Evaluating Machine Learning Algorithms for Forest Cover Extraction in Kailali, Nepal

Sudarshan Kumar Gautam[1], Sanjeevan Shrestha[2], Subhadra Joshi[3] & Jeshan Pokharel[3]

sudarshangtm01@gmail.com, shr.sanjeevan@gmail.com, joshisub7@gmail.com, jeshanpokharel123@gmail.com

[1] Survey Department, [2] Ministry of Land Management Cooperatives and Poverty Alleviation,
[3] Kathmandu University,

KEYWORDS

Forest Cover Mapping, Machine Learning, Sentinel-2, K-fold Cross Validation, Performance Evaluation

ABSTRACT

Forest cover mapping plays a critical role in environmental monitoring, biodiversity conservation, and sustainable land-use planning, especially in ecologically diverse regions like Nepal. This study evaluates the performance of ten supervised machine learning classifiers for forest cover extraction in the Kailali District using Sentinel-2 satellite imagery. The classifiers assessed include Random Forest, Support Vector Classifier, Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Tree, Gaussian Naïve Bayes, AdaBoost, Quadratic Discriminant Analysis, and Gaussian Process Classifier. Feature engineering involved the derivation of 17 vegetation and water indices alongside key spectral bands, followed by correlation analysis to optimize input variables. Ground truth data were collected through field surveys and high-resolution imagery to ensure accurate model training and validation. Classifier performance was evaluated using k-fold cross-validation and standard metrics, including accuracy, precision, recall, and F1-score. Among the models, Random Forest and Gaussian Process achieved the highest classification accuracies of 91.37% and 91.31%, respectively. The study demonstrates the effectiveness of machine learning techniques in forest cover classification and provides valuable insights for enhancing remote sensing-based monitoring frameworks in support of sustainable forest management in Nepal.

1. INTRODUCTION

1.1 Background

Forest cover extraction is an essential process in remote sensing and Geographic Information Systems (GIS) that involves identifying and mapping forested areas using satellite imagery and aerial photographs. This process plays a crucial role in environmental monitoring, conservation efforts, and sustainable land management. In Nepal, where forests contribute significantly to biodiversity and ecological stability, forest cover mapping provides valuable insights for policymakers, researchers, and conservationists. It helps in identifying biodiversity hotspots, tracking deforestation trends, and supporting sustainable forest management initiatives (Singh & Kushwaha, 2008).

Forests offer essential ecosystem services such as carbon sequestration, water regulation, and soil conservation, playing a key role in climate change mitigation and disaster risk reduction (Moomaw et al., 2019). In Nepal, where many communities depend on forests for their livelihoods, mapping forest cover helps in planning responsible resource utilization while maintaining ecological balance (Pokharel et al., 2023). Additionally, forest cover mapping supports land-use planning by guiding decisions on afforestation, reforestation, and agricultural expansion while minimizing negative environmental impacts. It provides essential data for identifying deforested or degraded areas suitable for restoration and helps balance ecological conservation with human development needs (Muinonen et al., 2012). As a committed participant in international agreements like the UNFCCC and REDD+, Nepal actively utilizes forest cover data to meet its climate action goals and support sustainable development efforts. This data is crucial for monitoring emissions reductions, managing forest resources, and reporting on progress under its Nationally Determined Contributions (NDCs) (Maraseni et al., 2020). By continuously monitoring and updating forest cover information, Nepal can strengthen conservation strategies, reduce environmental risks, and promote long-term ecological and economic sustainability.

1.2 Significance of Machine Learning for Forest Cover Extraction

Assessing different machine learning algorithms for forest cover extraction is essential for improving classification accuracy and ensuring reliable forest mapping. Various algorithms possess unique strengths and weaknesses, making it necessary to evaluate multiple models to determine the most effective approach. Selecting appropriate machine learning techniques enhances the precision of forest cover classification, providing dependable data for environmental monitoring and decision-making (Maxwell et al., 2018).

Accurate forest cover maps contribute to sustainable resource management by helping governments and conservation organizations plan forestry activities, monitor deforestation, and identify reforestation opportunities. Since forests play a significant role in carbon sequestration and climate regulation, improved forest mapping enables better estimation of carbon storage, analysis of deforestation trends, and assessment of land-use impacts on greenhouse gas emissions (DeFries et al., 2007).

