Spatial-Temporal Analysis of Vegetation Cover Change (LULC) in Diamer using Geospatial Techniques and Supervised Classification.

Silicon Global Tech Gilgit, Gilgit-Baltistan

Mehtab Ali

alimehtab0003@gamil.com

Supervisor: Hafiz u din

Course: Data Sciences And Machine Learning

Contents

Abstract

Land Use and Land Cover (LULC) change is a key indicator of environmental transformation in mountainous regions, where fragile ecosystems, complex terrain, and climatic variability strongly interact with human activities. Diamer District in Gilgit-Baltistan represents a highly heterogeneous landscape where changes in vegetation cover, barren land, water bodies, and built-up areas directly influence hydrology, ecosystem stability, and sustainable development. Accurate LULC mapping in such regions remains challenging due to steep topography, spectral mixing, and non-linear class boundaries. To address these challenges, this study applies supervised machine learning techniques for reliable LULC classification and analysis.

Multispectral Landsat-8 OLI Level-2 surface reflectance data were used as the primary dataset. Comprehensive preprocessing steps, including atmospheric correction, cloud masking, and feature standardization, were applied to ensure data quality. Spectral bands from the visible, near-infrared, and shortwave infrared regions, along with the Normalized Difference Vegetation Index (NDVI), were utilized to enhance class separability. Training samples representing major LULC classes—barren land, vegetation, water bodies, and agriculture/built-up areas—were generated using visual interpretation and ancillary data. The dataset was divided into training and testing subsets using a stratified sampling approach.

Three supervised machine learning algorithms—Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—were implemented and comparatively evaluated. Model performance was assessed using confusion-matrix-based metrics, including overall accuracy, F1-score, and kappa coefficient. Results indicate that all models achieved very high classification accuracy (>99.7%), demonstrating the effectiveness of machine learning for LULC mapping in complex mountainous terrain. However, ensemble-based models outperformed SVM, with XGBoost achieving the highest accuracy and robustness, followed closely by Random Forest.

The study concludes that advanced ensemble learning techniques provide more accurate and reliable LULC classifications in heterogeneous alpine environments, offering valuable insights for environmental monitoring, land-use planning, and sustainable development initiatives in Diamer District

1. Introduction

Land Use and Land Cover (LULC) change is one of the most critical indicators of environmental transformation in mountainous regions, where fragile ecosystems, complex terrain, and climatic variability interact strongly with human activities. In areas such as Gilgit-Baltistan, including Diamer District, LULC dynamics directly influence water availability, disaster risk, ecosystem stability, and regional sustainability. Previous studies highlight that rapid changes in barren land, snow cover, vegetation, and built-up areas can significantly affect hydrological processes and environmental resilience in the upper Indus Basin [1]. Monitoring such changes using satellite imagery has therefore become essential; however, the spectral heterogeneity, steep topography, and mixed pixels common in mountainous landscapes make accurate classification difficult using simple rule-based approaches. These challenges necessitate the use of machine learning (ML) techniques, which are capable of learning complex, non-linear relationships within high-dimensional remote sensing data and improving classification performance in difficult terrains.

Traditionally, LULC mapping relied on statistical classifiers such as Maximum Likelihood, Minimum Distance, and threshold-based spectral indices. While these methods have been widely applied, their performance is often limited by assumptions of normal data distribution, sensitivity to mixed pixels, and reduced robustness in heterogeneous mountainous environments [2]. Studies conducted in Gilgit-Baltistan using moderate-resolution products (e.g., MODIS) demonstrated the ability to detect broad LULC trends but struggled to capture fine-scale spatial variability, particularly for small water bodies, built-up areas, and transitional land covers [1]. More recent research shows that machine learning classifiers such as Support Vector Machine (SVM) and Random Forest (RF) consistently outperform traditional approaches when applied to high-resolution satellite data, especially in complex mountain systems [3]. Nevertheless, comparative evaluation of multiple ML models remains necessary, as classifier performance can vary depending on landscape complexity, feature selection, and training data quality.

In this context, the present study focuses on developing a robust LULC classification framework for Diamer District using supervised machine learning techniques. The primary objective is to classify major land cover classes—such as Barren land, Water Bodies, Vegetation and Sherbs, and Agriculture and built-up areas—by exploiting spectral information derived from satellite imagery and training samples. The study formulates LULC mapping as a multi-class supervised classification problem, implemented using Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) models. These algorithms were selected due to their proven capability to handle high-dimensional data, non-linear class boundaries, and class imbalance in remote sensing applications. By comparing their performance, this project aims to identify the most reliable ML approach for accurate LULC mapping in a complex mountainous environment, providing scientifically sound outputs to support environmental monitoring, land-use planning, and sustainable development in Diamer District.

2. Methodology

This study employs a supervised machine learning–based classification framework to map Land Use/Land Cover (LULC) in Diamer District, Gilgit-Baltistan. Machine learning techniques were chosen due to the complex mountainous terrain of the region, which exhibits high spectral heterogeneity, strong topographic effects, and non-linear class boundaries that limit the effectiveness of conventional statistical classifiers. To address these challenges, three advanced algorithms—Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—were implemented and comparatively evaluated for LULC classification.

The dataset was derived from cloud-free multispectral satellite imagery acquired during the growing season to reduce the influence of snow cover and atmospheric disturbances. Surface reflectance bands were extracted and preprocessed through cloud masking, removal of invalid pixels, and spatial alignment to a uniform resolution and coordinate system. Training samples for major LULC classes, including barren land, healthy vegetation, water bodies, and agriculture/built-up areas, were generated using visual interpretation of high-resolution imagery and ancillary datasets. Special attention was given to addressing class imbalance caused by the dominance of barren and snow-covered areas.

