



Rule-Based Vegetation Canopy Cover Classification Using Remote Sensing and Machine Learning Techniques in Marathwada Region, Maharashtra, India

¹Vikas Ghodke, ²Balaji Yadav, ³Sidheshwar Raut, ⁴Sayyad Shafiyoddin, ⁵Sandipan Sawant

¹Research Scholer, ²Research Scholer, ³Research Scholer, ⁴Ass. Professor, ⁵Professor

¹Name of Department of 1st Author,

^{1,2,3,5} Department of Physics, Shri. Chatrapati Shivaji College, Omerga, MS, India

²Microwave & Imaging Spectroscopy Laboratory, Milliya College, Beed, MS, India

Abstract: Monitoring vegetation canopy cover is vital for ecological balance, biodiversity preservation, and sustainable land management in the Marathwada region of Maharashtra, India, where agricultural growth, deforestation, and climate variability have driven significant environmental shifts. This study employs MODIS satellite imagery (2013–2023) to assess vegetation dynamics through integrated remote sensing and machine learning approaches. Pre-processing involved Dark Object Subtraction (DOS) for atmospheric correction, pixel normalization, and geometric alignment to ensure data accuracy. The Normalized Difference Vegetation Index (NDVI) served as the primary metric for vegetation density, with classification performed using supervised methods (Random Forest, XGBoost, Maximum Likelihood Classification) and unsupervised techniques (PCA + K-Means clustering). Random Forest achieved perfect classification accuracy (100%), followed closely by XGBoost (99.29%), surpassing Maximum Likelihood Classification (98.31%) and K-Means clustering. Temporal analysis revealed canopy cover fluctuations, with losses linked to urbanization and deforestation, alongside localized regrowth. PCA-based feature reduction preserved 94% of data variance, enhancing classification efficiency. The study achieved over 85% overall accuracy, highlighting the efficacy of ensemble classifiers (Random Forest and XGBoost) for precise land cover mapping and the utility of PCA + K-Means for initial vegetation analysis. These insights support environmental monitoring, climate impact evaluation, and sustainable land-use strategies in ecologically critical regions.

Index Terms - NDVI, Vegetation Canopy Cover (VCC), Remote Sensing, Maximum Likelihood Classification, MODIS Imagery, PCA + K-Means

I. INTRODUCTION

Vegetation canopy cover represents a critical ecological parameter that provides fundamental insights into ecosystem dynamics, climate regulation, and biodiversity conservation [1]. The spatial and temporal variations in vegetation cover significantly impact global environmental processes, including carbon sequestration, hydrological cycles, and climate change mitigation [2, 3]. Remote sensing (RS) technologies have emerged as powerful tools for monitoring and analyzing vegetation dynamics across large geographical scales, offering comprehensive and cost-effective approaches to landscape assessment [4-6].

In recent decades, the integration of satellite-based RS and advanced machine learning (ML) techniques has revolutionized vegetation monitoring methodologies [7]. Normalized Difference Vegetation Index (NDVI), a widely used spectral vegetation indicator, enables researchers to quantify vegetation density and health by leveraging the differential reflectance characteristics of vegetation in red and near-infrared spectral bands [8, 9]. Gandhi, G. Meera, et al.(2015) highlighted the significance of NDVI in assessing vegetation cover dynamics, particularly in regions experiencing complex land-use transformations [10, 11].

The Marathwada region in Maharashtra, India, presents a unique ecological landscape characterized by significant agricultural and environmental transitions. Previous studies have demonstrated substantial land-use changes in this region, underscoring the importance of comprehensive vegetation cover monitoring [12, 13]. However, existing classification methodologies face challenges related to spectral variability, mixed land cover types, and the requirement for accurate ground truth data [14].

This research aims to address these challenges by:

- Developing a robust vegetation canopy cover classification methodology
- Comparing unsupervised (K-Means) and supervised (Maximum Likelihood Classification) approaches
- Evaluating the effectiveness of NDVI-based classification techniques
- Contributing to advanced vegetation monitoring strategies in the Marathwada region

The study leverages a decade-long MODIS satellite imagery dataset (2013-2023), employing sophisticated preprocessing techniques and machine learning algorithms to extract precise vegetation cover information. By integrating Dark Object Subtraction (DOS), pixel normalization, and geometric correction, we enhance the accuracy and reliability of our classification process [15, 16].

Google Earth Engine (GEE) API for Python

The Google Earth Engine (GEE) Python API provides a strong interface for retrieving and examining satellite imagery and geographical data within a cloud-based system. The API, designed for researchers, scholars, and developers, enables comprehensive geospatial analysis, supporting applications such as remote sensing, environmental oversight, and resource management[17]. Users can effortlessly analyze, visualize, and obtain insights from a multi-petabyte dataset of satellite images. The Python API seamlessly works with major libraries like NumPy, Pandas, and TensorFlow, enabling advanced machine learning and data analysis tasks[18]. GEE remains free of charge for academic and research purposes, making it an essential tool for extensive environmental research and decision-making.

II. RELATED STUDIES

The classification of vegetation canopy cover by remote sensing and machine learning approaches is an advancing domain that utilizes sophisticated technologies to enhance precision and efficacy in environmental monitoring. Remote sensing offers a non-intrusive method for data collection across extensive regions, whereas machine learning methods improve the accuracy of canopy cover assessments. This integration is essential for applications spanning agriculture, urban planning, and forest management. The subsequent sections delineate diverse methodologies and conclusions derived from recent research.

2.1 Remote Sensing Techniques

High-resolution hyperspectral and multispectral images obtained from UAVs and gantry-based scanners is utilized to assess canopy cover. These methods employ the distinctive spectral patterns of vegetation in the visible and near-infrared wavelengths to differentiate vegetation from other components [19]. PlanetScope imagery, with its multispectral bands, is utilized to calculate vegetation indices for canopy cover modelling, illustrating the efficacy of satellite data in forest management [20].

2.2 Machine Learning Approaches

Rule-based approaches, Support Vector Machines (SVM), and Random Forest (RF) are frequently employed for canopy cover categorization. The rule-based approach, albeit responsive to mixed pixels, provides quicker processing speeds than SVM and RF, rendering it appropriate for high-throughput applications [21]. Deep learning methods, such as U-Net CNN, provide enhanced accuracy in analysing urban tree canopy coverage, albeit necessitating extended training duration [22]. The Light Gradient Boosting Machine has been recognized as the optimal model for predicting canopy cover in young boreal

forests, surpassing other regression models in fit statistics [20]. Whereas this research underscores the efficacy of remote sensing and machine learning in canopy cover assessment, problems persist. The susceptibility of specific approaches to mixed pixels and the requirement for substantial training data for deep learning models are significant constraints. The amalgamation of diverse data sources, including Landsat and Sentinel-2A, can augment temporal and spatial analysis, providing a more thorough comprehension of vegetation dynamics [23]. Table 1 provides a comparative review of multiple research articles that examine canopy cover assessment through remote sensing and machine learning methodologies. The research utilises many strategies, encompassing conventional machine learning classifiers like Support Vector Machines (SVM) and Random Forest (RF), in addition to sophisticated deep learning techniques such as U-Net CNN.