Machine learning-based forest cover extraction is also crucial for biodiversity conservation. Detailed maps help identify critical habitats, species distribution, and biodiversity hotspots, supporting the protection of endangered species. Additionally, forest cover information aids in disaster risk management, providing insights into the vulnerability of forests to wildfires, landslides, and floods.

Governments and environmental agencies rely on forest cover data to develop policies for land use, conservation, and sustainable development (Kissinger et al., 2012). Many countries utilize satellite-based monitoring systems, which can be enhanced through machine learning techniques to provide continuous and updated forest cover information (Pacheco-Pascagaza et al., 2022).

1.3 Objectives

The primary objectives of this study are to assess the effectiveness of different machine learning algorithms for forest cover extraction using remote sensing data in Nepal. This includes comparing the algorithms based on metrics such as accuracy, precision, recall, and F1 score to determine their performance. Additionally, the study aims to identify the most suitable algorithm for efficient and accurate forest cover mapping in Nepal, culminating in the creation of forest cover maps derived from various classification models.

2. METHODOLOGY

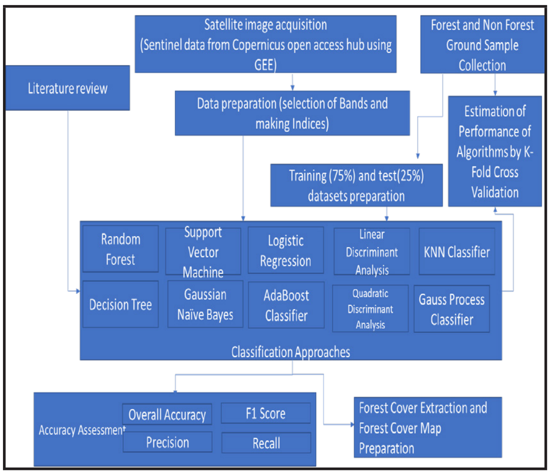


Figure 1:Overall Methodology of the Research

2.1 Study Area

The study was conducted in Kailali District, Nepal (Figure 2), covering an area of approximately 3,235 square kilometers. The district comprises diverse landscapes, from Terai plains to Siwalik foothills, making it an ideal study area for evaluating forest classification models.

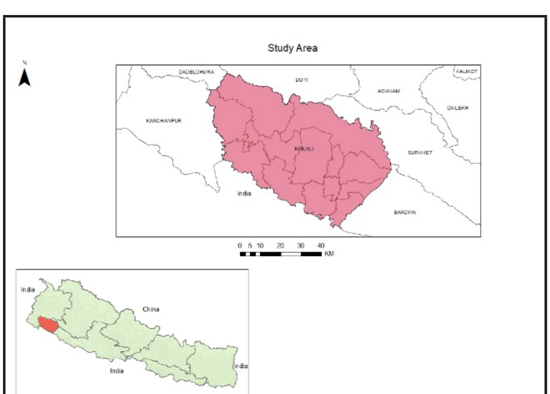


Figure 2: Study Area (Kailali District, Nepal) Map

2.2 Data Collection and Processing

2.2.1 Satellite Imagery

Sentinel-2 imagery was acquired from the Copernicus Open Access Hub. The four spectral bands (Blue Band (B2), Green Band (B3), Red Band (B4), and Near-Infrared Band (B8)) were used for classification.

2.2.2 Ground Truth Data Collection

The forest and non-forest point sample data utilized in this study forms the foundation for training and validating the machine learning algorithms employed to classify forested and non-forested regions. Ground truth data were gathered through field surveys using handheld GPS devices, along with the digitization of points from high-resolution Google Earth imagery.

2.2.3 Feature Engineering

Several spectral indices, as referenced from the Index Database, were derived from satellite bands, including vegetation indices such as NDVI, EVI2, SAVI, ARVI, OSAVI, and MSAVI2, as well as a water index (NDWI).

A total of 17 indices were computed using selected spectral bands (Red, Green, Blue, and NIR), as outlined in Table 1. When combined with the four original bands, this resulted in a total of 21 features, which were subsequently analyzed through correlation assessment. To minimize redundancy and enhance model performance, indices with high correlation values (>0.98) were excluded from further analysis (Figure 3).

