APPLICALIBITY OF SUPERVISED MACHINE LEARNING ALGORITHM FOR IDENTIFICATION OF VARIOUS GEOMORPHOLOGICAL FEATURES IN HIMALAYAN REGION

Report submitted for completion of training

(Satellite Meteorology and OceAnography Research and Training)
Programme of Space Applications Centre

Submitted by
AAYUSHI VIJAYBHAI GOHIL
(Registration No. RS00865)
M.Sc Integrated (Artificial Intelligence And Machine Learning)

Under the guidance of

Dr. Sandip R Oza
GROUP DIRECTOR, SCI/ENG – G
CHSG/EPSA
Space Applications Centre, (ISRO) Ahmedabad

Institution
Gujarat University
Ahmedabad – 380015
Gujarat

SRTD-RTCG-MISA
Space **A**pplications **C**entre (ISRO)
Ahmedabad, Guajrat

10 April 2023 to 10 July 2013

Acknowledgement

I would like to express my sincere gratitude to everyone who contributed to the completion of this report on Applicability of Supervised Machine Learning algorithms for identification of various geomorphological features in Himalayan region.

I am thankful to Dr. S.P Vyas, Head, SRTD for giving me the opportunity to work at their esteemed organization and encouraging me to push my boundaries during this training period.

I express my sincere gratitude to my mentor Dr.Sushil Kumar Singh, SCI/ENGR-SG, senior Madhukar Shrigyan and my colleagues who provided valuable guidance, advice, and kind support throughout this project.

I am also grateful to the director, Space Applications Centre (SAC), Indian Space Research Organisation (ISRO) for allowing me to carry out the research work.

I thank my all the faculty members of the University for their kind Support. Finally, I am also thankful to my friends and family who always supported me.

I would also like to extend my thanks to the authors of the papers and research articles that I consulted during my research. Their work provided me with the necessary background knowledge and insights to understand and analyse the topic in depth.

Once again, thank you to everyone who contributed to this report.

With Sincere regards,
Aayushi V. Gohil

Table of Contents

Acknowledgement 2
ABSTRACT 5
INTRODUCTION 6
Project Summary: 6
Objective: 7
Purpose: 7
Himalaya Region: 7
Remote Sensing of the Himalayan Region: 8
Geomorphological Classification: 9
Machine Learning: 9
Supervised Machine Learning: 10
Classification Module of Supervised Machine Learning: 11
STUDY AREA AND DATA 12
Study Area: 12
Data: 13
LISS-3 (Linear Imaging Self-Scanning Sensor): 13
LANDSAT-8: 14
SRTM (Shuttle Radar Topography Mission): 15
METHODOLOGY 16
Pre-Processing: 16
DN to Radiance Conversion: 16
Radiance to Reflectance Conversion: 17
ADDING LAYER OF SRTM MODEL: 18
GEO-REGISTRATION: 18
OVERLAYING OF DATA: 19
Algorithms: 20
Random Forest Algorithm: 20
Gradient Boost Algorithm: 22
Confusion Matrix: 24
RESULT AND DISCUSSION 26
The outcomes or findings obtained from Random Forest Classification: 26
Comparison of Random Forest and Gradient Boost Classifier: 30
Discussion: 30
CONCLUSION 32

REFERENCES 34

List of Figures

Figure 1 Study Area, in and around Chandra Basin depicted in FCC of LISS-3 data 12
Figure 2 Training and Testing of Random Forest Algorithm 20
Figure 3 The image of the study area was categorized using the Random Forest Classifier 26
Figure 4 The testing dataset's confusion matrix for a Random Forest Classifier 27
Figure 5 The image of study area was categorized using Gradient Boost Classifier 28
Figure 6 The testing dataset's confusion matrix for a Gradient Boost Classifier 29

List of Tables

Table 1 LISS-3 Sensor Spectral Bands information 13
Table 2 LANDSAT-8 Spectral Band Information 14

ABSTRACT

The Himalayan region is known for its diverse and complex geomorphological features, which play a crucial role in various environmental and geological studies. Identifying and mapping these features accurately is essential for effective land management, hazard assessment, and resource planning. In this project, we explored the applicability of supervised machine learning algorithms, specifically the random forest and gradient boost classifiers, for the identification of various geomorphological features in the Himalayan region.

Using a carefully curated dataset of remote sensing imagery, we trained and evaluated the random forest and gradient boost classifiers. Through the utilization of feature extraction techniques and appropriate preprocessing steps, we prepared the dataset for classification. The classifiers were then trained using labelled data and their performance was assessed through confusion matrices and accuracy metrics.

The results demonstrated the effectiveness of both classifiers in identifying geomorphological features. However, a comparative analysis revealed that the gradient boost classifier outperformed the random forest classifier in terms of accuracy, particularly in distinguishing challenging features such as shadows and debris. The gradient boost classifier achieved an overall accuracy of 98.5%, surpassing the random forest classifier's accuracy of 98%.

These findings highlight the potential of supervised machine learning algorithms for accurate feature identification in the Himalayan region. The successful application of these algorithms offers promising opportunities for improved land management, geological studies, and hazard assessment in this geologically dynamic region.

It is important to note that the performance of these classifiers may be influenced by factors such as dataset quality, feature selection, and parameter tuning. Further research and experimentation are recommended to validate and refine these results, potentially exploring the integration of other advanced machine learning techniques and additional datasets for enhanced accuracy and robustness.

Overall, this project contributes to the growing field of remote sensing and machine learning applications in geomorphology, providing insights into the applicability of supervised machine learning algorithms for the identification of various geomorphological features in the Himalayan region.

INTRODUCTION

Project Summary:

The study titled "Applicability of Supervised Machine Learning Algorithms to Identify Various Geomorphological Features in the Himalaya Region" investigates the application of supervised machine learning algorithms to classify and recognize various geomorphological characteristics in the Himalayas.

The project begins by extensively reviewing existing research and methodologies in this area to gain a comprehensive understanding. It then focuses on categorizing the diverse types of geomorphological features present in the Himalaya region, such as valleys, ridges, glaciers, landslides, and river channels.

To facilitate the analysis, a diverse dataset is collected, including satellite imagery and topographic data. The dataset undergoes preprocessing and cleaning procedures to ensure its quality and compatibility with the selected supervised machine learning algorithms.

For training and testing the machine learning algorithms, a file containing ROI (Region of Interest) samples is created using ARCGIS software. This file contains valuable information on various geomorphological features.

The Random Forest Classifier and Gradient Boost Classifier, both supervised machine learning algorithms, are employed to identify and classify the different geomorphological features. The accuracy score of each algorithm and a confusion matrix are calculated to evaluate the accuracy of the algorithms in identifying each geomorphological feature.

