN T)	Min (RF)	Max(RF)	MaxRF - Min RF	Min (MaxRF - MinRF)	PR
		Cambrian-Ordovician	30	2.74	586072	3.93	0.70	0.04	3.88	4.00					
		Neoproterozoic (Lower)	196	17.93	1574755	10.55	1.70	0.09	9.43	9.00					
		Eocene - Miocene	7	0.64	156864	1.05	0.61	0.03	3.38	3.00					
		Creataceous	0	0.00	3988	0.03	0.00	0.00	0.00	0.00					
		Jurassic - Creataceous	0	0.00	47278	0.32	0.00	0.00	0.00	0.00					
		Pliocene - Pleiostocene	3	0.27	89704	0.60	0.46	0.03	2.53	3.00					
		Palaeoproterozoic	18	1.65	45543	0.31	5.40	0.30	29.94	30.00					
		Mesoproterozoic - Neoproterozoic (undifferentiated)	74	6.77	980520	6.57	1.03	0.06	5.72	6.00					
		NLt=	1093		14925453		18.02	1.00			0.00	0.30	0.30	0.02	12.00
8	LULC	Barren Unculturable	132	12.08	1423862	9.54	1.27	0.14	13.60	14.00					
		Waterbody River/Reservoir	14	1.28	171408	1.15	1.12	0.12	11.98	12.00					
		Forest	288	26.35	4166409	27.91	0.94	0.10	10.14	10.00					
		Grass/Grazing	209	19.12	1890645	12.67	1.51	0.16	16.21	16.00					
		Builtup	45	4.12	264126	1.77	2.33	0.25	24.99	25.00					
		Agriculture	186	17.02	2113802	14.16	1.20	0.13	12.91	13.00					

Sr. No	Layer	Parameter	Nos. Of LSD (NLi)	LSD %	Area (No. Of Pixel) NCi	Area %	FR (FRi)	RF	RF%	RF(IN T)	Min (RF)	Max(RF)	MaxRF - Min RF	Min (MaxRF - MinRF)	PR
		Barren Rocky	215	19.67	3209298	21.50	0.91	0.10	9.83	10.00					
		Snow and Glacier	4	0.37	1685903	11.30	0.03	0.00	0.35	0.00					
		NLt=	1093		14925453		9.31	1.00			0.00	0.25	0.25	0.02	9.87
9	Geomorphology	Moderately Dissected Hills and Valleys	255	23.33	2017223	13.52	1.73	0.02	2.13	2.00					
		Low Dissected Hills and Valleys	0	0.00	1154	0.01	0.00	0.00	0.00	0.00					
		Piedmont Slope	15	1.37	352354	2.36	0.58	0.01	0.72	1.00					
		Mass Wasting Products	18	1.65	4100	0.03	59.95	0.74	73.99	74.00					
		Pediment Pediplain Complex	0	0.00	337	0.00	0.00	0.00	0.00	0.00					
		Alluvial Plain	10	0.91	184132	1.23	0.74	0.01	0.92	1.00					
		Flood Plain	0	0.00	9952	0.07	0.00	0.00	0.00	0.00					
		Waterbody - River	5	0.46	152560	1.02	0.45	0.01	0.55	1.00					
		Highly Dissected Hills and Valleys	776	71.00	9577979	64.17	1.11	0.01	1.37	1.00					
		Dam and Reservoir	0	0.00	549	0.00	0.00	0.00	0.00	0.00					
		Snow Cover	8	0.73	2198909	14.73	0.05	0.00	0.06	0.00					
		Glacial Valley	0	0.00	63322	0.42	0.00	0.00	0.00	0.00					
		Valley Glacier	0	0.00	355893	2.38	0.00	0.00	0.00	0.00					