Table 1: Summary of Canopy Cover Assessment Studies

Studied by	Methods Used	Results	Dataset
Maimaitijiang, M., et al 2021 [21]	Rule-based approach utilizing spectral patterns in the visible and near-infrared regions; evaluated against Support Vector Machine and Random Forest algorithms.	The rule-based technique demonstrated accuracy comparable to that of SVM and RF. Enhanced identification of diminutive leaves; Accelerated processing, although susceptible to mixed pixels and darkened regions.	High-resolution multispectral and hyperspectral imagery from agricultural fields in Maricopa (AZ), Columbia (MO), and Urbana (IL) collected between 2017 and 2020 using UAV and Gantry platforms.
Morgan, G.R., et al [22]	U-Net CNN, SVM, and RF classifiers for urban tree canopy assessment	U-Net CNN exhibited the highest accuracy (89.8%-91.4%) but necessitated extended training duration; urban tree loss correlated with rising temperatures.	Aerial photography from 2012 over two southern U.S. cities; Recent imagery utilized to assess a decade of canopy change; Landsat 9 thermal data employed for urban heat mapping.
Gyawali, A., et al [20]	Linear Regression, Elastic Net, SVM, RF, Extreme Gradient Boosting, LightGBM.	LightGBM exhibited superior performance (R^2 : 0.69 for training, 0.64 for testing; RMSE: 0.16; MAE: 0.12)	PlanetScope multispectral imaging (green, blue, red, NIR); 43 vegetation indices calculated
Rohith, K., et al. [24]	Machine learning and image processing for vegetation classification; Sentinel-2A	Enhanced categorisation precision; Efficient monitoring of vegetative alterations throughout time.	Landsat multispectral images; Sentinel-2A data for temporal analysis
Smith, O. and H. Cho [25]	CanoClass system employing RF and Extra Trees algorithms (scikit-learn); ARVI index for atmospheric adjustment	CanoClass exhibited a speed increase of 283% to 464% compared to Feature Analyst, with an accuracy ranging from 87.56% to 87.62%.	NAIP four-band (RGB + NIR) images from 2015, comprising 731 GB of data across 3913 quarter quad tiles. Compatible with Landsat 8 and Sentinel-2 satellites
Nasiri, Vahid, et al.[26]	RF, SVM, Elastic Net (ENET), and Extreme Gradient Boosting (XGBoost).	RF had superior performance (R^2 = 0.67, RMSE = 18.87%, MAE = 15.35%), but Elastic Net shown inferior performance (R^2 = 0.63, RMSE = 20.04%, MAE = 16.44%).	Sentinel-2 multispectral bands, UAV-derived canopy data, vegetation indices (NDVI, NDRE, NDI45).

Saini, Rashmi[27]	RF, AdaBoost	RF attained an accuracy of 91.23% (kappa = 0.89), whereas AdaBoost earned an accuracy of 88.59% (kappa = 0.86).
-------------------	--------------	---

III. METHODOLOGY

3.1 Study Area

The study focuses on the Marathwada region in Maharashtra, India. This region comprises eight districts: Aurangabad, Jalna, Bid (Beed), Parbhani, Latur, Nanded, Osmanabad, and Hingoli as showing figure 1. Marathwada has a semi-arid environment, characterized by undulating terrain and varied soil types, hence designating it as a significant agricultural area[28]. The region has monsoon rainfall from June to September, with an annual average ranging from 600 to 900 mm, depending on the district. Summers often exhibit high temperatures ranging from 40 to 45°C, and winters are characterized by moderate conditions, averaging from 10 to 15°C[29]. The southern and western regions include black soil, optimal for cultivating cotton and soybeans, whilst the eastern part is distinguished by red and lateritic soils, suitable for pulses and oilseeds[29]. The Godavari River Basin influences the region's hydrology, providing agricultural water and sustaining wildlife. The shapefile was obtained from the official survey of India online maps portal[30]. The geometries of the selected districts are amalgamated to create a unified research region, enabling efficient geospatial analysis.

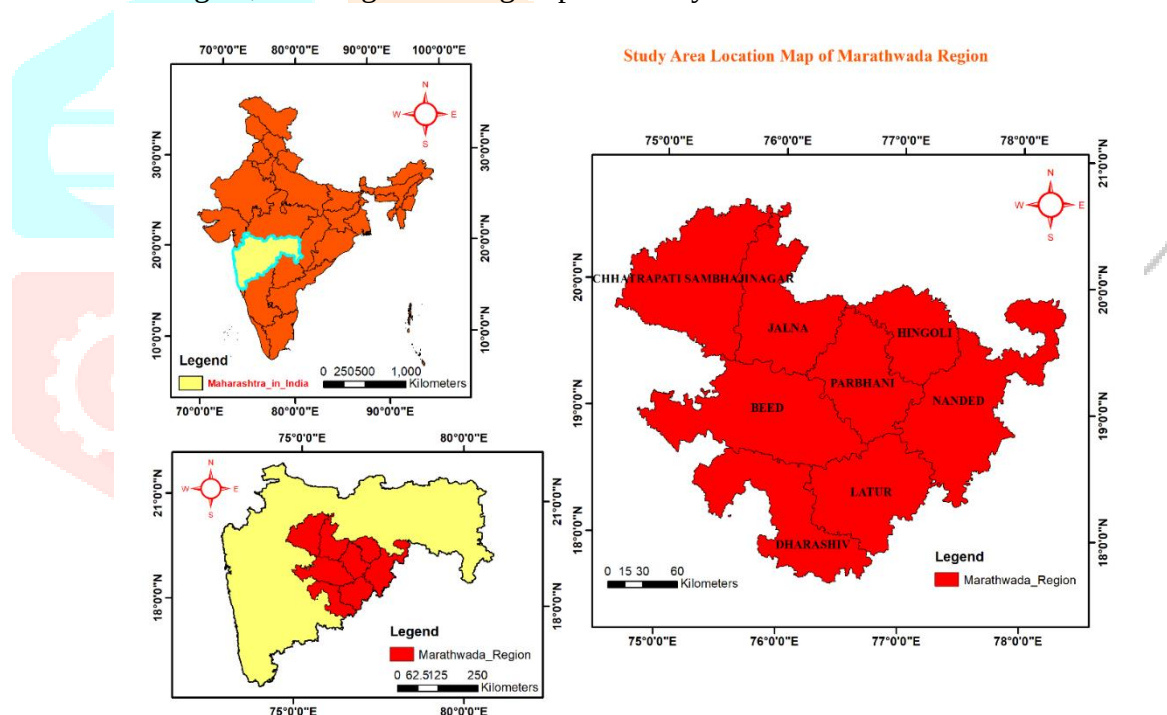


Figure 1: Study Area Map Location of Marathwada Region

3.2 Dataset

The present study uses NDVI imagery obtained from MODIS (Moderate Resolution Imaging Spectroradiometer) satellite data, particularly from the MOD09A1 product (Collection 061). The dataset covers the Marathwada area of Maharashtra, India, from 2013 to 2023. The MOD09A1 product offers 8-day surface reflectance composites with a spatial resolution of 500 meters, making it suitable for vegetation study. Data is obtained using GEE api using python by filtering photos according to geographical bounds (marathwada_geometry) and a temporal range (start_date to end_date). NDVI is calculated using three essential spectral bands: Red (Band 1 - sur_refl_b01), Near-Infrared (NIR) (Band 2 - sur_refl_b02), and Blue (Band 3 - sur_refl_b03). The NIR band (Band 2) is crucial for vegetation analysis, as healthy plants exhibit significant reflection of near-infrared light while absorbing visible light, particularly in the red spectrum. The pronounced reflectance in the near-infrared spectrum, relative to red light absorption, facilitates the differentiation of vegetation from other land cover types and the evaluation of plant health. Elevated NIR reflectance signifies abundant and robust vegetation, whereas decreased

values imply stress, sparse vegetation, or non-vegetated regions. To improve data quality, preprocessing involves cropping photos to the Marathwada area, calculating a median composite to reduce noise and cloud interference, and reprojecting the data to a uniform 1000m resolution for uniformity. The final processed images are exported in GeoTIFF format for subsequent analysis.