Based on the findings, the project provides recommendations for the practical implementation and utilization of supervised machine learning algorithms in identifying and classifying geomorphological features in the Himalaya region. Ultimately, the project aims to contribute to the understanding and application of machine learning techniques in studying and managing geomorphological features in the Himalayas.

Objective:

Main objective of the project is to perform a comprehensive analysis and categorization of the diverse geomorphological features in the Himalaya region. This includes identifying features like vegetation, water body, snow and ice, debris and bare terrain or bare land. . By gaining a deep understanding of these features and categorizing them effectively, the project establishes the basis for subsequent stages in training machine-learning algorithms to identify and classify them accurately.

Purpose:

Due to the challenging terrain of the Himalayan cryosphere, remote sensing plays a crucial role in this region, as it is impractical to gather ground truth data for every area on a regular basis. Hence, supervised machine learning classification algorithms can be employed to categorize specific satellite images into different geomorphological features.

Himalaya Region:

The Himalayas possess the largest reserves of snow, ice, and glaciers outside of the polar region. These glaciers are distributed throughout the Himalayas, extending from Kashmir in the West to Arunachal Pradesh in the East, covering the entire stretch of Himachal Pradesh, Uttarakhand, Nepal, Sikkim, and Bhutan.

The high-altitude regions of the Himalayas include glaciers, both the pro-glacial region and the areas below, which are covered with seasonal snow during the winter months. A significant portion of the Himalayas is consistently covered in snow and ice, while a larger part becomes snow-covered during winter. When summer arrives, the winter snow precipitation melts, contributing to the discharge of the rivers originating from these regions.

The Himalayan cryosphere plays a crucial role in global climate change studies and the regulation of freshwater supply to major Indian rivers like Ganga, Brahmaputra. Comprehending the Himalayan cryosphere holds great importance in evaluating the consequences of climate change, effectively managing water resources, and preserving the distinctive ecosystems found in this area.

Remote Sensing of the Himalayan Region:

Remote Sensing is a way of gathering information about Earth's surface physically being there. It involves using special sensors, like cameras, satellites to collect data from distance.

Himalayan region has very rugged topography. The Himalayas span a wide range of altitudes, from around 1000m to the towering height of 8848m at Mount Everest. The areas below 2500m to 3500m are mostly sparsely populated. Beyond this altitude, the climatic conditions in the remote parts of the Himalayas become extremely harsh, posing significant challenges in terms of accessibility.

The utilisation of remote sensing allows scientists to observe and analyse the size, movement, and transformations of glaciers, as well as to map and monitor the extent of snow cover. It also assists in estimating the balance of mass in glaciers, measuring surface temperatures of ice-covered areas, and monitoring the risks associated with glacial lakes.

Furthermore, remote sensing provides valuable insights into the challenging topography of the Himalayas, aiding in the creation of terrain maps and evaluations of accessibility.

By employing remote sensing techniques, researchers can gather comprehensive data on the Himalayan cryosphere, contributing to the comprehension of climate change impacts, the management of water resources, and conservation initiatives within the region.

Geomorphological Classification:

In non-technical language, Geomorphological classification is “the classification of various surface features of the earth.

Geomorphological classification entails the identification and categorization of diverse landforms and features within a particular geographic area or region. It offers a structured framework that facilitates the organisation and comprehension of the wide array of landforms found on the Earth's surface.

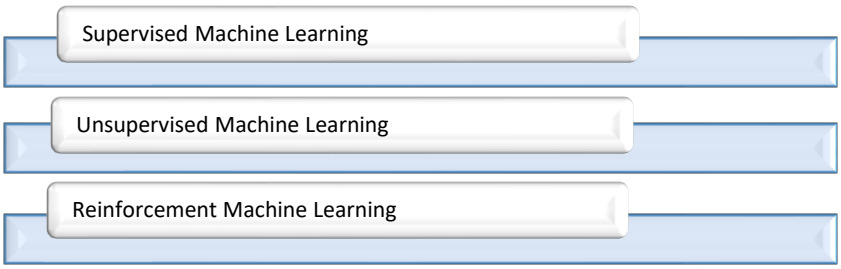
Machine Learning:

Machine Learning is a branch of Artificial Intelligence (AI) that is focusing on analysing and interpretation of the data.

The main goal of Machine Learning is “enables a machine to automatically learn from data, improve from experiences and predict things automatically without being explicitly programmed to do so.”

Machine learning employs diverse algorithms to construct mathematical models and make predictions based on historical data or information.

There are mainly three types of Machine Learning:



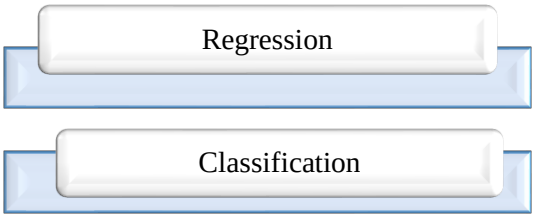
Supervised Machine Learning:

Supervised learning refers to a category of machine learning where machines are trained using labelled training data, enabling them to make predictions based on the provided data. The term "labelled data" implies that the input data is already associated with the correct output.

In supervised learning, the training data acts as a supervisor, instructing the machines “how to accurately predict the output”. This parallels the way a student learns under the guidance of a teacher.

Supervised learning involves providing both input data and corresponding output data to the machine-learning model. The objective of a supervised learning algorithm is to discover a mapping function that can associate the input variable (x) with the output variable (y).

Supervised Machine Learning can further divide into two parts:



Classification Module of Supervised Machine Learning:

Classification is a crucial task in which input data is assign to predefined classes or categories. The primary objective of classification algorithms is to establish a connection between input features and corresponding labels, allowing the algorithm to assign accurately new, unseen data points to their appropriate classes.

In the classification process, the algorithm receives training data that comprises labelled examples, with each example associated with a specific class or category. By learning from these labelled examples, the algorithm identifies common patterns and generalises them to make predictions on unseen data.

During the training phase, the classification algorithm carefully examines the input features and detects distinctive patterns or relationships that differentiate the various classes. It then constructs a model based on these patterns, which can subsequently be utilised to classify new instances effectively.

The ultimate goal of classification within supervised machine learning is to develop a dependable model capable of accurately categorising unseen data instances into their respective classes.