Sr. No	Layer	Parameter	Nos. Of LSD (NLi)	LSD %	Area (No. Of Pixel) NCi	Area %	FR (FRi)	RF	RF%	RF(IN T)	Min (RF)	Max(RF)	MaxRF - Min RF	Min (MaxRF - MinRF)	PR
		Glacial Moraine	0	0.00	1511	0.01	0.00	0.00	0.00	0.00					
		Glacial Lake and Nunatak	0	0.00	449	0.00	0.00	0.00	0.00	0.00					
		Anthropogenic terrain	6	0.55	4989	0.03	16.42	0.20	20.27	20.00					
		Waterbodies-Other	0	0.00	40	0.00	0.00	0.00	0.00	0.00					
		NLt=	1093		14925453		81.03	1.00			0.00	0.74	0.74	0.02	29.65
10	Soil Type	3661	396	36.23	4994030	33.46	1.08	0.14	14.38	14.00					
		3670	69	6.31	726873	4.87	1.30	0.17	17.21	17.00					
		3671	103	9.42	608755	4.08	2.31	0.31	30.68	31.00					
		3679	137	12.53	1041810	6.98	1.80	0.24	23.85	24.00					
		3717	368	33.67	5742914	38.48	0.88	0.12	11.62	12.00					
		6998	20	1.83	1603168	10.74	0.17	0.02	2.26	2.00					
		NLt=	1093		14925453		7.53	1.00			0.00	0.31	0.31	0.02	12.29

AHP Analysis

STEP-I : Pairwise Comparison Matrix

Factors	Geomorphology	Distance from Road	Elevation	Soil Type	Geology	Slope	LULC	NDVI	Distance from River	Curvature
Geomorphology	1.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00	4.00	5.00
Distance from Road	0.50	1.00	2.00	2.00	2.00	2.00	2.00	2.00	3.00	5.00
Elevation	0.50	0.50	1.00	2.00	2.00	2.00	2.00	2.00	3.00	5.00
Soil Type	0.33	0.50	0.50	1.00	2.00	2.00	2.00	2.00	2.00	4.00
Geology	0.33	0.50	0.50	0.50	1.00	2.00	2.00	2.00	2.00	4.00
Slope	0.33	0.50	0.50	0.50	0.50	1.00	2.00	2.00	2.00	4.00
LULC	0.33	0.50	0.50	0.50	0.50	0.50	1.00	2.00	2.00	4.00
NDVI	0.33	0.50	0.50	0.50	0.50	0.50	0.50	1.00	2.00	4.00
Distance from River	0.25	0.33	0.33	0.50	0.50	0.50	0.50	0.50	1.00	3.00
Curvature	0.20	0.20	0.20	0.25	0.25	0.25	0.25	0.25	0.33	1.00
Sum=	4.12	6.53	8.03	10.75	12.25	13.75	15.25	16.75	21.33	39.00

STEP-II : Normalized Pairwise Comparison Matrix

Factors	Geomorphology	Distance from Road	Elevation	Soil Type	Geology	Slope	LULC	NDVI	Distance from River	Curvature
Geomorphology	0.24	0.31	0.25	0.28	0.24	0.22	0.20	0.18	0.19	0.13
Distance from Road	0.12	0.15	0.25	0.19	0.16	0.15	0.13	0.12	0.14	0.13
Elevation	0.12	0.08	0.12	0.19	0.16	0.15	0.13	0.12	0.14	0.13
Soil Type	0.08	0.08	0.06	0.09	0.16	0.15	0.13	0.12	0.09	0.10
Geology	0.08	0.08	0.06	0.05	0.08	0.15	0.13	0.12	0.09	0.10
Slope	0.08	0.08	0.06	0.05	0.04	0.07	0.13	0.12	0.09	0.10
LULC	0.08	0.08	0.06	0.05	0.04	0.04	0.07	0.12	0.09	0.10
NDVI	0.08	0.08	0.06	0.05	0.04	0.04	0.03	0.06	0.09	0.10
Distance from River	0.06	0.05	0.04	0.05	0.04	0.04	0.03	0.03	0.05	0.08
Curvature	0.05	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.03
Sum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