Table 1:List of Indices Prepared using selected bands

Abbreviation	Full Name	Formula
NDVI	Normalized Difference Vegetation Index	$\frac{NIR - R}{NIR + R}$
EVI	Enhanced Vegetation Index	$\frac{2.5 \times (NIR - R)}{(NIR + 6 \times R - 7.5 \times B + 1)}$
SAVI	Soil-Adjusted Vegetation Index	$\frac{NIR - R}{NIR + R + L} \times (1 + L)$ Where L is a constant value (usually set to 0.5)
ARVI	Atmospherically Resistant Vegetation Index	$\frac{NIR - 2 \times R + B}{NIR + 2 \times R + B}$
OSAVI	Optimized Soil-Adjusted Vegetation Index	$\frac{NIR - R}{NIR + R + 0.16)} \times 1.16$
ARVI2	Modified Atmospherically Resistant Vegetation Index	$\frac{2 \times NIR + 1 - \sqrt{(2 \times NIR + 1)^2 - 8 \times (NIR - R)}}{2}$
EVI2	Enhanced Vegetation Index 2	$\frac{2.5 \times (NIR - R)}{(NIR + 2.4 \times R + 1)}$
EVI2.2	Enhanced Vegetation Index 2.2	$G_f \times \frac{(2.5 \times (NIR - R)}{(NIR + 2.4 \times R + 1)}$ Where G_f is a gain factor (Usually set to 2.5)
MCARI2	Modified Chlorophyll Absorption Ratio Index 2	$\frac{1.5 \times (2.5 \times (NIR - R) - 1.3 \times (NIR - G))}{\sqrt{2 \times NIR + 1 + 6 \times G - 7.5 \times R}}$
BNDVI	Blue Normalized Difference Vegetation Index	$\frac{NIR - B}{NIR + B}$
GNDVI	Green Normalized Difference Vegetation Index	$\frac{NIR - G}{NIR + G}$
GOSAVI	Green Optimized Soil-Adjusted Vegetation Index	$\frac{NIR - G}{NIR + G + 0.16} \times 1.16$
NDWI	Normalized Difference Water Index	$\frac{G - NIR}{G + NIR}$
MTVI2	Modified Triangular Vegetation Index 2	$\frac{1.5 \times (1.2 \times (NIR - G) - 2.5 \times (R - G)}{\sqrt{(2 \times NIR + 1)^2 - \left(6 \times NIR - 5 \times \sqrt{R}\right) - 0.5}}$
CIgreen	Green Chlorophyll Index	$\frac{NIR}{G} - 1$
GARVI	Green Atmospherically Resistant Vegetation Index	$\frac{NIR - (G - B)}{NIR + (G - B)}$
MSAVI2	Modified Soil-Adjusted Vegetation Index 2	$\frac{(2 \times NIR + 1 - \sqrt{(2 \times NIR + 1)^2 - 8 \times (NIR + R)}}{2}$