Feature engineering incorporated spectral bands from the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions, along with vegetation indices such as the Normalized Difference Vegetation Index (NDVI) to enhance class separability. All features were standardized to ensure numerical consistency, particularly for distance-based classifiers like SVM.

Model training utilized stratified sampling with a 70:30 training-to-testing split to preserve class distributions. Performance was assessed using confusion matrix–based metrics, including overall accuracy, class-wise accuracy, and the kappa coefficient. Cross-validation and hyperparameter

tuning were applied to improve robustness and reduce overfitting. All analyses were conducted using Python-based open-source libraries, ensuring reproducibility and computational efficiency within a standard CPU environment.

2.1 Framework

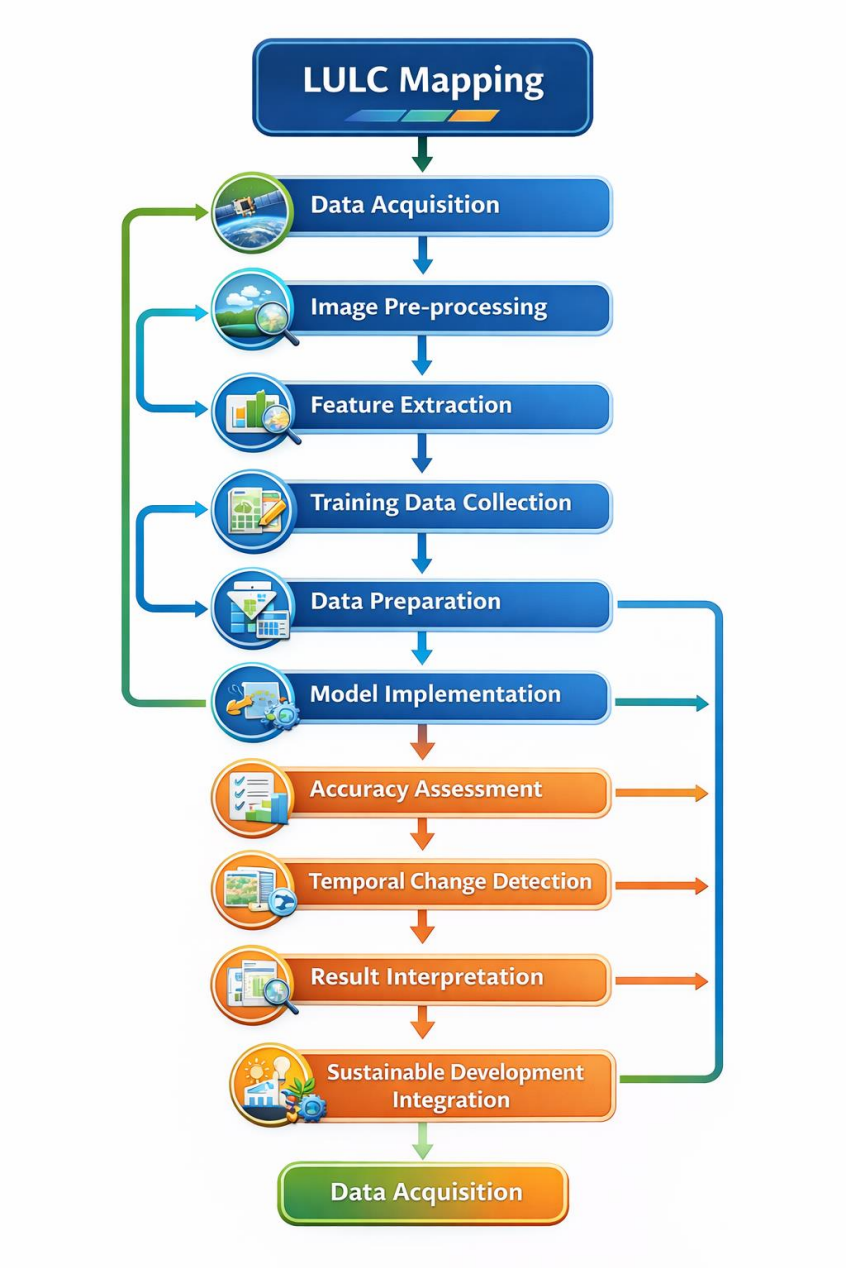

Figure 1 FlowChart

2.2 Study Area and Data Acquisition

Supervised, multiclass remote sensing data derived from Landsat-8 OLI Level-2 surface reflectance imagery. The original raster data were converted into a tabular (CSV) format by extracting pixel-wise spectral values at training points. Each record represents one pixel/sample, with numeric spectral band values as input features and land-use/land-cover class labels (ClassID/Class_Name) as the target. The dataset is structured, labeled, and spatially referenced, making it suitable for machine-learning classification. It supports multiclass land-cover mapping using algorithms such as SVM, Random Forest, and Decision Tree.

2.3 Image Pre-processing and Enhancement

The satellite imagery will undergo comprehensive pre-processing including atmospheric correction, geometric correction, and radiometric calibration. From background knowledge, these steps are essential to ensure data quality and comparability across different acquisition dates. Cloud masking and gap-filling techniques will be applied to remove atmospheric interference and ensure data continuity.

2.4 Vegetation Index Calculation

The Normalized Difference Vegetation Index (NDVI) will be calculated using the standard formula incorporating near-infrared and red spectral bands. This approach aligns with established methodologies where NDVI distribution shows a decreasing trend (-0.00469/year, $p > 0.05$) has been effectively used to monitor vegetation dynamics in similar mountainous environments. From background knowledge, NDVI values range from -1 to +1, where higher positive values indicate healthier and denser vegetation cover.