STUDY AREA AND DATA

Study Area:

Training and testing both of the classification algorithm done on particular scene which includes Chandra basin of Himachal Pradesh and its surrounding area

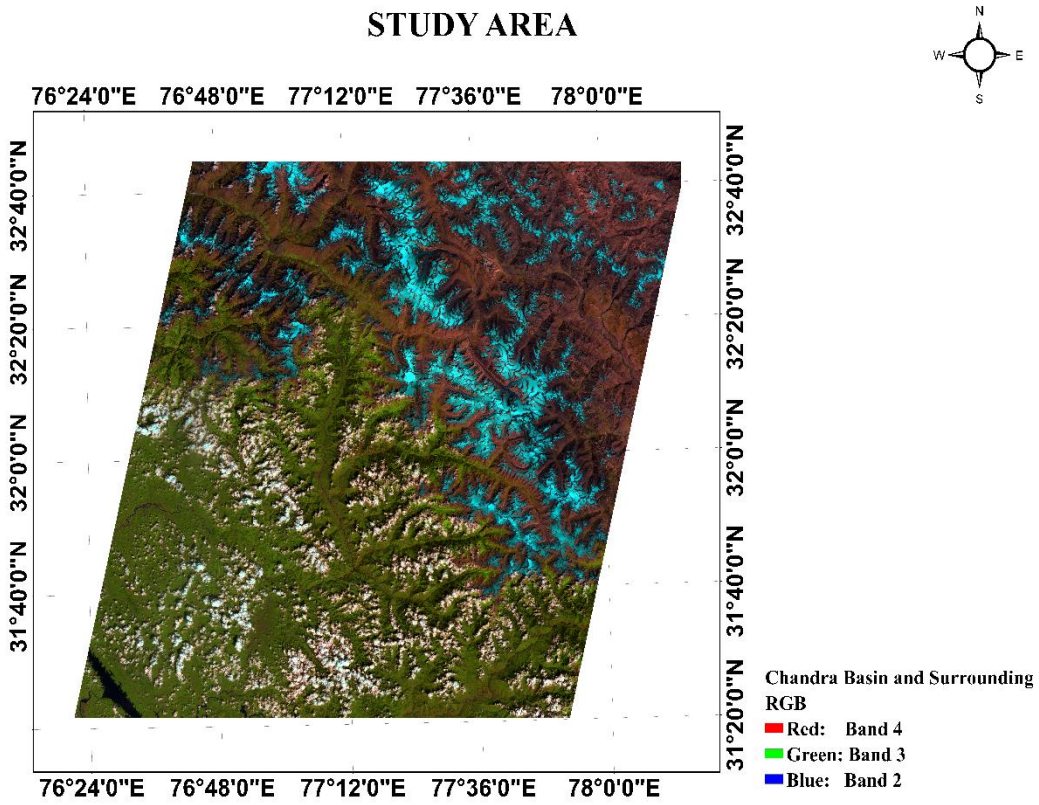


Figure 1 Study Area, in and around Chandra Basin depicted in FCC of LISS-3 data

Chandra basin is home to several glaciers and some notable glaciers of Chandra basin is Bara Shigri, Samudra Tapu, and Chhota Shigri etc. These glaciers contribute significantly to the flow of water in the Chandra River, and they play a vital role in the region's hydrology and ecosystem. They are also important indicators of climate change and are subject to scientific study to understand their behaviour and response to changing environmental conditions.

From visual Interpretation of study area, it has various geomorphological features such as Vegetation, Water Body, Ice, Snow, Bare Terrain etc.

Data:

Optical imagery from Resourcesat-2 satellite's LISS-3(Linear Imagery Self Scanning) sensor acquired in 8 September 2019 is used. OLI imagery data from LandSat-8 satellite acquired in September, 2019 was used. Scene from September month is used for snow free data.

LISS-3 (Linear Imaging Self-Scanning Sensor):

Sensor Type: LISS-3 is a multispectral sensor that operates in the visible and near-infrared spectrum.

Spatial Resolution: The LISS-3 sensor has a spatial resolution of 23.5 meters, which means each pixel in the image represents an area of 23.5 square meters on the ground.

Swath Width: LISS-3 has a swath width of 141 kilometres, which refers to the total width of the area covered by the sensor in a single pass.

The LISS-3 sensor has played a significant role in enhancing India's capabilities in remote sensing and has contributed to a wide range of applications, including environmental monitoring, agriculture, and natural resource management.

Bands	Wavelength(µm)	Spatial Resolution (m)
Band 2 - Green	**0.52 - 0.59**	**23.5**
Band 3 - Red	**0.62 - 0.68**	**23.5**
Band 4 - Near Infrared (NIR)	**0.77 - 0.86**	**23.5**
Band 5 - Shortwave Infrared(SWIR)	**1.55 – 1.70**	**23.5**

Table 1 LISS-3 Sensor Spectral Bands information

LANDSAT-8:

Sensor Type: Landsat-8 is equipped with the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). These sensors capture data across multiple spectral bands, providing a comprehensive view of the Earth's surface.

Spatial Resolution: The OLI sensor achieves a spatial resolution of 30 meters for most spectral bands, while the panchromatic band (Band 8) offers a higher resolution of 15 meters. The TIRS sensors provide a spatial resolution of 100 meters.

Bands	Wavelength (µm)	Spatial Resolution (m)
Band 1 – Costal aerosol	**0.43 - 0.45**	**30**
Band 2 -Blue	**0.45 - 0.51**	**30**
Band 3 - Green	**0.53 - 0.59**	**30**
Band 4 - Red	**0.64 - 0.67**	**30**
Band 5 - Near Infrared (NIR)	**0.85 - 0.88**	**30**
Band 6 - SWIR 1	**1.57 - 1.65**	**30**
Band 7 - SWIR 2	**2.21 - 2.29**	**30**
Band 8 - Panchromatic	**0.50 - 0.68**	**30**
Band 9 - Cirrus	**1.36 - 1.38**	**30**
Band 10 - Thermal Infrared(TIRS) 1	**10.60 - 11.19**	**30**
Band 11 - Thermal Infrared(TIRS) 2	**11.50 - 12.51**	**30**

Table 2 LANDSAT-8 Spectral Band Information

SRTM (Shuttle Radar Topography Mission):

The SRTM Digital Elevation Model (DEM) is a highly detailed dataset that offers precise elevation information for the Earth's surface. It was generated through the Shuttle Radar Topography Mission carried out by the Space Shuttle Endeavour in 2000 under NASA.

Utilizing radar technology, the SRTM DEM captures accurate elevation data, even in areas with dense vegetation or cloud cover. By employing a specially designed radar system, the mission measured the time taken for radar signals to bounce off the Earth's surface and return to the spacecraft. Analysis of these signals enabled the creation of a precise digital representation of the topography.

With global coverage, including the Himalaya region, the SRTM DEM provides a spatial resolution of approximately 30 meters. This means that each pixel in the dataset corresponds to an area of 30 meters by 30 meters on the ground, allowing for the identification of various geomorphological features such as valleys, mountains, ridges, and other landforms.