STEP-III : Determination of Critical Weight

Factors	Geomorphology	Distance from Road	Elevation	Soil Type	Geology	Slope	LULC	NDVI	Distance from River	Curvature	Criteria Weight
Geomorphology	0.24	0.31	0.25	0.28	0.24	0.22	0.20	0.18	0.19	0.13	0.22
Distance from Road	0.12	0.15	0.25	0.19	0.16	0.15	0.13	0.12	0.14	0.13	0.15
Elevation	0.12	0.08	0.12	0.19	0.16	0.15	0.13	0.12	0.14	0.13	0.13
Soil Type	0.08	0.08	0.06	0.09	0.16	0.15	0.13	0.12	0.09	0.10	0.11
Geology	0.08	0.08	0.06	0.05	0.08	0.15	0.13	0.12	0.09	0.10	0.09
Slope	0.08	0.08	0.06	0.05	0.04	0.07	0.13	0.12	0.09	0.10	0.08
LULC	0.08	0.08	0.06	0.05	0.04	0.04	0.07	0.12	0.09	0.10	0.07
NDVI	0.08	0.08	0.06	0.05	0.04	0.04	0.03	0.06	0.09	0.10	0.06
Distance from River	0.06	0.05	0.04	0.05	0.04	0.04	0.03	0.03	0.05	0.08	0.05
Curvature	0.05	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.03	0.02

STEP-IV : Determination of Weighted Sum

Criteria Weight	0.22	0.15	0.13	0.11	0.09	0.08	0.07	0.06	0.05	0.02	
	Geomorphology	Distance from Road	Elevation	Soil Type	Geology	Slope	LULC	NDVI	Distance from River	Curvature	Weighted Sum Value
Geomorphology	0.223	0.308	0.267	0.321	0.282	0.248	0.217	0.190	0.185	0.119	2.360
Distance from Road	0.112	0.154	0.267	0.214	0.188	0.165	0.145	0.126	0.139	0.119	1.629
Elevation	0.112	0.077	0.134	0.214	0.188	0.165	0.145	0.126	0.139	0.119	1.419
Soil Type	0.074	0.077	0.067	0.107	0.188	0.165	0.145	0.126	0.093	0.095	1.138
Geology	0.074	0.077	0.067	0.053	0.094	0.165	0.145	0.126	0.093	0.095	0.990
Slope	0.074	0.077	0.067	0.053	0.047	0.083	0.145	0.126	0.093	0.095	0.861
LULC	0.074	0.077	0.067	0.053	0.047	0.041	0.072	0.126	0.093	0.095	0.747
NDVI	0.074	0.077	0.067	0.053	0.047	0.041	0.036	0.063	0.093	0.095	0.647
Distance from River	0.056	0.051	0.045	0.053	0.047	0.041	0.036	0.032	0.046	0.072	0.479
Curvature	0.045	0.031	0.027	0.027	0.024	0.021	0.018	0.016	0.015	0.024	0.246

STEP-V : Determination of Eigen Value

Weighted Sum Value	Criteria Weight	Consistency Vector (Ratio)
2.36	0.22	10.58
1.63	0.15	10.60
1.42	0.13	10.62
1.14	0.11	10.65
0.99	0.09	10.53
0.86	0.08	10.41
0.75	0.07	10.31
0.65	0.06	10.24
0.48	0.05	10.34
0.25	0.02	10.32