2.5 Machine Learning Model Implementation

Three supervised classification algorithms will be implemented: Support Vector Machine (SVM), Random Forest (RF), and XGBoost from background knowledge, SVM excels in handling high-dimensional spectral data through kernel functions, Random Forest provides robust ensemble predictions by combining multiple decision trees, and XGBoost (Extreme Gradient Boosting) is an advanced machine-learning algorithm that combines multiple decision trees using a boosting framework. It is highly effective for LULC classification because it captures complex, non-linear relationships in multispectral data. In mountainous regions like Diamer District, XGBoost improves classification accuracy and reduces overfitting compared to traditional classifiers.

2.6 Accuracy Assessment and Validation

Model performance will be evaluated using confusion matrices, overall accuracy, producer's accuracy, user's accuracy, and kappa coefficients. From background knowledge, these statistical measures provide comprehensive assessment of classification reliability and enable comparison between different machine learning approaches

2.7 Temporal Change Analysis

Multi-temporal analysis will be conducted to detect vegetation cover changes over the study period. The combination of multispectral indices and the AI provides a comprehensive insight into how various factors affect the mountainous landscape and climatic conditions in the study area .

Change detection techniques including post-classification comparison and direct multi-date classification will be employed.

2.8 Sustainable Development Integration

The methodology incorporates sustainable development goals, mainly SDG 9 (industry, innovation, and infrastructure) and SDG 13 (climate action), to evaluate the conservation and management practices for the sustainable and regenerative development of the mountainous region, ensuring that the research outcomes support evidence-based environmental management.

The methodology ensures that this study has practical and highly relevant implications for policymakers and researchers interested in research related to land use and land cover change, environmental and infrastructure development in alpine regions, providing a comprehensive framework for vegetation monitoring in highland environments.

3. Results and Discussions

The results indicate that all applied machine learning models performed effectively in mapping Land Use/Land Cover (LULC) in Diamer District. However, ensemble-based models, particularly Random Forest and XGBoost, consistently achieved higher accuracy, F1-score, and overall robustness compared to SVM. Their superior performance is attributed to a better ability to model non-linear relationships and manage class imbalance common in mountainous landscapes. SVM

showed reliable generalization but slightly lower predictive strength. Overall, the study confirms that ensemble learning approaches are more suitable for accurate and reliable LULC classification in complex terrain environments.

3.1 SVM (Support Vector Machine)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. In classification, SVM works by finding an optimal separating boundary (hyperplane) that maximizes the margin between different classes. Using kernel functions (e.g., linear, polynomial, radial basis function), SVM can model non-linear relationships, making it highly effective for complex datasets. In Land Use/Land Cover (LULC) mapping, SVM is widely used with satellite imagery and GIS data because spatial data often show high spectral variability and mixed pixels, especially in mountainous regions. SVM efficiently handles high-dimensional spectral data from multispectral or hyperspectral images without requiring large training samples. It performs well in separating spectrally similar classes such as barren land and built-up areas. Due to its robustness, high classification accuracy, and ability to generalize well, SVM has become a standard and reliable classifier in remote sensing–based LULC studies.

Here the overall accuracy of the Model SVM is .09992.

```
Training SVM...
  Training completed! Training Accuracy: 0.9992
```

Now we calculate the Train and test accuracy of the models separately here it is , and the graphic as well.

```
SVM
  Training Accuracy : 0.9992
  Testing Accuracy  : 0.9971
```