The SRTM DEM has become an invaluable resource for diverse applications, including geospatial analysis, hydrological modelling, terrain mapping, and natural resource management. Its wide availability and accuracy make it an indispensable tool for researchers, planners, and decision-makers working in fields related to geology, geography, environmental sciences, and engineering.

Overall, the SRTM DEM provides a reliable and comprehensive dataset of elevation information, encompassing the Himalaya region, which plays a crucial role in understanding and analysing the Earth's surface topography.

METHODOLOGY

Pre-Processing:

DN to Radiance Conversion:

DN (Digital Number) to radiance conversion is a process used in remote sensing and satellite imagery analysis to convert the digital numbers recorded by sensors into radiance values. Digital numbers are essentially the discrete values that represent the brightness or intensity of pixels in an image.

To convert DN to radiance, a calibration process is performed using sensor-specific calibration coefficients. These coefficients are derived during the sensor calibration process and are provided by the satellite or sensor manufacturer. They account for factors such as sensor gain, offset, and other parameters specific to the sensor.

The conversion formula typically involves a linear transformation, where the DN values are multiplied by a scaling factor (gain) and then adjusted by an offset. This process ensures that the resulting radiance values are in appropriate units (e.g., watts per square meter per steradian per micrometre).

The calibration coefficients are specific to each sensor and can vary depending on the spectral band being considered. Therefore, the DN to radiance conversion is typically band-dependent. Multiple conversions may be required if working with multispectral or hyperspectral imagery, as each band might have its own set of calibration coefficients.

Formula of DN to Radiance Conversion:

$$L_\lambda = \frac{LMAX_\lambda - LMIN_\lambda}{QCALMAX - QCALMIN}(QCAL - QCALMIN) + LMIN_\lambda$$

L_λ = Spectral Radiance at the Sensor

QCAL = the quantized calibrated pixel value in DN

QCALMIN = the minimum quantized calibrated pixel value in DN

QCALMAX = the maximum quantized calibrated pixel value in DN

$LMIN_\lambda$ = the spectral radiance that is scaled to QCALMIN

$LMAX_\lambda$ = the spectral radiance that is scaled to QCALMAX

Radiance to Reflectance Conversion:

Radiance to reflectance conversion is a process used in remote sensing and satellite imagery analysis to transform radiance measurements captured by sensors into reflectance values. Radiance refers to the amount of electromagnetic radiation emitted or reflected by a surface, while reflectance represents the fraction of incident radiation that is reflected by a surface.

The conversion from radiance to reflectance involves accounting for various factors such as atmospheric conditions, solar angle, sensor characteristics, and surface properties. These factors can introduce variations in radiance measurements, making it challenging to compare and analyse different images.

By converting radiance to reflectance, the resulting values become more standardized and can be directly compared between different images, sensors, and over time. Reflectance values are particularly useful for quantitative analysis, such as vegetation indices, land cover classification, and monitoring environmental changes.

Formula for Radiance to Reflectance Conversion:

$$\rho_p = \frac{\pi \cdot L_\lambda \cdot d^2}{ESUN_\lambda \cdot cos\theta_s}$$

ρ_p = Unitless planetary reflectance
L_λ = Spectral Radiance at the sensor's aperture
d = Earth-Sun distance in astronomical units
$ESUN_\lambda$ = Mean solar exoatmospheric irradiance
θ_s = Solar Zenith angle in degree

ADDING LAYER OF SRTM MODEL:

The characteristics of debris and bare terrain appear similar in satellite imagery. To differentiate between them, the slope of the terrain is determined using the SRTM (Shuttle Radar Topography Mission) Digital Elevation Model (DEM) data. This is accomplished by utilizing the QGIS software to calculate the slope values. Subsequently, the slope information is incorporated as a fifth layer in the data, alongside the four bands derived from the LISS-3 sensor.

GEO-REGISTRATION:

In simple terms, geo-registration involves aligning digital images or spatial data with their accurate geographic positions on the Earth's surface. It requires matching the coordinates of the image or data to known reference points or a coordinate system.

Put simply, geo-registration is like placing a digital image or dataset onto a map so that it accurately corresponds to the real-world locations it represents. This is crucial for precise analysis, visualization, and integration of spatial data from different sources.

The process of geo-registration typically involves using control points, which are identifiable features in the image or dataset that have known geographic coordinates. By matching these control points to their corresponding locations on a reference map or coordinate system, the image or dataset can be properly aligned in a georeferenced manner.

Once the geo-registration is completed, the digital image or dataset can be overlaid with other spatial data, such as maps or satellite imagery. This allows for confident spatial analysis and measurements, as everything is now correctly aligned in geographic space.

Geo-registration is done using ARCGIS software automatically.

OVERLAYING OF DATA:

Overlaying two satellite images is crucial because it enables us to observe temporal changes, validate data accuracy, merge diverse datasets, and enhance the visualization of specific features. It facilitates the monitoring of landscape transformations, confirms data reliability, leverages dataset strengths, and improves our comprehension of areas or phenomena. Ultimately, overlaying satellite images supports efficient analysis, monitoring, and decision-making across various domains.

Due to the aforementioned benefits, after performing geo-registration, the data from the LISS-3 sensors, along with the slope layer of SRTM DEM model, is superimposed or overlaid onto the Landsat-8 imagery.

Algorithms:

Random Forest Algorithm:

The Random Forest Algorithm is a widely used supervised machine learning technique for addressing Classification and Regression problems in the field of Machine Learning. It is highly favoured due to its popularity.

In this algorithm, a forest is created, consisting of multiple trees, and the algorithm becomes more robust as the number of trees increases. The Random Forest Algorithm is particularly effective in achieving high accuracy and enhancing problem-solving capabilities.

It functions as a classifier by utilising several decision trees that are built on different subsets of the given dataset. The algorithm takes the average of the outputs from these trees to enhance the predictive accuracy of the dataset.

The underlying principle of the Random Forest Algorithm is based on ensemble learning, which involves combining multiple classifiers to tackle complex problems and improve the overall performance of the model.