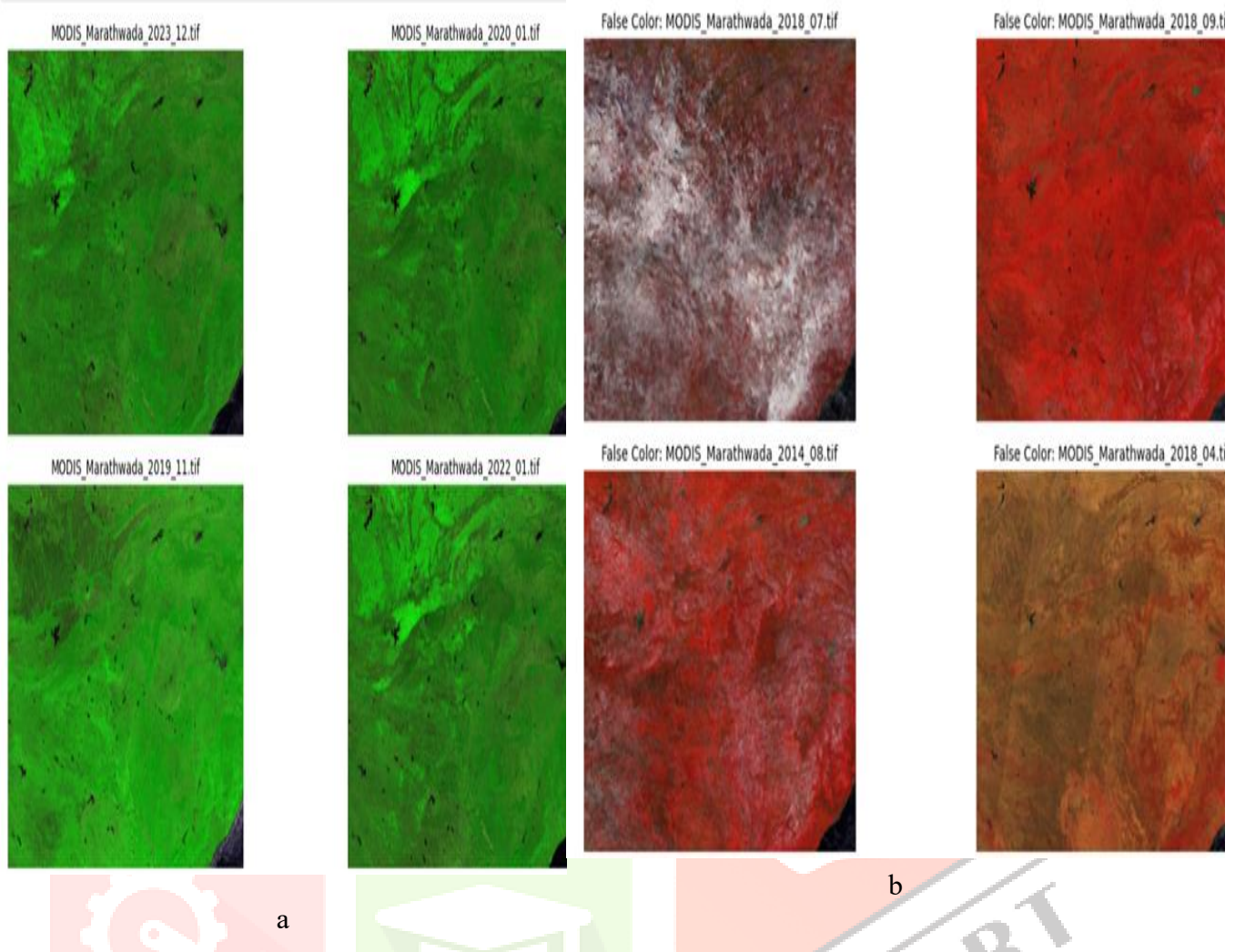


Figure 2: Samples images of True-color composite and False Color-color composite

True color composites are generated by allocating distinct spectral bands to the red, green, and blue channels to improve the interpretability of satellite images. The conventional true-color composite (RGB: 1, 2, 3) closely mirrors natural human perception, with flora shown in green, aquatic bodies in blue, and arid areas in brown as shown in figure 2 (a). Furthermore, false-color composites (RGB: 2, 1, 3) are used to emphasize vegetation health, with near-infrared (NIR) reflectance allocated to the red channel, resulting in healthy vegetation appearing brilliant red, whereas non-vegetated regions manifest in hues of brown or gray as shown in figure 2 (b).

3.3 Methodology

Figure 3 illustrates a systematic process for classifying vegetative canopy cover (VCC) utilising remote sensing data and machine learning algorithms. The procedure has five primary stages: Data Collection, Pre-processing, Calculation of Vegetation Indices, Selection of Training Samples, and Classification.

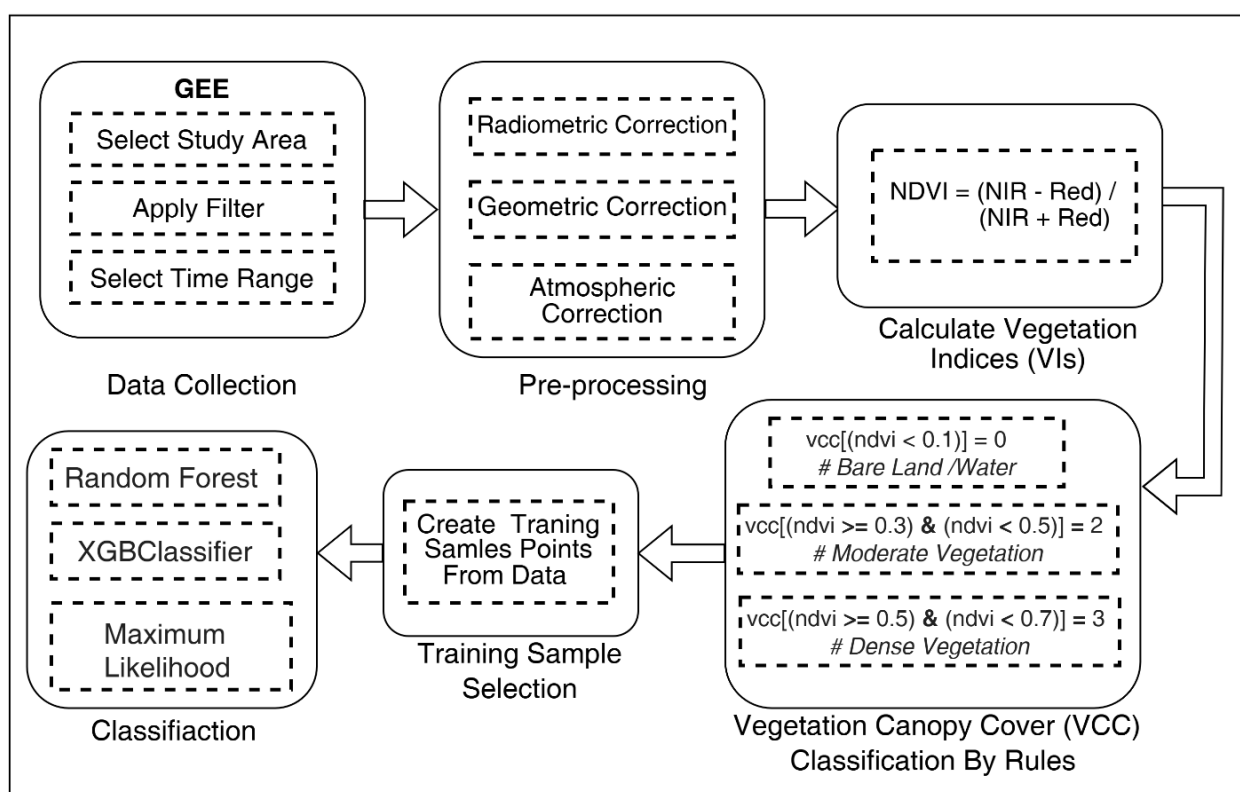


Figure 3:Methodology for Vegetation Canopy Cover (VCC) Classification

Data Collecting: The data collecting step use GEE, a cloud-based platform for acquiring remote sensing data. The study area is delineated according to established geographic boundaries. Filters are utilised to enhance the dataset by eliminating undesirable variations, and a designated time frame is chosen to maintain temporal consistency in the data.