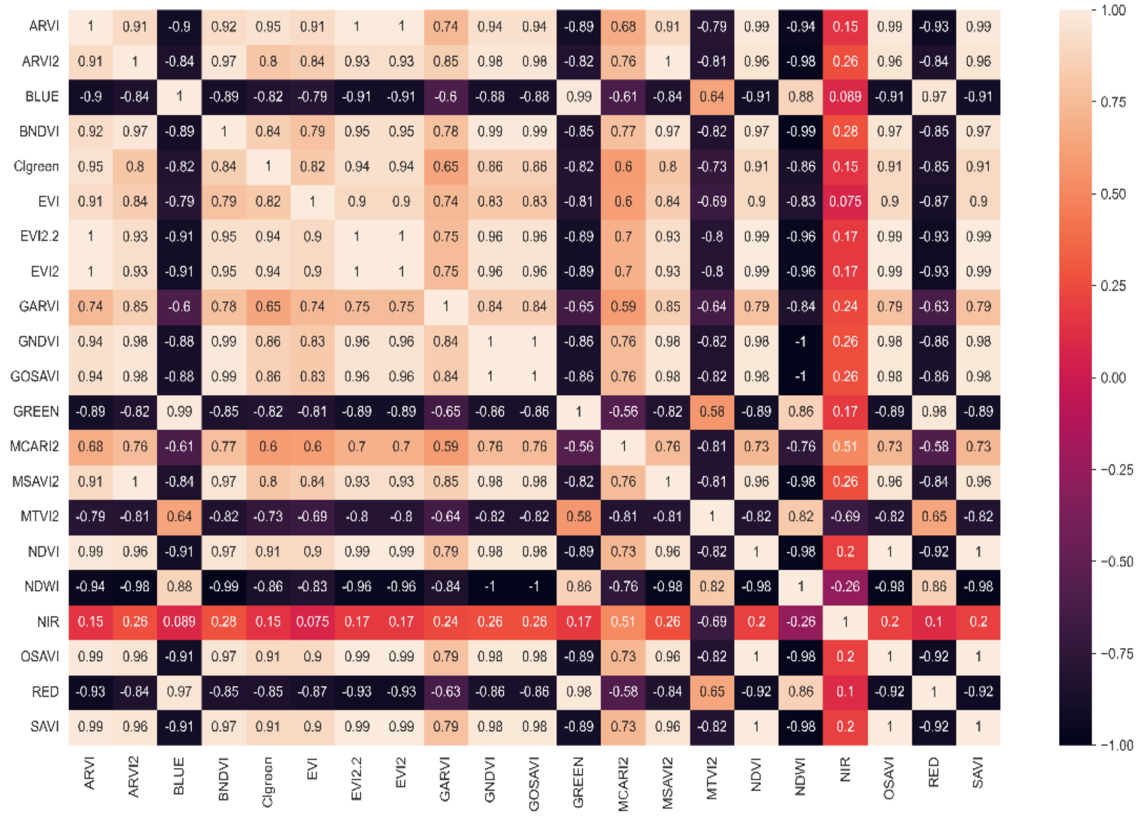
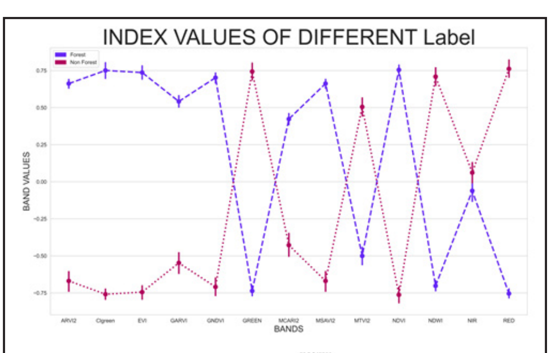
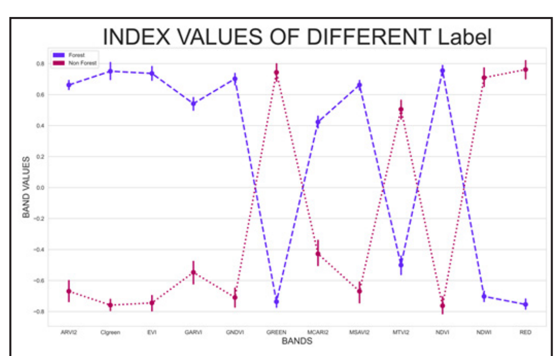


Figure 3 : Correlation Analysis between Different Vegetation Indices

The selected spectral bands and indices (ARVI2, CIgreen, EVI, GARVI, GNDVI, GREEN, MCARI2, MSAVI2, MTVI2, NDVI, NDWI, NIR, and RED) were further examined through a line plot (Figure 4(a)), illustrating their variation across forest and non-forest classes. The plot shows that the Near-Infrared (NIR) band exhibits minimal differentiation between the two classes, suggesting limited discriminatory capability in this specific context. As a result, NIR was excluded from further analysis to streamline the process without compromising classification effectiveness.



(a)



(b)

Figure 4 : Line plot including NIR (a) and Excluding NIR (b)

2.3 Machine Learning Models

Ten supervised machine learning classifiers (as presented in Table 2) were initially assessed using 10-fold cross-validation to identify any models demonstrating weak performance that might warrant early exclusion. Among them, Random Forest and Gaussian Process Classifier achieved the highest average accuracies—91.37% and 91.31%, respectively—with low standard deviations, suggesting strong and consistent performance across folds. In contrast, the Decision Tree classifier recorded the lowest accuracy (86.59%) and showed relatively higher variability. Nevertheless, all classifiers demonstrated satisfactory performance and were retained for subsequent analysis.

Table 2 : K-Fold Cross Validation Accuracy of Machine Learning Models

Classifier	Accuracy (%)	Deviation (%)
Random Forest	91.37	2.88
Gaussian Process	91.31	2.55
Support Vector Machine	91.19	2.53
Logistic Regression	90.68	3.26
Linear Discriminant Analysis	90.62	2.69
K-Nearest Neighbors	90.68	3.28
Decision Tree	86.59	3.16
Gaussian Naïve Bayes	90.62	2.49
AdaBoost	89.92	2.82
Quadratic Discriminant Analysis	88.28	2.51

Final evaluations were performed using a standardized train-test split (75% training, 25% testing) with a fixed random state to ensure reproducibility and fair comparison across models.