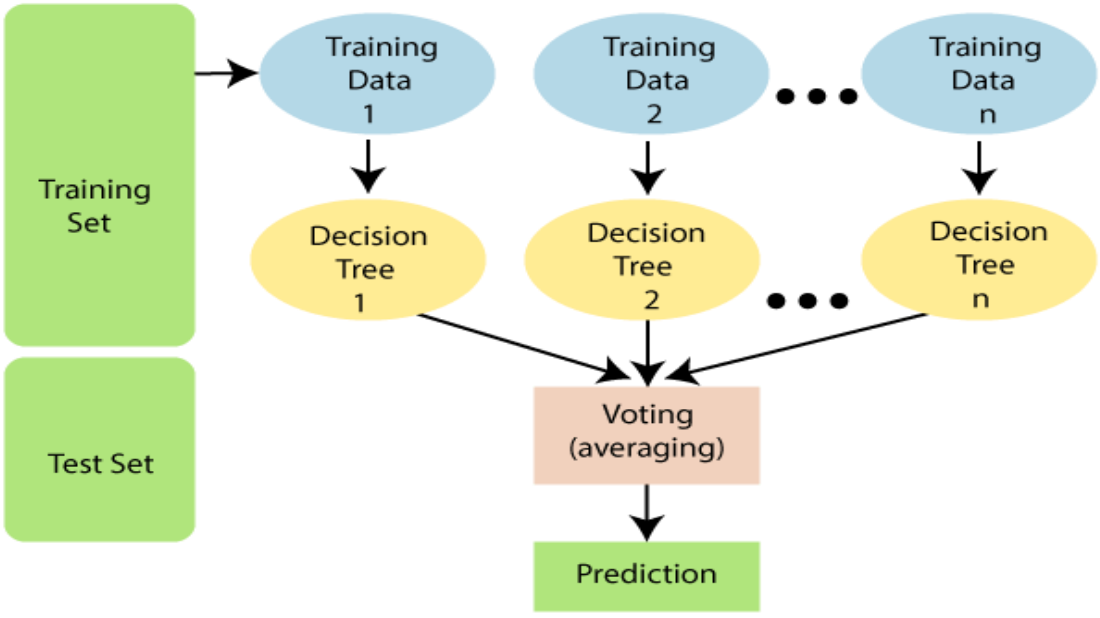


Figure 2 Training and Testing of Random Forest Algorithm

The Random Forest Algorithm's operation is described in the phases that follow:

Step 1: Choose random samples from a specified data collection or training set.

Step 2: For each training set of data, this algorithm will build a decision tree.

Step 3: Voting will be conducted using an average of the decision tree.

Step 4: Finally, choose the prediction result that received the most votes as the final prediction result.

The predictive power of the algorithm is improved by utilising the following hyperparameters:

- - n_estimators: This hyperparameter determines the number of trees that are constructed by the algorithm before averaging their outputs.
- - max_features: It specifies the maximum number of features that the random forest considers when deciding to split a node.
- The following hyperparameters are employed to optimise the speed of the model:
- - n_jobs: This hyperparameter informs the engine about the number of processors it can utilise. A value of 1 indicates that only one processor can be used, while a value of -1 implies that there is no limit.
- - random_state: It manages the randomness of the sample. If the random_state has a specific value and the model is provided with the same hyperparameters and training data, it will consistently produce the same results.

Gradient Boost Algorithm:

Gradient Boosting is an iterative algorithm that utilizes functional gradients to sequentially choose a function that moves in the direction of a weak hypothesis or negative gradient. This process aims to minimize a given loss function. The Gradient Boosting classifier merges multiple weak learning models to generate a robust predictive model with enhanced performance.

"Gradient Boosting comprises of three fundamental components:

1. Loss Function:

The loss function is responsible for assessing the predictive performance of the model based on the available data. Its specific form may vary depending on the particular problem at hand.

2. Weak Learner:

A weak learner classifies the data but tends to make numerous mistakes in the process. Typically, decision trees are employed as weak learners.

3. Additive Model:

The additive model represents the iterative and sequential addition of trees. Each iteration brings the model closer to its final form.

Steps in Gradient Boosting:

The Gradient Boosting classifier involves the following steps:

1. Model Fitting: The initial model is fitted to the data.

2. Hyperparameter and Parameter Adaptation: The model's hyperparameters and parameters are adjusted accordingly.

3. Prediction Making: Predictions are generated using the adapted model.

4. Interpretation of Results: The findings obtained from the predictions are interpreted.

Intuitive Understanding of Gradient Boost Algorithm:

(1) Initially, the algorithm computes the logarithm of the odds to make early predictions on the data. Typically, this is achieved by taking the ratio of the number of True values to the number of False values.

(2) For instance, if we have a dataset of six cancer occurrences, with four individuals having cancer and three being cancer-free, the log(odds) would be log (4/3) = 1.3. A cancer-free individual would have a log(odds) value of 0, while an individual with cancer would have a value of 1.

(3) To make predictions, the log(odds) is transformed into a probability using a logistic function. In this case, it would be approximately equal to 1.3, which matches the log(odds) value of 1.3.

(4) As the value is greater than 0.5, the algorithm employs 1.3 as the baseline estimate for each occurrence.

(5) The above formula is used to calculate the residuals for each instance in the training set.

(6) Subsequently, a Decision Tree is constructed to predict the estimated residuals.

(7) While creating the decision tree, a maximum number of leaves can be specified, resulting in two possible outcomes: instances being grouped together in the same leaf or the leaf containing more than one instance. The values are adjusted using the following formula:

ΣResidual / Previous Probability (1 - Previous Probability)

(8) The next steps involve obtaining the logarithmic prediction for each instance in the training set and transforming it into a probability.

(9) The final prediction is generated using the following formula:

base_log_odds + (learning_rate * predicted residual value)

The Gradient Boosting classifier has several important parameters that can significantly impact its performance. Here are some key parameters to consider when using the Gradient Boosting classifier:

(1) **n_estimators**: This parameter determines the number of boosting stages or trees to be built. Increasing the number of estimators generally improves the

model's performance, but it also increases the training time. It's important to find a balance to avoid overfitting or excessive computation.

(2) **learning_rate**: The learning rate controls the contribution of each tree in the ensemble. A lower learning rate means each tree has a smaller impact, requiring more trees to achieve the same overall effect. It helps prevent overfitting but may lead to longer training times.

(3) **max_depth**: It defines the maximum depth or the maximum number of levels in each individual decision tree. Deeper trees can capture more complex relationships in the data but may also overfit. Setting an appropriate value based on the complexity of the problem and available data is crucial.

Confusion Matrix:

A confusion matrix is a performance evaluation metric that is commonly used in machine learning and classification tasks. It provides a detailed breakdown of the predictions made by a classification model, allowing us to assess its accuracy and understand the types of errors it is making.

The confusion matrix is usually presented as a table with four different components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Each component represents a specific outcome of the classification process:

- True Positives (TP): These are the cases where the model correctly predicted the positive class (e.g., a disease is present) when it was indeed present in the actual data.

- True Negatives (TN): These are the cases where the model correctly predicted the negative class (e.g., a disease is absent) when it was indeed absent in the actual data.