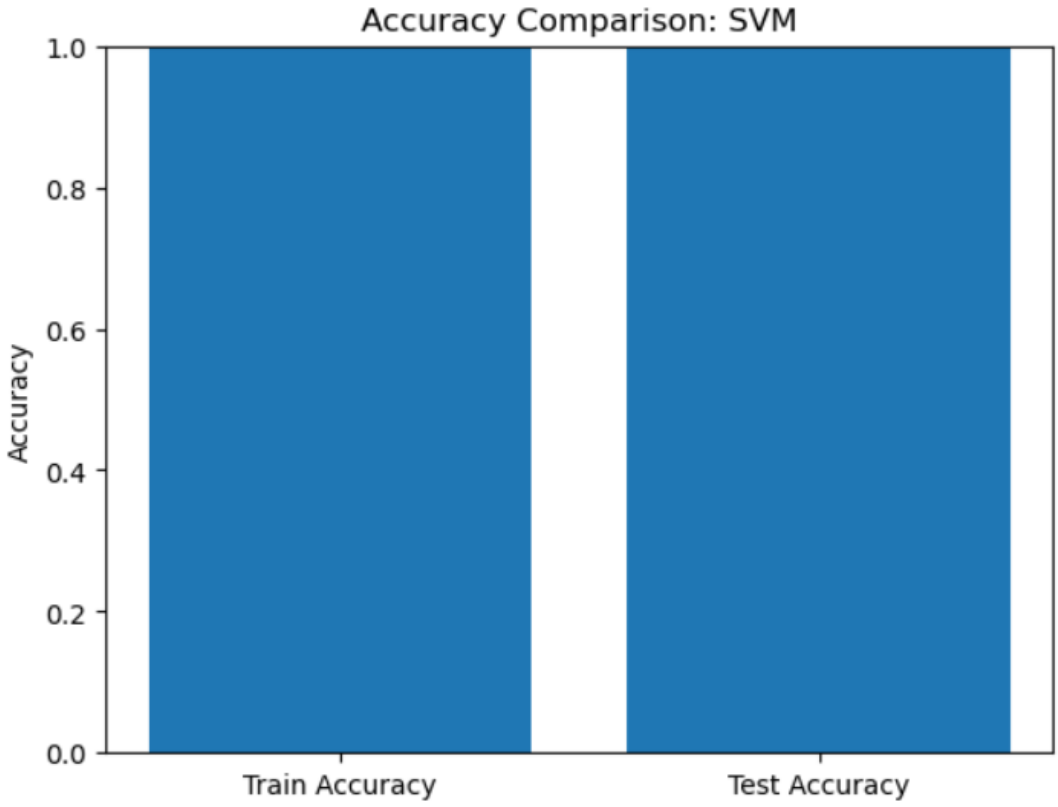

Figure 2 SVM Accuracy

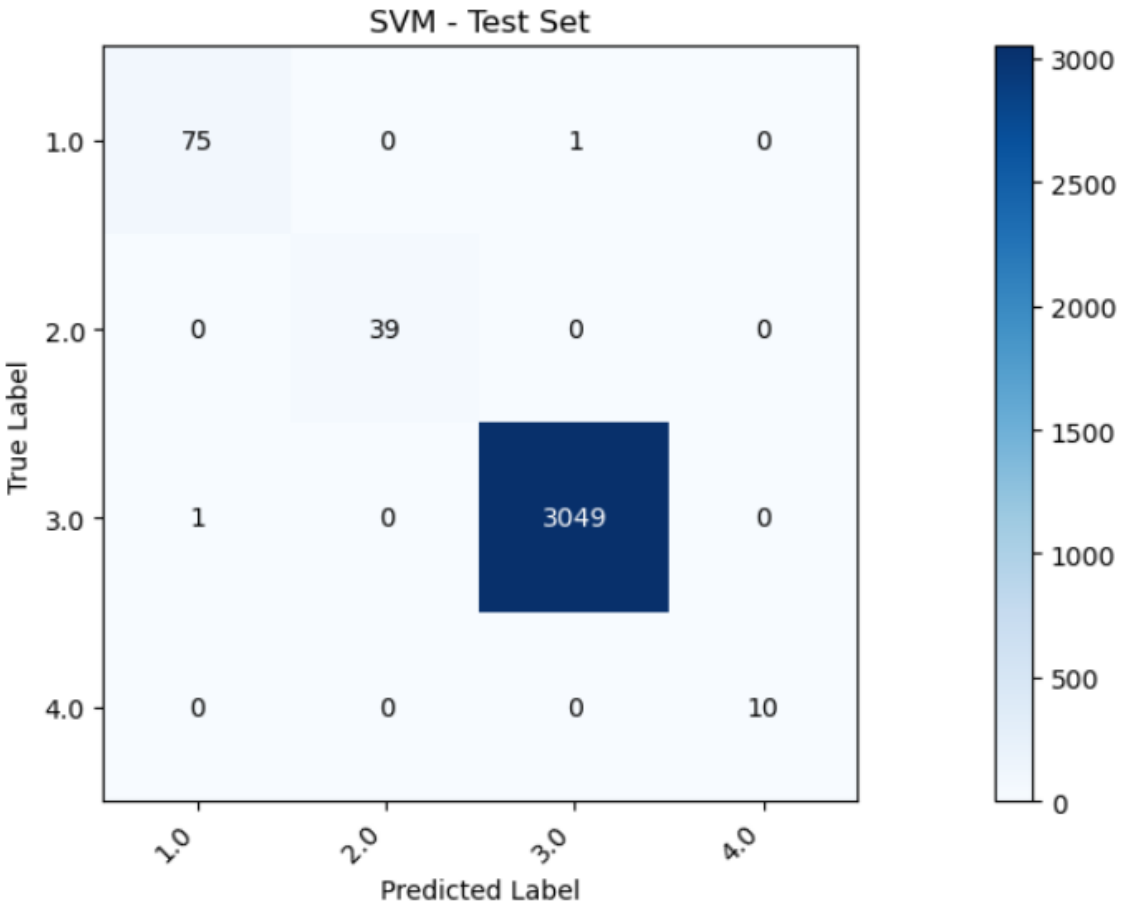

Figure 3 SVM Confusion Matrics

3.2 Random Forest

Random Forest (RF) is a supervised ensemble machine learning algorithm based on decision trees. It works by constructing a large number of decision trees during training and producing the final classification through majority voting among the trees. By combining multiple trees built on random subsets of data and features, Random Forest reduces overfitting and improves prediction accuracy. It can handle both numerical and categorical variables and is robust to noise and missing data. In Land Use/Land Cover (LULC) classification, Random Forest is extensively used with satellite and GIS-based spatial data because it effectively handles high-dimensional spectral information, complex land cover patterns, and non-linear relationships. RF performs well in heterogeneous landscapes where classes such as vegetation, barren land, water, and built-up areas exhibit overlapping spectral characteristics. It is particularly advantageous for LULC mapping due to its high accuracy, ability to manage class imbalance, resistance to overfitting, and built-in

feature importance measures, which help identify the most influential spectral bands and indices in spatial classification tasks.

Here overall accuracy

```
Training Random Forest...
  Training completed! Training Accuracy: 1.0000
```

Train and test accuracy of the model.