Preprocess the Images: The pre-processing of satellite imagery entails several correction approaches to improve data quality and precision. Dark Object Subtraction (DOS) is a technique for atmospheric correction employed to reduce the impact of haze and aerosols in satellite images[31]. It establishes the lowest atmospheric influence by determining the first percentile pixel value of the image and removing it from each pixel value. To avert distortions, all adverse events are assigned a value of zero. This procedure enhances the precision of remote sensing data by mitigating atmospheric effects. This is accomplished by deducting the darkest pixel value at the 1st percentile (L_{min}) from the raw satellite pixel value (L_{raw}). The adjusted pixel value (L_{DOS}) is subsequently derived using the equation (1).

$$L_{DOS} = L_{raw} - L_{min} \quad \dots \dots \dots (1)$$

For sure accurate outcomes, any negative pixel values resulting from subtraction are assigned a value of zero, mathematically represented as equation (2)

$$L_{DOS} = \max(0, L_{raw} - L_{min}) \quad \dots \dots \dots (2)$$

Radiometric Correction is performed to correct sensor-related variations and ensure consistency in satellite imagery by converting raw digital numbers into reflectance values[32]. This process involves applying a gain and offset transformation, where the pixel values obtained after atmospheric correction[33, 34]. L_{DOS} are modified by a sensor-specific calibration coefficient (Gain) and refined using an additional correction factor (Offset). This is mathematically represented as Equation (3).

$$L_{corrected} = (L_{DOS} \times \text{Gain}) + \text{Offset} \quad \dots \dots \dots (3)$$

Here, Gain corrects sensor sensitivity differences; Offset fixes systematic measurement biases. This change ensures that data from different imaging conditions or sensors remains consistent. Moreover, any negative post-transformation results are given a value of zero to prevent false values, hence ensuring

physically relevant reflectance values. Exact Top of Atmosphere (TOA) reflectance is obtained by means of this process, which is required for quantitative remote sensing study.

Geometric correction is an essential procedure that aligns satellite imagery with actual geographic coordinates through the use of Ground Control Points (GCPs)[35]. GCPs are reference markers with established latitude and longitude coordinates that facilitate precise mapping between picture pixels and corresponding locations on the Earth's surface. Utilising these control points, the distorted or misaligned satellite image is rectified to ensure its characteristics correspond accurately with established geographic locations. A transformation matrix is utilised to execute geometric correction by repositioning pixels according to their original coordinates (x, y) and converting them into corrected coordinates (x', y'). The transformation parameters (a, b, c, d, e, f) are ascertained by interpolation methods, such Thin Plate Spline (TPS) interpolation or Affine Transformation. The TPS approach guarantees seamless and adaptable image warping to rectify distortions while preserving geometric precision. The transformation equation for geometric correction is represented as equation (4)

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \dots \dots \dots (4)$$

Here, (x', y') represent the adjusted pixel coordinates. Coordinates (x, y) represent the original pixel locations. a, b, c, d, e, f are transformation parameters obtained via Thin Plate Spline (TPS) or Affine Transformation.

Compute Vegetation Indices: Following to executing geometric correction, vegetation indices (VIs) are computed to assess vegetation health and density in satellite data. The Normalised Difference Vegetation Index (NDVI) is the most prevalent statistic among these. NDVI evaluates vegetation vitality by quantifying the disparity between near-infrared (NIR) and red-light reflectance, as healthy plants exhibit high NIR reflectance and absorb red light for photosynthesis[36-38]. The NDVI is calculated using the following equation (5). NDVI values go from -1 to +1, offering insights into surface attributes[39]. Negative readings, generally near -1, signify the existence of water, snow, or clouds[40]. Values near 0 indicate desolate terrain, stony landscapes, or urbanised regions with scant flora[41]. NDVI values ranging from 0.2 to 0.5 indicate sparse vegetation or grasslands, signifying moderate plant coverage[42]. Values over 0.6 denote dense and thriving vegetation, including woods or well-irrigated agricultural areas, reflecting vigorous plant growth and chlorophyll activity[43, 44].

$$NDVI = \frac{(NIR + Red)}{(NIR - Red)} \dots \dots \dots (5)$$

Here, NIR (Near-Infrared Band): Indicates the reflectance in the near-infrared spectrum, which is significantly reflected by healthy plants. Red (Red Band): Denotes the reflectance within the visible red spectrum, which is absorbed by chlorophyll in vegetation.

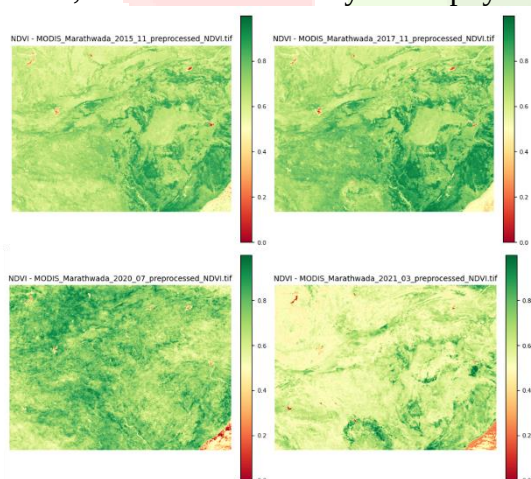


Figure 4: Normalised Difference Vegetation Index (NDVI)

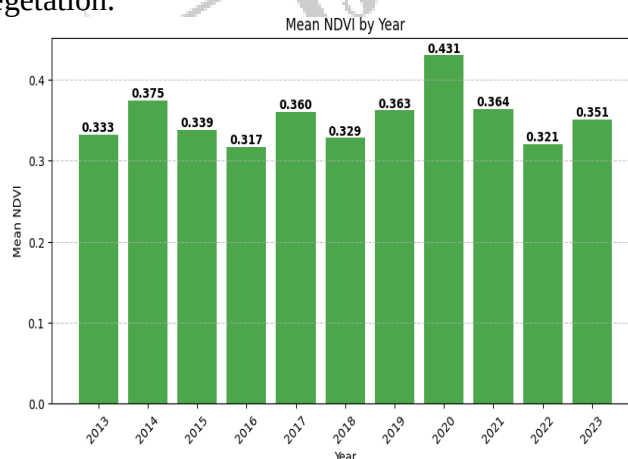


Figure 5: Mean NDVI by Year

The NDVI maps for several months and years are shown in Figure 4, which shows the density and health of the vegetation in the research area. Healthy vegetation is indicated by higher NDVI values (green), whereas sparse or stressed vegetation is indicated by lower values (red). The yearly mean NDVI values from 2013 to 2023 are displayed in Figure 5, emphasising the interannual differences in vegetative health.

While preceding years exhibit minor swings, 2020 saw the greatest mean NDVI (0.431), indicating a time of improving vegetative conditions.

Vegetation Canopy Cover (VCC) Classification: After to computing the Normalised Difference Vegetation Index (NDVI), the ensuing procedures are executed to identify vegetation and analyse land cover proficiently. The Vegetation Canopy Cover (VCC) classification employs a rule-based methodology to classify vegetation density levels according to NDVI readings as show in table 2. This classification aids in recognising various land cover types and evaluating vegetation health.

Table 2: Classification Rules for Vegetation Canopy Cover Assessment[45, 46].