- False Positives (FP): These are the cases where the model incorrectly predicted the positive class when it was actually negative. It is also known as a Type I

error. For example, the model predicted the presence of a disease, but the person was actually healthy.

- False Negatives (FN): These are the cases where the model incorrectly predicted the negative class when it was actually positive. It is also known as a Type II error. For example, the model predicted the absence of a disease, but the person was actually sick.

The confusion matrix allows us to calculate various performance metrics based on these four components, including:

- **Accuracy**: It measures the overall correctness of the model's predictions, calculated as (TP + TN) / (TP + TN + FP + FN).

- **Precision**: It quantifies the proportion of correctly predicted positive cases out of all predicted positive cases, calculated as TP / (TP + FP). Precision is useful when the cost of false positives is high.

- **Recall (Sensitivity or True Positive Rate)**: It represents the proportion of correctly predicted positive cases out of all actual positive cases, calculated as TP / (TP + FN). Recall is helpful when the cost of false negatives is high.

- **Specificity (True Negative Rate)**: It measures the proportion of correctly predicted negative cases out of all actual negative cases, calculated as TN / (TN + FP). Specificity complements recall and is valuable when the focus is on correctly identifying negative cases.

- **F1 Score**: It is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is calculated as 2 * (Precision * Recall) / (Precision + Recall).

By examining the confusion matrix and these performance metrics, we can gain insights into the strengths and weaknesses of the classification model. It helps in understanding the types of errors made by the model and allows us to make informed decisions about potential improvements or adjustments to the model or the classification threshold.

RESULT AND DISCUSSION

- The classifier successfully identified various geomorphological features, including vegetation, water bodies, debris, snow, ice, bare terrain, terrain's shadow and land. Additionally, the satellite image incorporated the presence of clouds and their shadows, which influenced the classification. Consequently, the classified image encompasses additional classes such as clouds and their shadows on vegetation, snow or ice, as well as bare terrain or land.

The outcomes or findings obtained from Random Forest Classification:

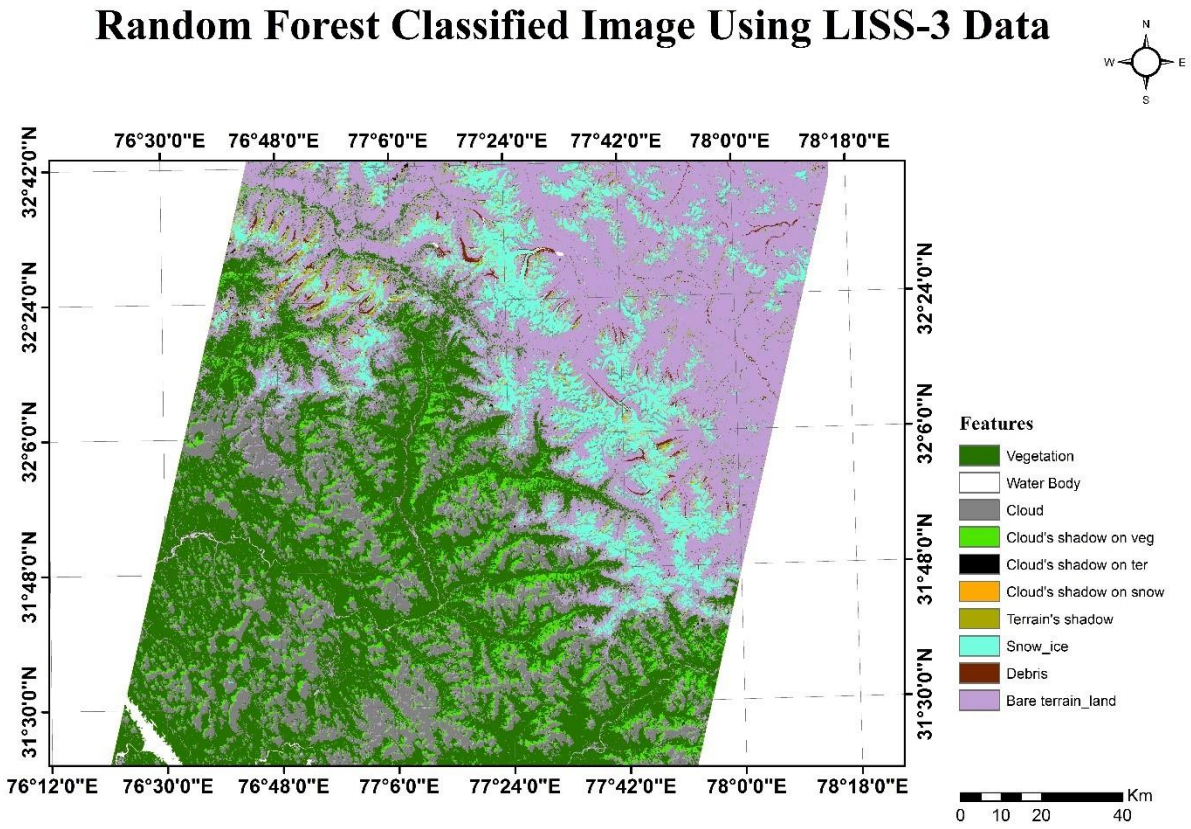


Figure 3 The image of the study area was categorized using the Random Forest Classifier

Figure 3 displays the classified image of diverse geomorphological features using the Random Forest classifier, achieving an impressive 98% accuracy for the testing dataset.

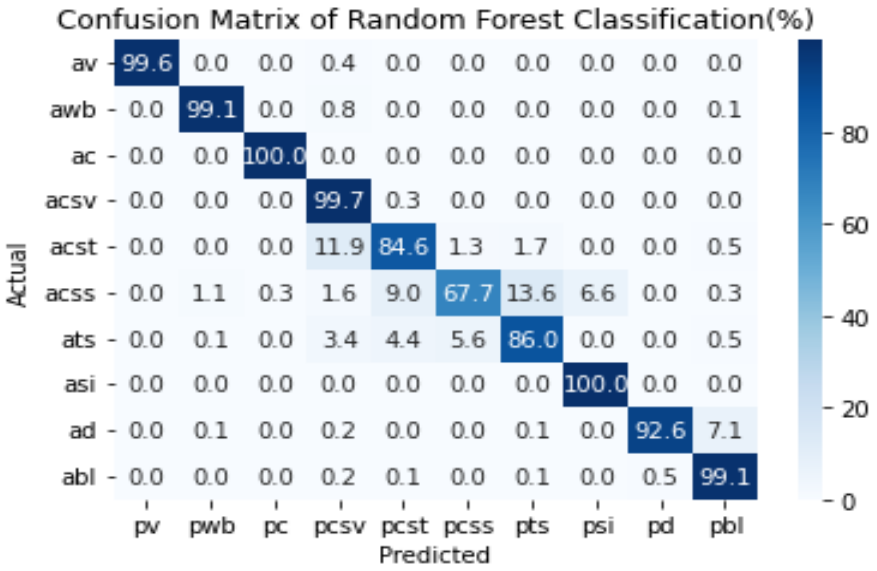


Figure 4 The testing dataset's confusion matrix for a Random Forest Classifier

Figure 4 illustrates the confusion matrix of random forest classification applied to the testing dataset. The x-axis represents the predicted values, while the y-axis corresponds to the actual values for the testing dataset of the random forest classifier. In the confusion matrix, the labels are abbreviated as follows: "v" for vegetation, "wb" for water body, "c" for cloud, "csv" for cloud's shadow on vegetation, "cst" for cloud's shadow on bare terrain or bare land, "css" for cloud's shadow on snow or ice, "ts" for terrain's shadow, "si" for snow and ice, "d" for debris, and "bl" for bare land or bare terrain.