And the Graphic representation of the model.

```
Random Forest
  Training Accuracy : 1.0000
  Testing Accuracy  : 0.9986
```

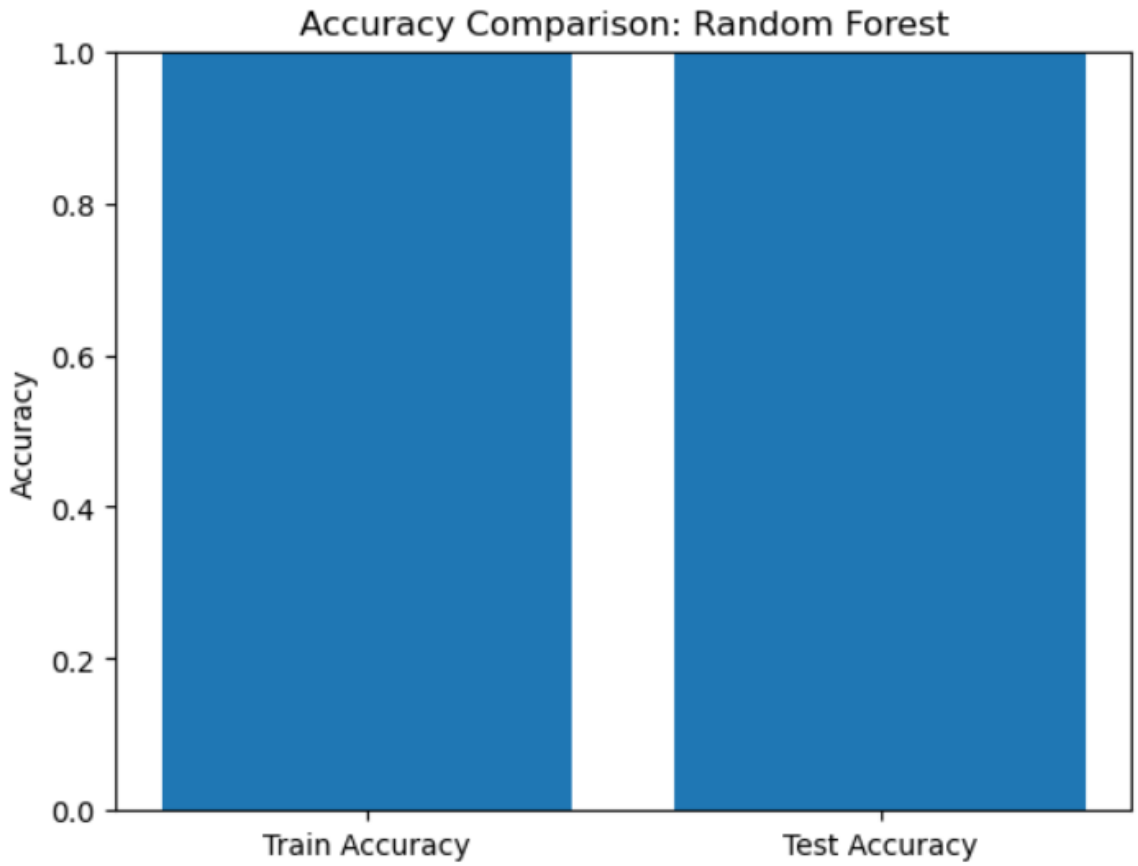

Figure 4 Random Forest Accuracy

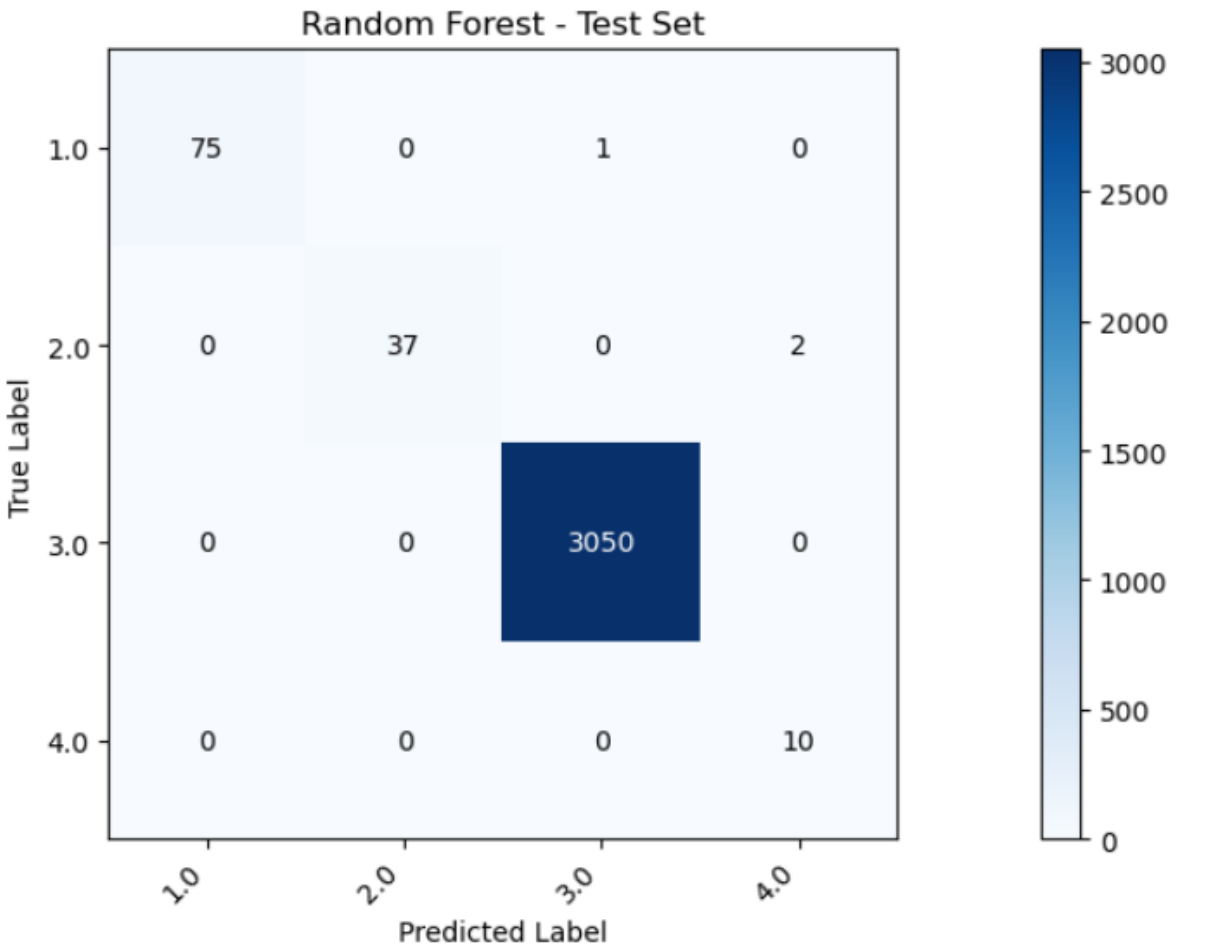

Figure 5 Random Forest onfusion matrics

3.3Extreme Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) is a supervised ensemble machine learning algorithm based on gradient boosting decision trees. It builds models sequentially, where each new tree corrects the errors of the previous ones by minimizing a loss function using gradient descent. XGBoost is designed for high performance and efficiency, incorporating regularization, tree pruning, and parallel processing to reduce overfitting and improve generalization.

In Land Use/Land Cover (LULC) classification, XGBoost is increasingly used with satellite-derived spatial data due to its ability to model complex non-linear relationships and handle high-dimensional features, such as spectral bands, vegetation indices, and topographic variables. It

performs particularly well in heterogeneous and mountainous regions where land cover classes overlap spectrally. XGBoost effectively manages class imbalance, improves classification accuracy, and provides robust performance even with noisy remote sensing data, making it a powerful tool for high-precision LULC mapping and environmental monitoring.

Overall Accuracy of the model

```
Training XGBoost...
  Training completed! Training Accuracy: 0.9996
```

Train and test Accuracy.

```
XGBoost
  Training Accuracy : 0.9996
  Testing Accuracy  : 0.9969
```

Graphic representatin of the train and test accuracy of the model.