NDVI Range	VCC Value	Vegetation Type
$\text{NDVI} < 0.1$	0	Bare Land / Water
$0.1 \leq \text{NDVI} < 0.3$	1	Sparse Vegetation
$0.3 \leq \text{NDVI} < 0.5$	2	Moderate Vegetation
$0.5 \leq \text{NDVI} < 0.7$	3	Dense Vegetation
$\text{NDVI} \geq 0.7$	4	Very Dense Vegetation

Once the VCC classification criteria are established, the next critical step is training sample selection to ensure accurate vegetation categorization. The process involves creating labeled training sample points from the pre-processed NDVI and VCC datasets, which serve as reference or ground truth data for training and evaluating machine learning classification models. A total of 124,548,600 samples were generated, consisting of NDVI values as input features and corresponding VCC classes as reference labels to guide the learning process and assess model performance. To build a well-balanced and diverse dataset, representative training samples are carefully chosen across different vegetation categories, including bare land, sparse vegetation, moderate vegetation, dense vegetation, and very dense vegetation. These samples accurately reflect real-world land cover conditions, ensuring that the dataset captures natural variations[47, 48]. Once selected, these training samples are used to prepare training data for machine learning algorithms, enabling classifiers to learn patterns associated with different vegetation types. A well-structured training dataset enhances model performance, classification accuracy, and generalization ability, leading to more reliable vegetation mapping and land cover analysis. These ground truth samples help machine learning models precisely differentiate among various vegetation classes, ultimately improving the robustness and dependability of the classification system. In this process, NDVI images and VCC classified images are loaded from predefined directories (preprocessed_data/NDVI and preprocessed_data/VCC). The script ensures that NDVI and VCC files are correctly matched, and it extracts NDVI values as features (X) and VCC classes as labels (y). Invalid or missing values (NaNs) are removed before appending the cleaned data to the training set. This method results in a well-structured dataset, as demonstrated by the successful loading of 124,548,600 samples for training, ready for further classification using machine learning techniques.

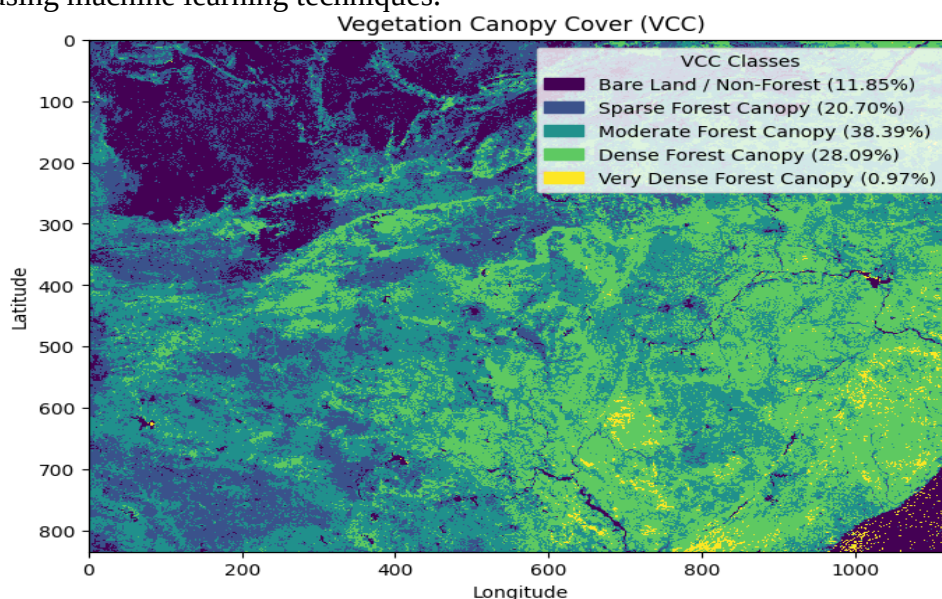


Figure 6: Percentage Values of Each Vegetation Canopy Cover (VCC) Class Sample Image

Figure 6 shows the VCC classification based on NDVI data. The map divides vegetation into five types: Bare Land/Non-Forest (11.85%), Sparse Forest Canopy (20.70%), Moderate Forest Canopy (38.39%), Dense Forest Canopy (28.09%), and Very Dense Forest Canopy (0.97%). Dark purple areas represent barren land with little to no vegetation, while green to yellow shades indicate increasing vegetation density. Most of the area is covered by moderate and dense forest, while very dense forest regions are limited. This distribution highlights variations in forest cover influenced by climate, land use, and ecological factors.

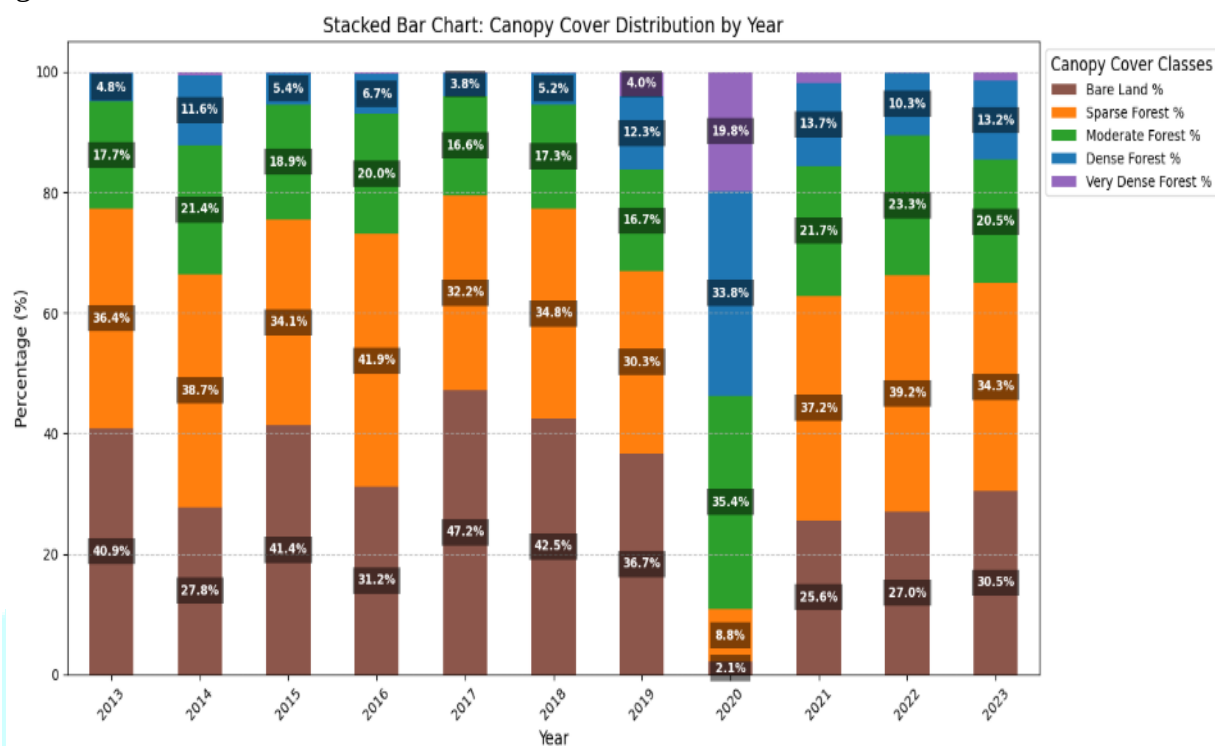


Figure 7: Yearly Distribution of Canopy Cover

Figure 7 shows the yearly distribution of canopy cover using a stacked bar chart. It categorizes vegetation into five types: Bare Land, Sparse Forest, Moderate Forest, Dense Forest, and Very Dense Forest, displaying their percentage changes over the years. Bare Land fluctuates between 25.6% (2021) and 47.2% (2016), while Sparse Forest varies from 27.8% (2014) to 41.9% (2016). Moderate Forest remains relatively stable, ranging between 16.6% (2017) and 23.3% (2022). Dense Forest shows notable fluctuations, peaking at 33.8% in 2020. Very Dense Forest remains minimal, reaching its highest percentage at 19.8% in 2020. These yearly changes highlight the impact of climate, land use practices, and environmental factors on vegetation dynamics.