3. RESULTS AND DISCUSSION

3.1 Performance of Classifiers

The performance of ten supervised machine learning classifiers was evaluated using precision, recall, F1-score, and overall accuracy. Logistic Regression and AdaBoost achieved the highest accuracy of 0.8967, followed closely by Support Vector Classifier, K-Nearest Neighbors, and Gaussian Process Classifier with accuracies around 0.89. Quadratic Discriminant Analysis had the lowest performance with an accuracy of 0.7884. The comparison highlights the strengths and weaknesses of each model for forest cover classification tasks in the study area.

Table 3 : Accuracy, Precision, Recall and F1-Score of different ML Algorithms

Classifier	Accuracy	Forest			Non-Forest		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	89.17	87	92	89	92	87	89
SVC	89.42	87	92	89	92	87	89
LR	89.67	86	93	90	93	86	90
LDA	89.17	87	92	89	92	87	89
K-NN	89.42	87	92	89	92	87	90
Decision Tree	85.64	85	85	85	86	86	86
Gaussian Process	89.17	86	93	89	93	86	89
Gaussian NB	88.92	85	93	89	93	85	89
AdaBoost	89.67	86	94	90	94	85	90
QDA	78.84	80	75	77	78	82	80

3.2 Forest Cover Maps

Forest cover maps were produced by applying each of the aforementioned classifiers to the dataset, resulting in binary classifications of forest and non-forest areas. These classifications were subsequently used to generate thematic maps, providing a visual representation of the spatial distribution of forest and non-forest regions. An example of the resulting maps is presented in Figure 5 and 6.

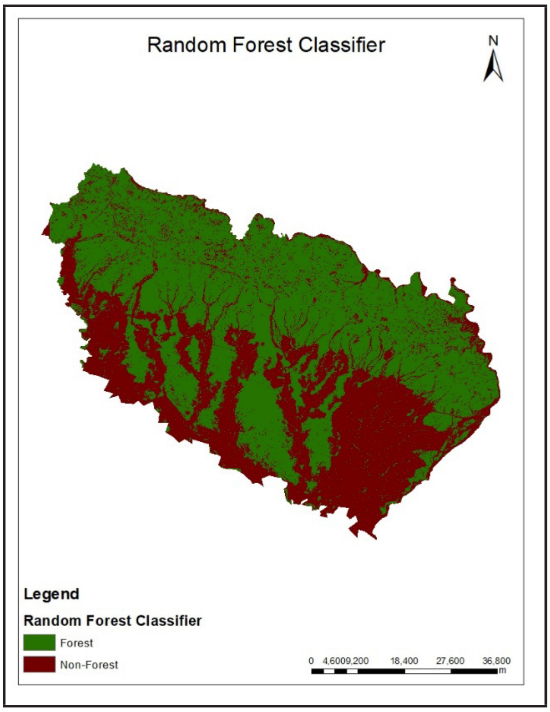


Figure 5: Forest and Non-Forest Regions of Study Area by Random Forest Classifier

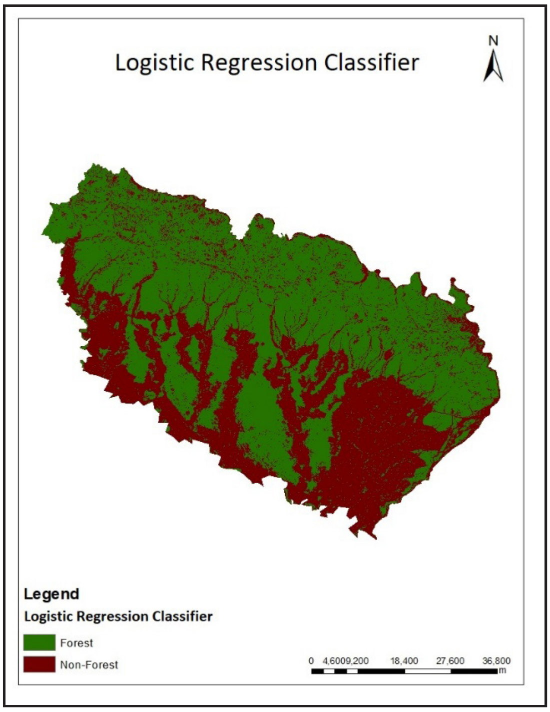


Figure 6: Forest and Non-Forest Regions of Study Area by Logistic Regression Classifier