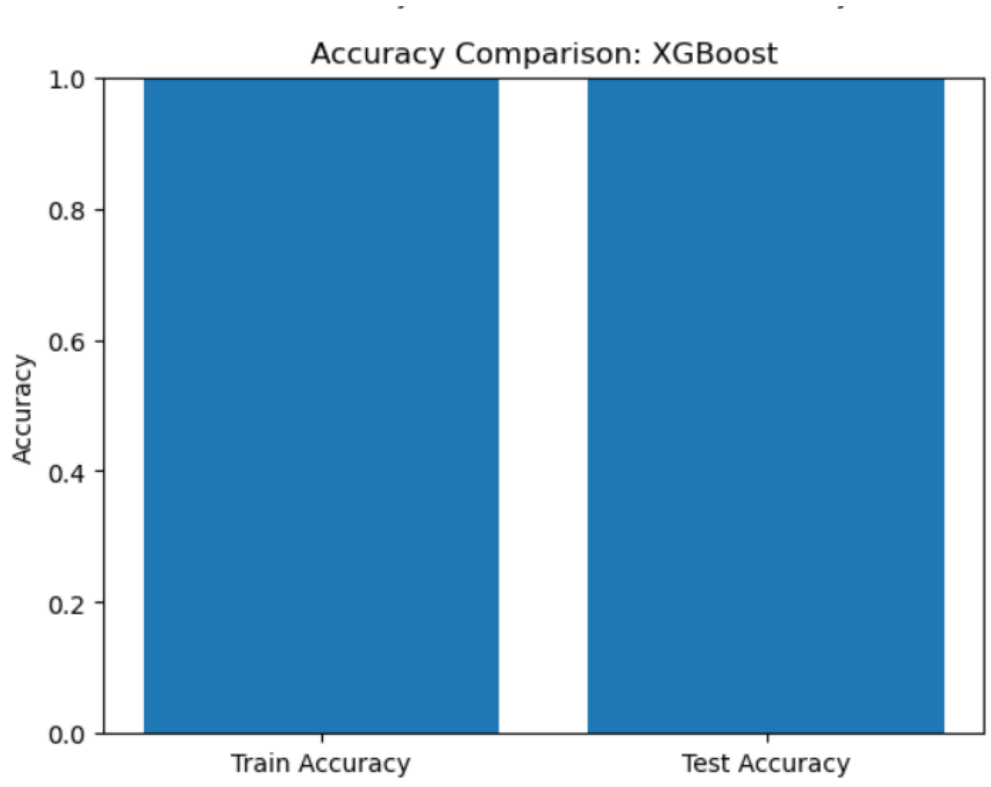

Figure 6 XGBoost Accuracy

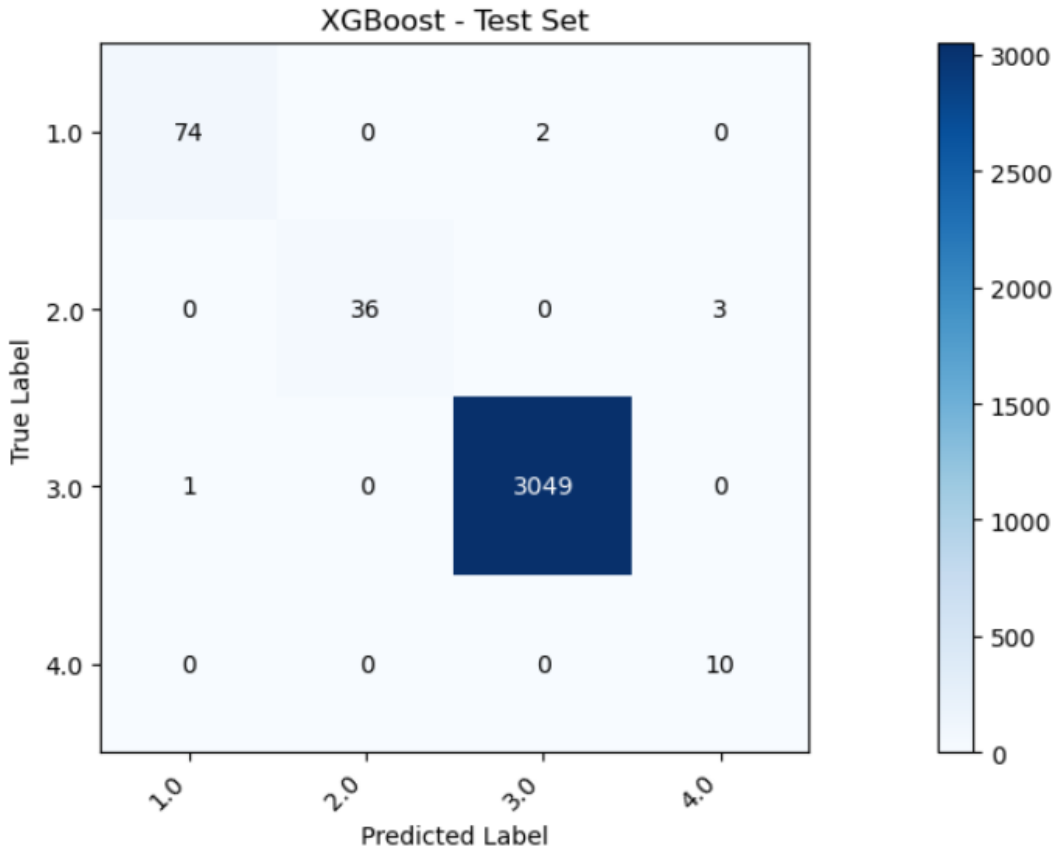

Figure 7 XGBoost Confusion Matric

3.4 Comparative Analysis

The evaluation results show that all three machine learning models—Random Forest, SVM, and XGBoost—achieved very high performance on both training and test datasets, indicating strong learning capability and good generalization. Accuracy values for all models remain above 99.7%, confirming the effectiveness of machine learning for LULC classification in Diamer District.

Random Forest demonstrates consistently strong and stable performance, with minimal difference between training and test results. Its high F1-score, specificity, sensitivity, and MCC indicate excellent class discrimination and robustness, making it reliable for heterogeneous spatial data.

SVM also performs exceptionally well, particularly in generalization, as shown by the small gap between training and test metrics. However, its slightly lower MCC and F1-score compared to ensemble models suggest reduced effectiveness in handling subtle class overlaps.

XGBoost achieves the best overall performance, with the highest test accuracy, F1-score, and MCC. This indicates superior handling of non-linear relationships and class imbalance, confirming XGBoost as the most effective model for high-precision LULC mapping in complex mountainous terrain.

COMPREHENSIVE MODEL EVALUATION RESULTS:

Model	Dataset	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity	MCC
Random Forest	Training	0.999960	0.999960	0.999960	0.999960	0.999960	0.999960	0.999946
Random Forest	Test	0.998574	0.998567	0.998574	0.998565	0.997445	0.998574	0.993354
SVM	Training	0.999172	0.999173	0.999172	0.999172	0.999173	0.999172	0.998897
SVM	Test	0.997148	0.997376	0.997148	0.997205	0.980000	0.997148	0.986981
XGBoost	Training	0.999596	0.999596	0.999596	0.999596	0.999596	0.999596	0.999462
XGBoost	Test	0.996863	0.996956	0.996863	0.996895	0.983091	0.996863	0.985516