The confusion matrix provides insights into the accuracy of the testing dataset across different features. It reveals that the accuracy is generally low for the shadow of cloud on bare terrain or land, shadow of cloud on snow or ice, and terrain's shadow. The primary reason behind this low accuracy is that these three classes share similar features, leading to incorrect predictions and consequently lower classification accuracy. This pattern is also observed for debris and bare terrain or bare land. According to the confusion matrix, the accuracy for the debris feature is 92.6%; however, it also misclassifies 7.1% of instances as bare terrain or bare land since the features of these two classes are similar.

The accuracy of certain features, namely vegetation, water body, cloud, cloud's shadow on vegetation, snow and ice, and bare terrain or bare land, is relatively higher compared to other features.

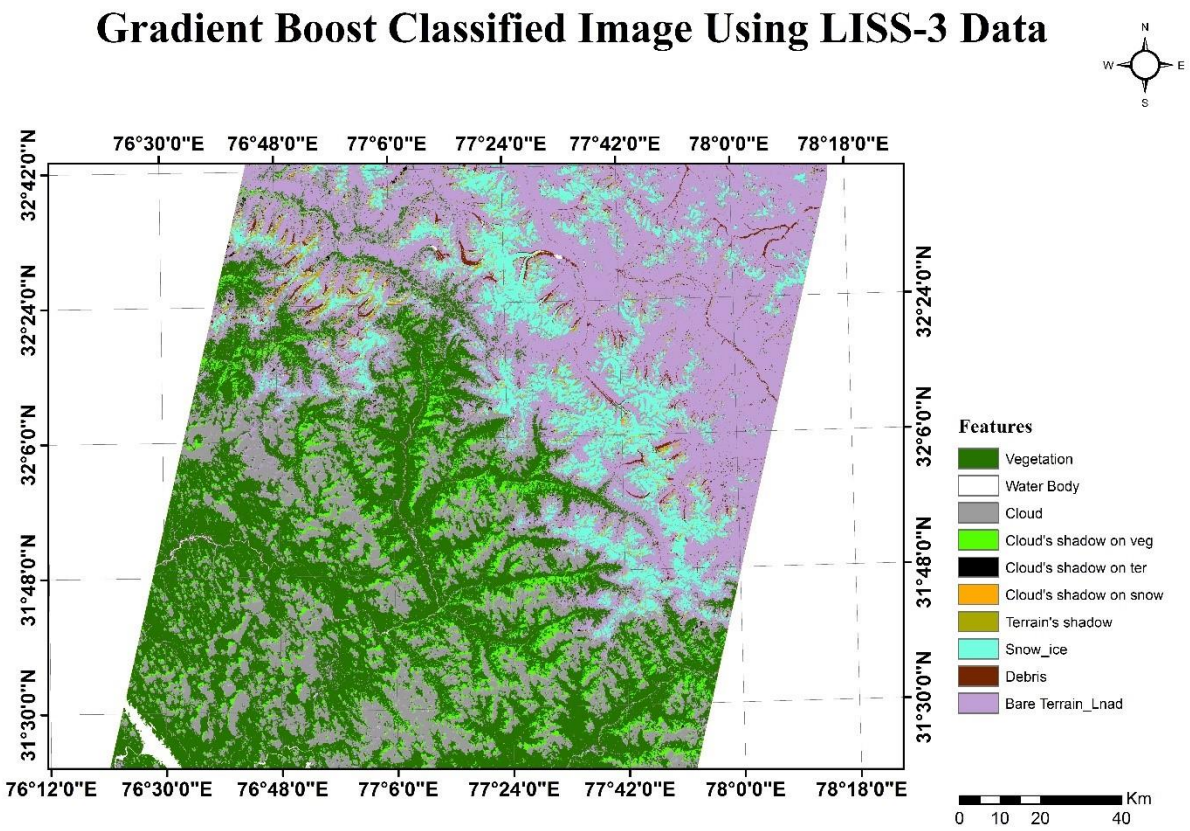


Figure 5 The image of study area was categorized using Gradient Boost Classifier

Figure 5 depicts the classified map of geomorphological features achieved through the utilization of a gradient boost classifier. The accuracy of this classifier in predicting geomorphological features stands at an impressive 98.5%.

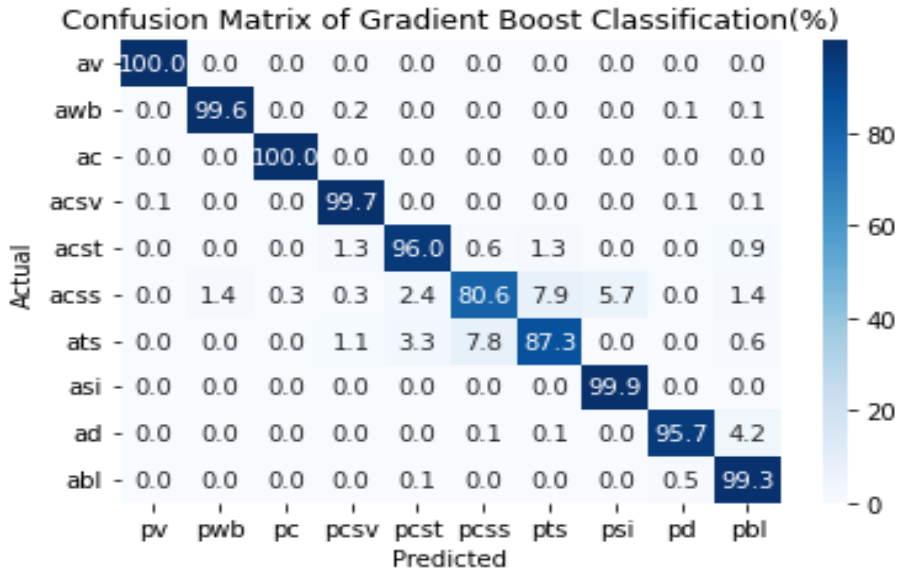


Figure 6 The testing dataset's confusion matrix for a Gradient Boost Classifier

Figure 6 illustrates the confusion matrix of Gradient Boost classification applied to the testing dataset. The x-axis represents the predicted values, while the y-axis corresponds to the actual values for the testing dataset of the Gradient Boost classifier. In the confusion matrix, the labels are abbreviated as follows: "v" for vegetation, "wb" for water body, "c" for cloud, "csv" for cloud's shadow on vegetation, "cst" for cloud's shadow on bare terrain or bare land, "css" for cloud's shadow on snow or ice, "ts" for terrain's shadow, "si" for snow and ice, "d" for debris, and "bl" for bare land or bare terrain.