APPENDIX B : Results of various Machine Learning Models

Statistical Evaluation results of Machine Learning models on Testing datasets

No	Metric	LR	KNN	SVM	Kernel SVM	Naïve Bayes	Decision Tree	Random Forest	XGBoost	ANN
1	TN	453	428	455	459	414	433	467	465	459
2	TP	419	405	406	413	422	440	431	429	420
3	FN	48	62	61	54	45	27	36	38	47
4	FP	16	41	14	10	55	36	2	4	10
5	Accuracy Rate (AR)	93.16%	89.00%	91.99%	93.16%	89.32%	93.27%	95.94%	95.51%	93.91%
6	Error Rate (ER)	6.84%	11.00%	8.01%	6.84%	10.68%	6.73%	4.06%	4.49%	6.09%
7	Accuracy - K-fold cross validation	92.80%	89.00%	92.90%	92.80%	89.32%	91.66%	94.64%	94.87%	93.32%
8	Sd - K-fold cross validation	1.17%	2.37%	1.76%	1.87%	3.14%	1.52%	1.36%	1.26%	0.81%
9	Sensitivity (SST)/Recall	89.72%	86.72%	86.94%	88.44%	90.36%	94.22%	92.29%	91.86%	89.94%
10	Specificity (SPF)	96.59%	91.26%	97.01%	97.87%	88.27%	92.32%	99.57%	99.15%	97.87%
11	Positive Predictive Value (PPV)/Precision	96.32%	90.81%	96.67%	97.64%	88.47%	92.44%	99.54%	99.08%	97.67%
12	Negative Predictive Value (NPV)	90.42%	87.35%	88.18%	89.47%	90.20%	94.13%	92.84%	92.45%	90.71%
13	Kappa (K)	0.863	0.780	0.840	0.863	0.786	0.865	0.919	0.910	0.878
14	AUC	0.975	0.949	0.974	0.980	0.960	0.933	0.982	0.984	0.982

Landslide Density Index calculation for various ML models

XGB						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	6	0.38	1912828	12.82	0.03
2	Low	130	8.33	4539109	30.41	0.27
3	Moderate	362	23.19	4428627	29.67	0.78
4	High	416	26.65	2353323	15.77	1.69
5	Very high	647	41.45	1691566	11.33	3.66
		1561		14925453		

KNN						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	5	0.32	2808305	18.82	0.02
2	Low	61	3.91	2705449	18.13	0.22
3	Moderate	162	10.38	3105502	20.81	0.50
4	High	283	18.13	2879028	19.29	0.94
5	Very high	1050	67.26	3427169	22.96	2.93
		1561		14925453		

NB						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	16	1.02	3084398	20.67	0.05
2	Low	248	15.89	2717890	18.21	0.87
3	Moderate	181	11.60	1897733	12.71	0.91
4	High	351	22.49	3917930	26.25	0.86
5	Very high	765	49.01	3307502	22.16	2.21
		1561		14925453		

K-SVM						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	231	14.80	3457755	23.17	0.64
2	Low	117	7.50	1915883	12.84	0.56
3	Moderate	259	16.59	4394400	29.44	0.58
4	High	430	27.55	3613985	24.21	1.14
5	Very high	524	33.57	1543430	10.34	3.25
		1561		14925453		

LR						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	108	6.92	3090139	20.70	0.33
2	Low	189	12.11	3466489	23.23	0.52
3	Moderate	215	13.77	2485314	16.65	0.83
4	High	412	26.39	3497764	23.43	1.13
5	Very high	637	40.81	2385747	15.98	2.55
		1561		14925453		

ANN						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	123	7.88	3156442	21.15	0.37
2	Low	141	9.03	2751690	18.44	0.49
3	Moderate	344	22.04	4949755	33.16	0.66
4	High	425	27.23	2650777	17.76	1.53
5	Very high	528	33.82	1416789	9.49	3.56
		1561		14925453		

SVM						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	172	11.02	5221248	34.98	0.31
2	Low	93	5.96	1341287	8.99	0.43
3	Moderate	137	8.78	3017585	20.22	0.66
4	High	444	28.44	3122877	20.92	1.36
5	Very high	715	45.80	2222456	14.89	3.08
		1561		14925453		

DT						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	314	20.12	6740163	45.16	0.45
2	Low	396	25.37	4202956	28.16	0.90
3	Moderate	256	16.40	1837021	12.31	1.33
4	High	280	17.94	1291984	8.66	2.07
5	Very high	315	20.18	853329	5.72	3.53
		1561		14925453		

RF						
	Susceptibility	Nos of landslide	LSD %	Nos of pixels	Area %	LDI
1	Very low	1	0.06	3450033	23.12	0.00
2	Low	80	5.12	3585676	24.02	0.21
3	Moderate	106	6.79	1895082	12.70	0.53
4	High	450	28.83	3664347	24.55	1.17
5	Very high	924	59.19	2330315	15.61	3.79
		1561		14925453		