Classification Using Machine Learning: Machine learning techniques are employed for vegetation classification, including Random Forest (RF), XGBoost Classifier (XGBC), and Maximum Likelihood Classification (MLC). These models use labeled training data to learn vegetation patterns effectively. RF is an ensemble learning approach that generates multiple decision trees and consolidates their results, ensuring strong and dependable categorization. It also excels in defining intricate relationships in vegetation classification by incorporating various spectral indices and spatial attributes. XGBC is a powerful machine learning technique that is tuned for speed and accuracy, able to handle large, high-dimensional datasets and handle missing data. It is widely used in vegetation classification and land cover analysis due to its quick execution time and superior performance. MLE classifier is a probabilistic supervised classification method that uses Gaussian probability distributions to distinguish between different types of vegetation. It provides confidence levels for classification decisions and is an excellent choice for vegetation mapping. The study also employs an unsupervised machine learning approach to classify vegetation canopy cover using Principal Component Analysis (PCA) and K-means clustering. PCA reduces the dimensionality of multi-temporal NDVI data while preserving important temporal trends in vegetation dynamics. The compressed data is then grouped using the K-means method, classifying pixels into distinct vegetation categories based on their temporal NDVI profiles. Validation metrics such as the Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, Dunn Index, and spatial autocorrelation are used to evaluate classification performance. The classification results are further examined through confusion matrices that juxtapose the clusters with traditional NDVI threshold-defined

categories. Visual depictions of spatial distribution and class-specific percentage analyses facilitate the transformation of uncontrolled observations into ecologically relevant categories, improving forest monitoring and land cover evaluation.

Model Evaluation: After concluding the vegetation classification through machine learning models, including RF, XGBC and MLE, it is imperative to evaluate the accuracy and reliability of the classification outcomes. This is accomplished by quantitative evaluation criteria, such as the categorization report, confusion matrix, and Cohen's Kappa coefficient.

Classification Report: The classification report offers a comprehensive overview of the model's performance through essential evaluation metrics:

- **Precision (P):** The ratio of accurately predicted positive observations to the total anticipated positive observations, calculated as follows equation (6). where TP denotes true positives and FP signifies false positives[49, 50].

$$Precision = \frac{TP}{TP + FP} \dots \dots \dots (6)$$

- **Recall (R) or Sensitivity:** The ratio of accurately predicted positive observations to actual positives is shown in Equation (7), where FN denotes false negatives[50, 51].

$$Recall = \frac{TP}{TP + FN} \dots \dots \dots (7)$$

- **F1-Score:** The harmonic mean of precision and recall, which balances both metrics using equation (8), reveals that a greater F1-score reflects superior model performance[50, 52].

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \dots \dots \dots (8)$$

- **Overall Accuracy:** The proportion of accurately identified samples to the overall number of samples. Perform calculations utilizing equation (9)[50, 53].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots \dots \dots (9)$$

Confusion Matrix: The confusion matrix shows a comprehensive juxtaposition of the projected and actual categories. It is a square matrix that elucidates the distribution of accurate and erroneous classifications among various vegetation classes[54]. The matrix structure is represented by equation (10). where: TP: True Positives, FP: False Positives, FN: False Negatives and TN :True Negatives[55].

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \dots \dots \dots (10)$$

Cohen's Kappa Coefficient (κ): Cohen's Kappa Score (κ) is a statistical measure that evaluates the concordance between anticipated and actual classifications, while controlling for chance agreement[56]. It offers a more comprehensive assessment than accuracy alone. The Kappa coefficient is computed using equation (11).

$$k = \frac{p_o - p_e}{1 - p_e} \dots \dots \dots (11)$$

The Kappa coefficient (κ) varies from -1 to 1, signifying varying degrees of concordance between projected and actual classifications[57]. A κ value of 1 signifies perfect agreement, whereas values beyond 0.8 denote significant agreement. A score ranging from 0.6 to 0.8 indicates considerable agreement, whereas values between 0.4 and 0.6 denote fair agreement. A Kappa score below 0.4 indicates weak agreement in categorization, whereas a κ value of 0 signifies that the agreement is entirely attributable to chance.

IV. RESULTS:

The study evaluated three classification algorithms, RF Extreme XGBC and MLC, for vegetation classification using NDVI, VCC, and spectral metrics. The dataset consisted of 124,548,600 samples, divided into 80% for training and 20% for testing. The RF classifier, consisting of 100 decision trees, achieved 100% accuracy on the test data, demonstrating exceptional capability in capturing intricate relationships between NDVI values and vegetation types. The XGBC model, trained using 100 estimators with a learning rate of 0.1, completed training in 7,557.18 seconds, reaching a 99.29% accuracy with high precision and recall across all classes. The MLC model, utilizing Gaussian assumptions, was the fastest to train at 7.14 seconds and achieved 98.31% accuracy. Despite minor performance drops in the Dense Forest

class, its ability to correct previous errors and manage missing data efficiently makes it highly suitable for large-scale remote sensing applications. The findings underscore the effectiveness of ensemble-based classifiers for vegetation classification using NDVI data. RF and XGBC proved highly effective for forest canopy classification, while MLC offered faster computation with moderate accuracy. Table 3 depicts the Classification Performance and Accuracy Assessment

Table 4: Classification Performance and Accuracy Assessment

Classifier	Accuracy (%)	Precision	Recall	F1-Score
Random Forest (RF)	100.00	1.00	1.00	1.00
XGBoost (XGBC)	99.29	~0.99	~0.99	~0.99
MLC	98.31	~0.99	~0.98	~0.98

The figure 6 shows confusion matrices for RF, XGBC, and MLC models illustrate the effectiveness of rule-based vegetation canopy cover classification using NDVI. Each model was tested across five canopy classes: Bare Land (0), Sparse Vegetation (1), Moderate Vegetation (2), Dense Vegetation (3), and Very Dense Vegetation (4). All three classifiers show strong diagonal dominance, reflecting high accuracy in assigning pixels to their correct classes. The RF model achieves near-perfect predictions with minimal misclassifications, especially in Bare Land, Sparse, and Very Dense categories, yielding a Cohen's Kappa Score of 1.00, indicating perfect agreement with ground truth. XGBC performs comparably well (Kappa Score: 0.99), with only minor confusions, particularly between Moderate and Dense classes. MLC, while slightly less accurate, still produces reliable classification outcomes (Kappa Score: 0.98), though it shows some overlap between adjacent classes, especially in Moderate and Dense Vegetation zones.

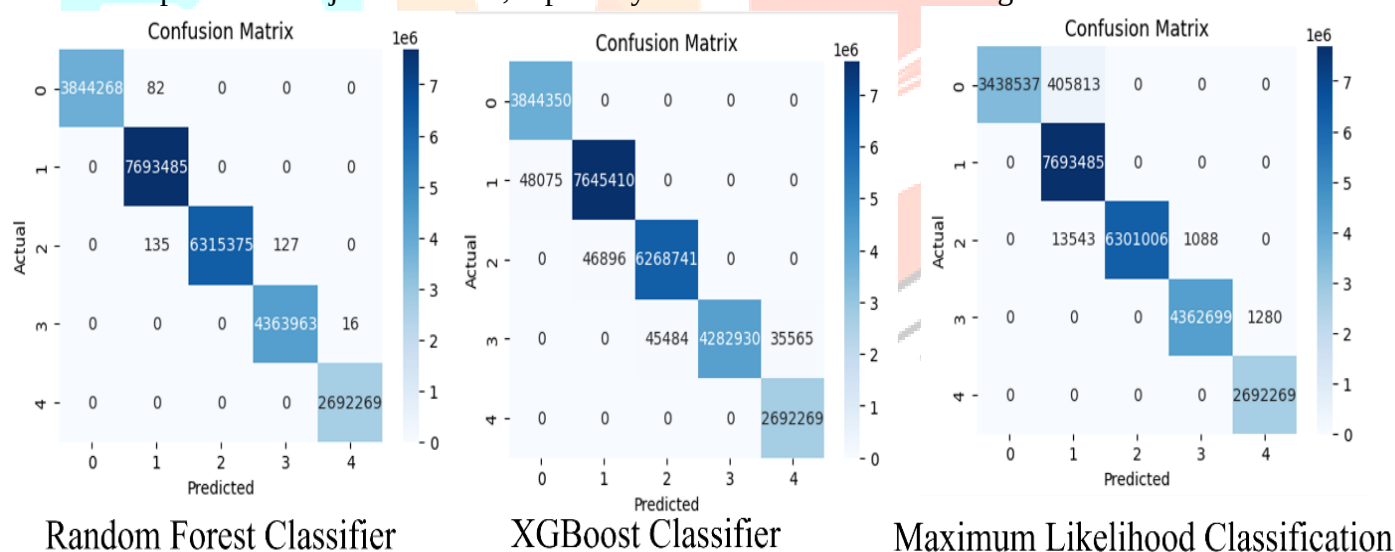


Figure 6: Confusion Matrices for Rule-Based Vegetation Canopy Cover Classification Using NDVI