The provided confusion matrix offers valuable information about the accuracy of the gradient boost classifier across different features. Notably, the accuracy of cloud's shadow on snow or ice and terrain's shadow is relatively lower compared to other features. While most of the geomorphological features achieve an accuracy above 90%, these two features exhibit accuracies of 80.6% and 87.3% respectively. On the other hand, features such as vegetation, water body, cloud, debris, bare terrain or bare land, and snow or ice demonstrate favourable accuracy levels for identification.

Comparison of Random Forest and Gradient Boost Classifier:

Upon comparing Figure 4 (confusion matrix of random forest classifier) and Figure 5 (confusion matrix of gradient boost classifier), it can be inferred that the gradient boost classifier exhibits higher accuracy in predicting cloud's shadow on terrain, cloud's shadow on snow, terrain's shadow, and debris compared to the random forest classifier. Additionally, the overall accuracy of the gradient boost classifier (98.5%) surpasses that of the random forest classifier (98%).

Discussion:

The study aimed to evaluate the applicability of random forest and gradient boost classifiers for the identification of various geomorphological features in the Himalayan region. The results obtained from the confusion matrices of both classifiers provided valuable insights into their performance and accuracy.

Analysing Figure 4, which represents the confusion matrix of the random forest classifier, it is evident that the accuracy of the classifier varied across different features. Notably, the accuracy was relatively lower for the shadow of cloud on bare terrain or land, shadow of cloud on snow or ice, and terrain's shadow. This lower accuracy can be attributed to the similarities in the features of these classes, leading to misclassifications. Similarly, the debris and bare terrain or bare land features showed some confusion, with a small percentage of instances being misclassified.

In Figure 5, the confusion matrix of the gradient boost classifier, it was observed that the classifier demonstrated improved accuracy compared to the random forest classifier, particularly in the prediction of cloud's shadow on terrain, cloud's shadow on snow, terrain's shadow, and debris. The accuracy for these features was comparatively higher, indicating the effectiveness of the gradient boost classifier in distinguishing these classes accurately.

Moreover, the overall accuracy of the gradient boost classifier was found to be 98.5%, which outperformed the random forest classifier with an accuracy of 98%. This indicates that the gradient boost classifier is better suited for the identification of various geomorphological features in the Himalayan region.

These findings suggest that both random forest and gradient boost classifiers can be applied for the identification of geomorphological features. However, the gradient boost classifier demonstrated higher accuracy and performed better in distinguishing challenging features such as shadows and debris. Therefore, it can

be considered as a more suitable choice for accurate feature identification in the Himalayan region.

It is important to note that the applicability and performance of these classifiers may vary depending on the specific dataset, feature extraction techniques, and parameter tuning. Further research and experimentation should be conducted to validate these findings and explore potential enhancements to improve the accuracy and efficiency of feature identification in the Himalayan region.

CONCLUSION

In conclusion, the applicability of supervised machine learning algorithms for the identification of various geomorphological features in the Himalayan region holds significant promise. The study demonstrates that these algorithms can effectively analyse and classify complex terrain characteristics, providing valuable insights into the region's geomorphology.

The results indicate that supervised machine learning algorithms, such as random forests and Gradient Boost exhibit a high level of accuracy in detecting and classifying geomorphological features. By leveraging labelled training datasets and utilizing robust feature engineering techniques, these algorithms showcase their capability to accurately identify features such as Vegetation, Water Body, Snow, Ice, Debris, Bare Land, Bare Terrain.

The advantages of utilizing supervised machine learning algorithms in this context are evident. They offer a systematic and automated approach to processing large volumes of geospatial data, enabling efficient feature extraction and classification. Additionally, these algorithms can handle various data types, including remote sensing imagery, digital elevation models, and other geospatial datasets.

The successful application of supervised machine learning algorithms in the identification of geomorphological features in the Himalayan region opens avenue for numerous practical applications. These include mapping and monitoring of landscape changes, hazard assessment, resource management, and land-use planning.

However, it is essential to acknowledge the challenges and limitations associated with these algorithms. The availability and quality of training datasets, as well as the selection of appropriate features and algorithms, significantly impact the accuracy and generalizability of the models. Addressing these challenges requires careful data collection, preprocessing, and validation processes.

Further research and advancements in supervised machine learning techniques can enhance the applicability and accuracy of feature identification in the Himalayan region. Exploring novel algorithms, integrating multi-source data, and refining feature extraction methods can contribute to more robust and reliable results.

In conclusion, the utilization of supervised machine learning algorithms for identifying geomorphological features in the Himalayan region offers great potential for improving our understanding of the region's landscape dynamics.

With continued research and technological advancements, these algorithms can serve as valuable tools for geomorphological analysis, environmental management, and sustainable development in the Himalayas.

REFERENCES

1. Qian, Y., Zhou, W., Yan, J., Li, W. and Han, L., 2014. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing*, *7*(1), pp.153-168.
2. Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.A. and Rahman, A., 2020. Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote Sensing*, *12*(7), p.1135.
3. Kulkarni, A.D. and Lowe, B., 2016. Random forest algorithm for land cover classification.
4. Thanh Noi, P. and Kappas, M., 2017. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, *18*(1), p.18.
5. Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R., 2006. Random forests for land cover classification. *Pattern recognition letters*, *27*(4), pp.294-300.
6. Alifu, H., Vuillaume, J.F., Johnson, B.A. and Hirabayashi, Y., 2020. Machine-learning classification of debris-covered glaciers using a combination of Sentinel-1/-2 (SAR/optical), Landsat 8 (thermal) and digital elevation data. *Geomorphology*, *369*, p.107365.
7. Zhang, T., Su, J., Xu, Z., Luo, Y. and Li, J., 2021. Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier. *Applied Sciences*, *11*(2), p.543.