District wise Class Distribution

RF				
LABEL	KINNAUR	PART_OF_KULLU	PART_OF_MANDI	SHIMLA
1	2404278	65751	258816	721188
2	1308355	187783	421999	1667539
3	695774	121412	188566	889330
4	1621027	221550	315087	1506683
5	1140049	159215	140753	890298

XGB				
LABEL	KINNAUR	PART_OF_KULLU	PART_OF_MANDI	SHIMLA
1	1515676	14913	96019	286220
2	2526424	115530	439256	1457899
3	1789633	215949	453910	1969135
4	702228	223525	210886	1216684
5	636259	185605	125046	744656

ANN				
LABEL	KINNAUR	PART_OF_KULLU	PART_OF_MANDI	SHIMLA
1	791784	138401	562866	1663391
2	1513440	95213	232470	910567
3	2967645	212133	262914	1507063
4	1288971	206340	137028	1018438
5	606457	103670	130174	576488

DT				
LABEL	KINNAUR	PART OF KULLU	PART OF MANDI	SHIMLA
1	3915837	201415	452866	2170045
2	1622341	239975	452003	1888637
3	691978	134515	205029	805499
4	514208	108487	142659	526630
5	424359	71068	72604	285298

KSVM				
LABEL	KINNAUR	PART OF KULLU	PART OF MANDI	SHIMLA
1	847398	169305	546493	1894559
2	826186	82956	211676	795065
3	2622757	190140	276004	1305499
4	2190586	209301	152170	1061928
5	681084	104137	139074	619135

SVM				
LABEL	KINNAUR	PART OF KULLU	PART OF MANDI	SHIMLA
1	2477618	224890	555394	1963346
2	1178199	13134	31557	118397

3	1381523	139153	306437	1190472
4	1181206	204505	284132	1453034
5	950396	174058	147844	950158

KNN				
LABEL	KINNAUR	PART OF KULLU	PART OF MANDI	SHIMLA
1	1561277	99647	318495	828886
2	1146381	147598	286277	1125193
3	1364316	170260	278345	1292581
4	1398464	153067	218042	1109455
5	1698196	185282	224208	1319483

LR				
LABEL	KINNAUR	PART OF KULLU	PART OF MANDI	SHIMLA
1	2026610	72652	159928	830949
2	1809755	120833	274952	1260949
3	996074	126144	260667	1102429
4	1305862	237576	387312	1567014
5	1031344	198239	242078	914086

NB				
LABEL	KINNAUR	PART OF KULLU	PART OF MANDI	SHIMLA
1	2802689	36120	56926	188663
2	1094052	381819	101143	1140876
3	1076775	60417	256913	503628
4	1401077	97965	506766	1912122
5	794080	179319	403465	1930638

Combined Pixel Percentage for all Districts for zonal statistics

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
Kinnaur	48.03	28.19	RF	Very low	16.11
				Low	8.77
				Moderate	4.66
				High	10.86
				Very high	7.64
			XGB	Very low	10.15
				Low	16.93
				Moderate	11.99
				High	4.7
				Very high	4.26
			ANN	Very low	5.3
				Low	10.14
				Moderate	19.88
				High	8.64
				Very high	4.06
			DT	Very low	26.24
				Low	10.87
				Moderate	4.64
				High	3.45
				Very high	2.84
			KSVM	Very low	5.68
				Low	5.54
				Moderate	17.57
				High	14.68
				Very high	4.56
			SVM	Very low	16.6

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
				Low	7.89
				Moderate	9.26
				High	7.91
				Very high	6.37
			KNN	Very low	10.46
				Low	7.68
				Moderate	9.14
				High	9.37
				Very high	11.38
			LR	Very low	13.58
				Low	12.13
				Moderate	6.67
				High	8.75
				Very high	6.91
			NB	Very low	18.78
				Low	7.33
				Moderate	7.21
				High	9.39
				Very high	5.32
Part of Kullu	5.06	8.01	RF	Very low	0.44
				Low	1.26
				Moderate	0.81
				High	1.48
				Very high	1.07
			XGB	Very low	0.1
				Low	0.77
				Moderate	1.45
				High	1.5