Figure 7 presents the results of Principal Component Analysis (PCA) and K-Means clustering applied to vegetation classification. The PCA Explained Variance plot shows that the first five principal components retain 94% of the total variance, with PC1 contributing 35.2%, PC2 21.8%, PC3 15.4%, PC4 12.3%, and PC5 9.3%. The PCA + K-Means Vegetation Class Distribution histogram reveals that the dataset is divided into five vegetation classes with the following proportions: Class 0 (23.7%), Class 1 (18.5%), Class 2 (21.9%), Class 3 (15.8%), and Class 4 (20.1%). The 3D PCA Space with Cluster Assignments confirms that the clusters exhibit distinguishable separations in the transformed space. The Temporal NDVI Profiles by Vegetation Class graph highlights seasonal variations, with Class 4 exhibiting the highest NDVI fluctuation range between 0.4 and 0.9, while Class 0 remains consistently low, ranging from 0.1 to 0.3.

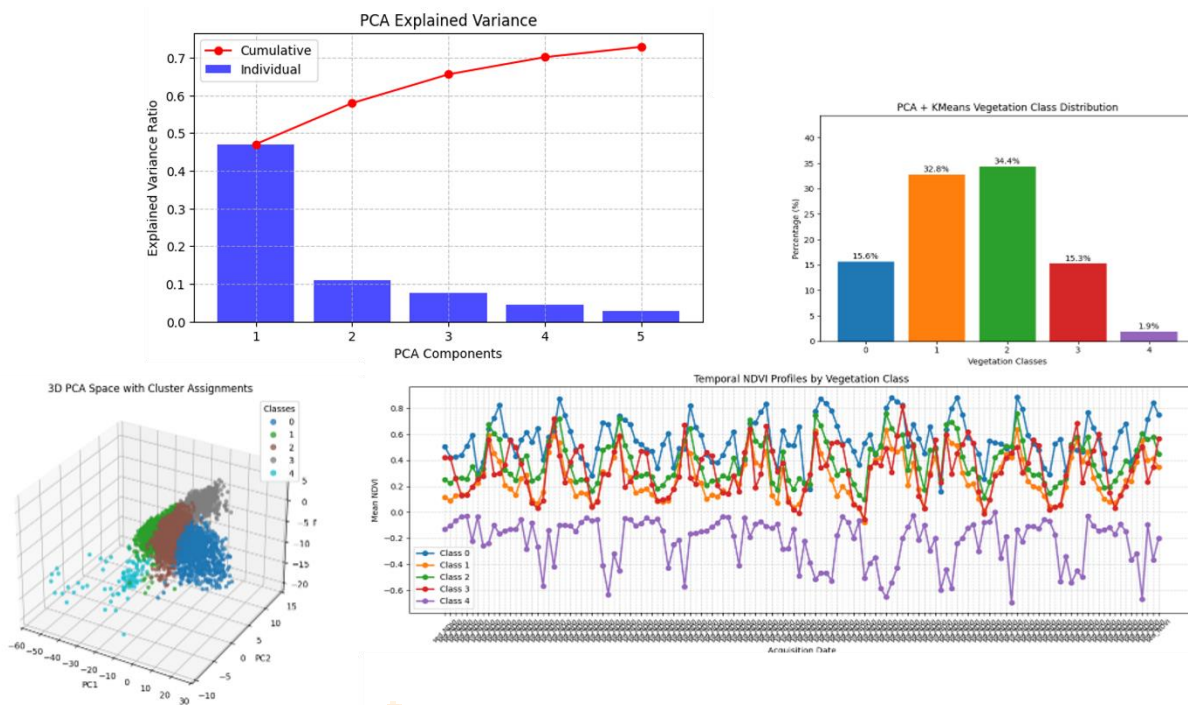


Figure 7: PCA + K-Means Vegetation Class Distribution, 3D PCA Space with Cluster and Temporal NDVI Profiles by Vegetation Class

Figure 8 provides a spatial distribution of vegetation classes, where vegetation clusters are spatially well-distributed, reflecting expected ecological patterns. The PCA Component Scatter Plots illustrate distinct separations among clusters, confirming that K-Means has effectively segmented the dataset. The histograms of PCA components further reveal that Class 4 have the highest concentration of high NDVI values, whereas Class 0 predominantly consists of low NDVI regions. The silhouette score of 0.31 suggests moderate clustering performance, indicating reasonable separation between clusters.

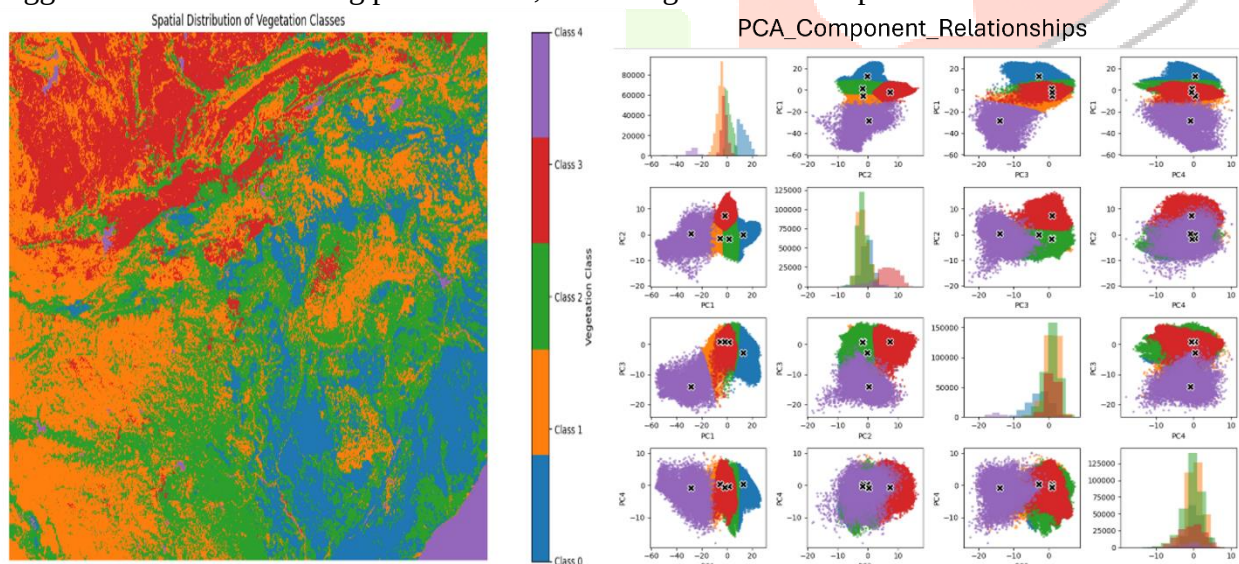


Figure 8: Spatial Distribution of Vegetation Classes and PCA Component Scatter Plots

In conclusion, the PCA + K-Means analysis on multi-temporal NDVI data produced a silhouette score of 0.31, which indicates moderate clustering quality. This suggests that while certain vegetation patterns are separable, the clusters are not highly distinct. Therefore, these unsupervised clusters should be considered exploratory groupings rather than definitive vegetation classes.