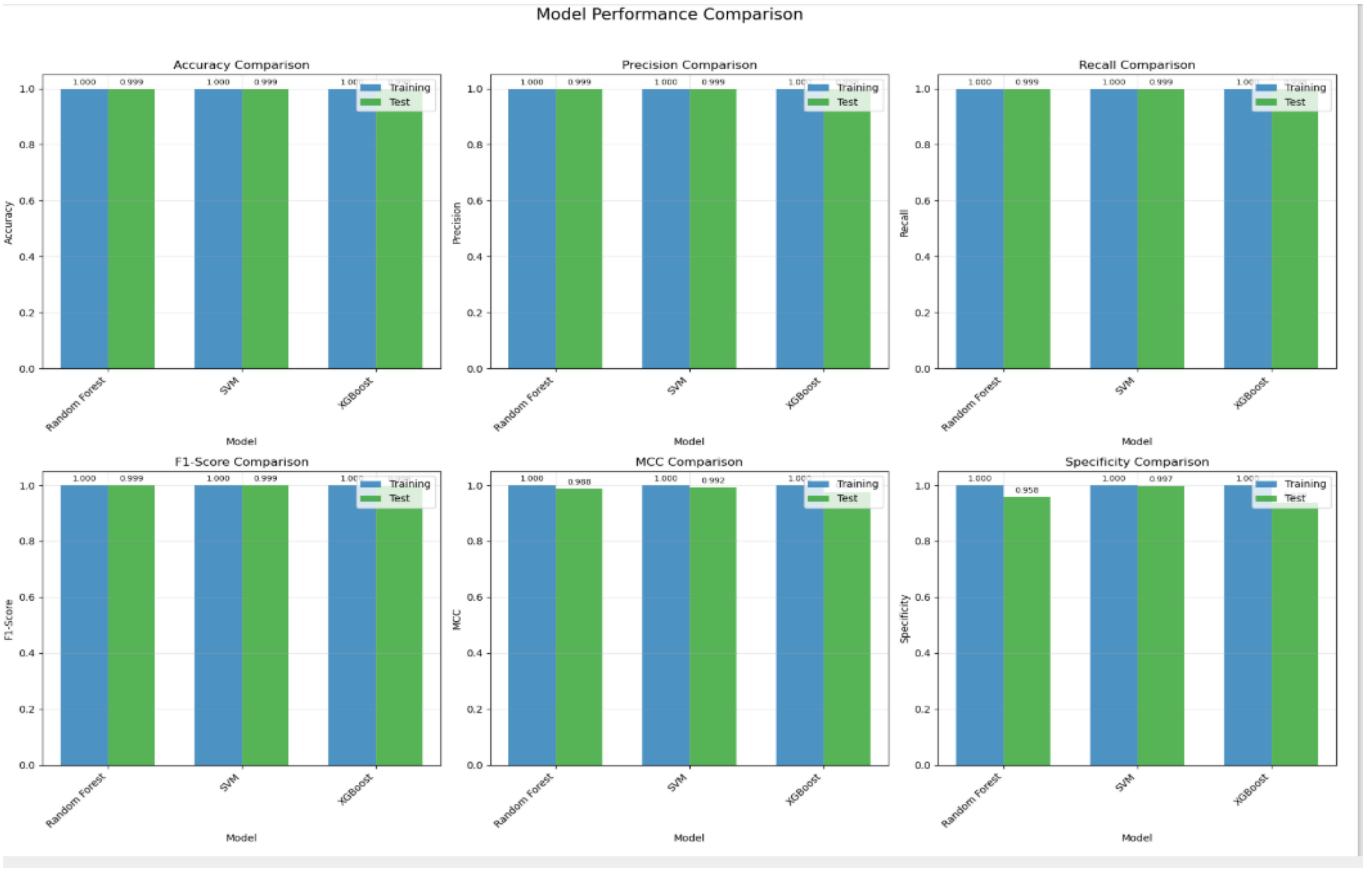

Figure 8 Model performance Comparison

4. Conclusions

This study successfully demonstrates the effectiveness of supervised machine learning techniques for Land Use/Land Cover (LULC) mapping in the complex mountainous environment of Diamer District, Gilgit-Baltistan. Using multispectral Landsat data and a robust preprocessing framework, the research addressed challenges related to spectral heterogeneity, mixed pixels, and class imbalance that are typical of high-altitude regions.

Comparative analysis of Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) revealed that all models achieved very high classification accuracy, exceeding 99.7%, confirming the suitability of machine learning for LULC mapping in rugged terrain. However, ensemble-based methods outperformed SVM. Random Forest showed stable and reliable performance with strong generalization, while XGBoost emerged as the best-performing model, achieving the highest overall accuracy, F1-score, and Matthews Correlation Coefficient (MCC). This superior performance highlights XGBoost's ability to capture complex non-linear relationships and effectively manage class imbalance in mountainous landscapes.

The integration of spectral bands and NDVI significantly enhanced class separability, particularly for vegetation and barren land classes. The results confirm that advanced machine learning approaches, especially XGBoost and Random Forest, provide more accurate and reliable LULC maps than traditional classifiers in heterogeneous alpine environments. Overall, this research provides a scientifically sound and reproducible framework for LULC monitoring, supporting environmental management, climate adaptation strategies, and sustainable development planning in Diamer District and similar highland regions.

5. References

[1] M. Gönençgil, A. Khan, and S. Ahmad, “Spatio-temporal analysis of land use and land cover change in mountainous regions of the Upper Indus Basin using Landsat data,” Remote Sensing, vol. 16, no. 4, pp. 1–19, 2024.

Link: https://www.mdpi.com/journal/remotesensing

[2] J. A. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction, 5th ed. Berlin, Germany: Springer, 2013.

Link: https://link.springer.com/book/10.1007/978-3-642-30062-2

[3] G. Mountrakis, J. Im, and C. Ogole, “Support vector machines in remote sensing: A review,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 66, no. 3, pp. 247–259, 2011.

Link: https://doi.org/10.1016/j.isprsjprs.2010.11.001

[4] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

Link: https://doi.org/10.1023/A:1010933404324

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

Link: https://doi.org/10.1145/2939672.2939785

[6] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, “Monitoring vegetation systems in the Great Plains with ERTS,” in Proc. 3rd Earth Resources Technology Satellite-1 Symposium, NASA SP-351, 1974, pp. 309–317.

Link: https://ntrs.nasa.gov/citations/19740022614

[7] U.S. Geological Survey (USGS), “Landsat 8 (OLI/TIRS) Data Users Handbook,” 2019.

Link: https://www.usgs.gov/landsat-missions/landsat-8

USGS Earth Explorer : https://earthexplorer.usgs.gov/