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
				Very high	1.24
			ANN	Very low	0.93
				Low	0.64
				Moderate	1.42
				High	1.38
				Very high	0.69
			DT	Very low	1.35
				Low	1.61
				Moderate	0.9
				High	0.73
				Very high	0.48
			KSVM	Very low	1.13
				Low	0.56
				Moderate	1.27
				High	1.4
				Very high	0.7
			SVM	Very low	1.51
				Low	0.09
				Moderate	0.93
				High	1.37
				Very high	1.17
			KNN	Very low	0.67
				Low	0.99
				Moderate	1.14
				High	1.03
				Very high	1.24
			LR	Very low	0.49
				Low	0.81

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
				Moderate	0.85
				High	1.59
				Very high	1.33
			NB	Very low	0.24
				Low	2.56
				Moderate	0.4
				High	0.66
				Very high	1.2
Part of Mandi	8.88	7.69	RF	Very low	1.73
				Low	2.83
				Moderate	1.26
				High	2.11
				Very high	0.94
			XGB	Very low	0.64
				Low	2.94
				Moderate	3.04
				High	1.41
				Very high	0.84
			ANN	Very low	3.77
				Low	1.56
				Moderate	1.76
				High	0.92
				Very high	0.87
			DT	Very low	3.03
				Low	3.03
				Moderate	1.37
				High	0.96
				Very high	0.49

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
			KSVM	Very low	3.66
				Low	1.42
				Moderate	1.85
				High	1.02
				Very high	0.93
			SVM	Very low	3.72
				Low	0.21
				Moderate	2.05
				High	1.9
				Very high	0.99
			KNN	Very low	2.13
				Low	1.92
				Moderate	1.86
				High	1.46
				Very high	1.5
			LR	Very low	1.07
				Low	1.84
				Moderate	1.75
				High	2.59
				Very high	1.62
			NB	Very low	0.38
				Low	0.68
				Moderate	1.72
				High	3.4
				Very high	2.7
Shimla	38.03	56.12	RF	Very low	4.83
				Low	11.17
				Moderate	5.96

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
				High	10.09
				Very high	5.96
			XGB	Very low	1.92
				Low	9.77
				Moderate	13.19
				High	8.15
				Very high	4.99
			ANN	Very low	11.14
				Low	6.1
				Moderate	10.1
				High	6.82
				Very high	3.86
			DT	Very low	14.54
				Low	12.65
				Moderate	5.4
				High	3.53
				Very high	1.91
			KSVM	Very low	12.69
				Low	5.33
				Moderate	8.75
				High	7.11
				Very high	4.15
			SVM	Very low	13.15
				Low	0.79
				Moderate	7.98
				High	9.74
				Very high	6.37
			KNN	Very low	5.55

District	Area percentage	Landslide percentage	Model	Susceptibility	Pixel percentage
				Low	7.54
				Moderate	8.66
				High	7.43
				Very high	8.84
			LR	Very low	5.57
				Low	8.45
				Moderate	7.39
				High	10.5
				Very high	6.12
			NB	Very low	1.26
				Low	7.64
				Moderate	3.37
				High	12.81
				Very high	12.94

APPENDIX C : Publications

1. Pathak, Yashodhar & Prakash, Indra & Dholakia, M. (2023). CONVENTIONAL AND MODERN APPROACHES IN LANDSLIDE SUSCEPTIBILITY MAPPING: A METHODOLOGICAL REVIEW. International journal of applied engineering and technology (London). 5. 2633-4828. (Scopus Indexed)
2. Pathak, Y. P., Dholakia, M. B., Prakash, I., Jagad, G. D., & Gohil, K. B. (2024). Evaluating landslide susceptibility in the Lower Sutlej Basin, Himachal Pradesh, India using GIS: A comparative study of the Frequency Ratio and Analytical Hierarchy Process methods. *Nanotechnology Perceptions*, 20(S5), 719–742. <https://doi.org/10.62441/nano-ntp.v20iS5.67>. (Scopus Indexed)