4. CONCLUSION AND RECOMMENDATIONS

4.1 Conclusion

This study investigated the performance of ten supervised machine learning classifiers for forest cover extraction in the context of Nepal's diverse landscape. Through rigorous evaluation using 10-fold cross-validation, the Random Forest classifier emerged as the top performer, demonstrating both high accuracy and consistency across folds. Logistic Regression, Gaussian Process, and AdaBoost classifiers also showed strong and balanced results in terms of precision and recall for both forest and non-forest classes.

On the other hand, classifiers such as Decision Tree and Quadratic Discriminant Analysis recorded relatively lower performance, which may be attributed to their limitations in modeling complex, nonlinear relationships within the data. While the Gaussian Naïve Bayes classifier achieved moderate accuracy, its reduced precision for non-forest areas indicates a potential risk of false positives in real-world applications.

4.2 Recommendations

Given their reliable performance, Logistic Regression, Gaussian Process, and AdaBoost classifiers are recommended for practical forest cover classification tasks in the study region. These models strike a favorable balance between interpretability and predictive capability, making them suitable for use in operational mapping and decision support systems.

Future research could benefit from exploring advanced ensemble techniques or hybrid approaches that integrate multiple models to further enhance classification accuracy. In addition, periodic retraining of models with updated datasets is encouraged to reflect changes in land cover and ensure continued relevance over time.

Finally, the study underscores the importance of hyperparameter optimization. Notably, fine-tuning the Random Forest classifier led to a measurable improvement in accuracy (increasing to 92%), suggesting that careful model calibration can significantly boost performance.

REFERENCES

DeFries, R., Achard, F., Brown, S., Herold, M., Murdiyarso, D., Schlamadinger, B., & de Souza Jr, C. (2007). Earth observations for estimating greenhouse gas emissions from deforestation in developing countries. *Environmental Science & Policy*, *10*(4), 385–394.

Kissinger, G., Herold, M., & Sy, V. De. (2012). *Drivers of deforestation and forest degradation: a synthesis report for REDD+ policymakers.*

Maraseni, T. N., Poudyal, B. H., Rana, E., Khanal, S. C., Ghimire, P. L., & Subedi, B. P. (2020). Mapping national REDD+ initiatives in the Asia-Pacific region. *Journal of Environmental Management*, *269*, 110763.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817.

Moomaw, W. R., Masino, S. A., & Faison, E. K. (2019). Intact forests in the United States: Proforestation mitigates climate change and serves the greatest good. *Frontiers in Forests and Global Change*, *2*, 449206.

Muinonen, E., Parikka, H., Pokharel, Y. P., Shrestha, S. M., & Eerikäinen, K. (2012). Utilizing a multi-source forest inventory technique, MODIS data and Landsat TM images in the production of forest cover and volume maps for the Terai Physiographic Zone in Nepal. *Remote Sensing*, *4*(12), 3920–3947.

Pacheco-Pascagaza, A. M., Gou, Y., Louis, V., Roberts, J. F., Rodríguez-Veiga, P., da Conceição Bispo, P., Espírito-Santo, F. D. B., Robb, C., Upton, C., & Galindo, G. (2022). Near real-time change detection system using Sentinel-2 and machine learning: A test for Mexican and Colombian forests. *Remote Sensing*, *14*(3), 707.

Pokharel, B., Chotikarn, P., & Gywali, S. (2023). Changing the perception of forest value and attitude toward management in the conservation area in Nepal and sustainable forest energy: A Review. *Asian Journal of Conservation Biology*, *12*(2), 239–247.

Singh, J. S., & Kushwaha, S. P. S. (2008). Forest biodiversity and its conservation in India. *International Forestry Review*, *10*(2), 292–304. https://doi.org/10.1505/ifor.10.2.292

Author's Information

Name	:	Sudarshan Kumar Gautam
Academic Qualification	:	Masters of Engineering in Geoinformatics (Kathmandu University)
Organization	:	Survey Department
Current Designation	:	Survey Officer
Work Experience	:	12 years