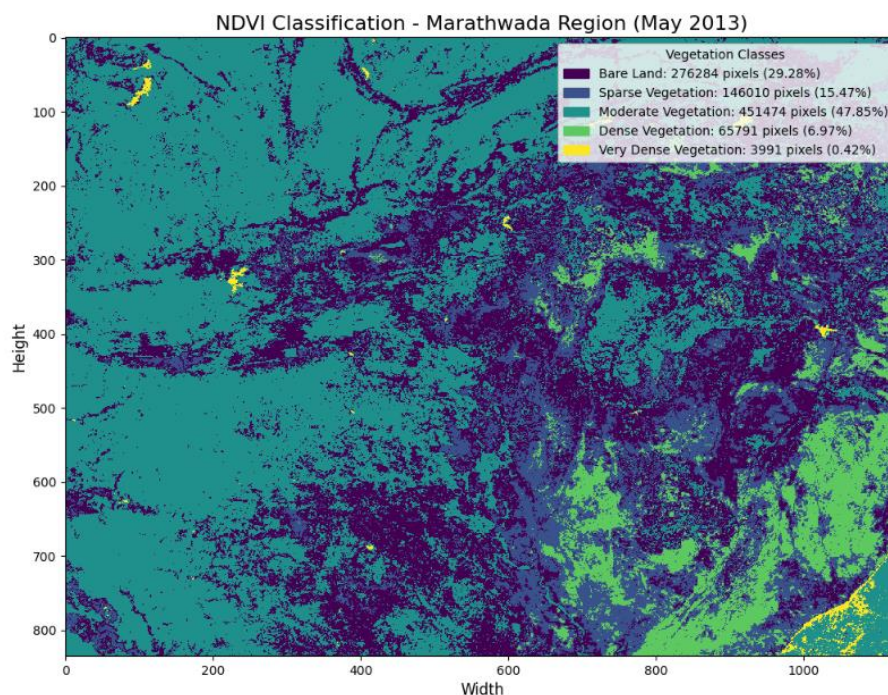


Figure 9: Sample Output of NDVI-based vegetation classification using PCA and K-means clustering

Figure 9 depicts Marathwada area's vegetation classification for May 2013 is divided into five groups: Bare Land (27.28%), Sparse Vegetation (15.47%), Moderate Vegetation (47.85%), Dense Vegetation (6.97%), and Very Dense Vegetation (0.42%). Moderate vegetation is prevalent, indicating a balanced ecosystem. Denser vegetation is found in lesser patches, indicating forested zones or well-watered agricultural lands. This categorization aids in environmental supervision, land-use administration, and agricultural strategies.

V. DISCUSSION

The study evaluates supervised and unsupervised classification methods for vegetation mapping using NDVI, VCC, and spectral metrics. The supervised classifiers RF, XGBC, and MLC showed high accuracy, with RF achieving flawless classification (100%), XGBC (99.29%), and MLC (98.31%). RF was the most resilient model, despite longer training duration. XGBC provided a balanced compromise between computational efficiency and predictive capability, making it suitable for remote sensing applications. MLC showed good performance but faced challenges with bare land and sparse vegetation classes due to spectral overlaps. The unsupervised PCA + K-Means clustering method differentiated vegetation categories using NDVI values, providing a viable option when labeled data is unavailable. Despite being computationally efficient, the PCA + K-Means approach lacked classification precision, making it suitable for preliminary vegetation mapping and exploratory data analysis. RF was the most effective classifier for NDVI-based vegetation mapping, followed closely by XGBC. The accuracy values primarily reflect internal consistency within the dataset rather than true real-world predictive performance. In heterogeneous regions such as Marathwada, factors like mixed pixels, seasonal variability, and spectral overlap may lead to lower accuracy when validated against independent or higher-resolution datasets.

VI. CONCLUSION

The study evaluated the effectiveness of supervised and unsupervised classification algorithms for vegetation classification using NDVI, VCC, and spectral metrics. Results showed that supervised classifiers (RF, XGBC, and MLC) achieved the highest accuracy (100%), followed by XGBC (99.29%) and MLC (98.31%). XGBC was the most efficient but had slight misclassifications due to spectral overlaps. Unsupervised PCA combined with K-Means effectively classified vegetation types but had moderate cluster separability. Despite PCA + K-Means being effective for preliminary vegetation mapping, supervised classifiers (RF and XGBC) are still preferred for precise vegetation classification and land-use observation. Future research should focus on merging deep learning methods and adding more spectral bands.

REFERENCES

- [1] S. W. Running, "Ecosystem disturbance, carbon, and climate," *Science*, vol. 321, no. 5889, pp. 652-653, 2008.
- [2] G. B. Bonan, "Forests and climate change: forcings, feedbacks, and the climate benefits of forests," *science*, vol. 320, no. 5882, pp. 1444-1449, 2008.
- [3] J. A. Foley *et al.*, "Global consequences of land use," *science*, vol. 309, no. 5734, pp. 570-574, 2005.
- [4] P. Kowe, O. Mutanga, and T. Dube, "Advancements in the remote sensing of landscape pattern of urban green spaces and vegetation fragmentation," *International Journal of Remote Sensing*, vol. 42, no. 10, pp. 3797-3832, 2021.
- [5] S. Sharma *et al.*, "Remote Sensing and GIS in Natural Resource Management: Comparing tools and emphasizing the importance of in-situ data," *Remote Sensing*, vol. 16, no. 22, p. 4161, 2024.
- [6] M. J. Mashala, T. Dube, B. T. Mudereri, K. K. Ayisi, and M. R. Ramudzuli, "A systematic review on advancements in remote sensing for assessing and monitoring land use and land cover changes impacts on surface water resources in semi-arid tropical environments," *Remote Sensing*, vol. 15, no. 16, p. 3926, 2023.
- [7] Z. Zhu *et al.*, "Benefits of the free and open Landsat data policy," *Remote sensing of environment*, vol. 224, pp. 382-385, 2019.
- [8] A. De La Iglesia Martinez and S. Labib, "Demystifying normalized difference vegetation index (NDVI) for greenness exposure assessments and policy interventions in urban greening," *Environmental Research*, vol. 220, p. 115155, 2023.
- [9] Q. Zhao and Y. Qu, "The retrieval of ground NDVI (normalized difference vegetation index) data consistent with remote-sensing observations," *Remote Sensing*, vol. 16, no. 7, p. 1212, 2024.
- [10] A. M. Márquez-Romance, B. E. Fariás-de Márquez, and E. Guevara-Pérez, "Land use and land cover change detection using satellite remote sensing techniques in a tropical basin," *Environmental Quality Management*, vol. 31, no. 4, pp. 183-196, 2022.
- [11] G. M. Gandhi, S. Parthiban, N. Thummalu, and A. Christy, "Ndv: Vegetation change detection using remote sensing and gis—A case study of Vellore District," *Procedia computer science*, vol. 57, pp. 1199-1210, 2015.
- [12] S. Swain, S. K. Mishra, A. Pandey, and D. Dayal, "Assessment of drought trends and variabilities over the agriculture-dominated Marathwada Region, India," *Environmental Monitoring and Assessment*, vol. 194, no. 12, p. 883, 2022.
- [13] S. S. Kulkarni, B. D. Wardlow, Y. A. Bayissa, T. Tadesse, M. D. Svoboda, and S. S. Gedam, "Developing a remote sensing-based combined drought indicator approach for agricultural drought monitoring over Marathwada, India," *Remote Sensing*, vol. 12, no. 13, p. 2091, 2020.
- [14] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International journal of Remote sensing*, vol. 28, no. 5, pp. 823-870, 2007.
- [15] P. S. Chavez Jr, "An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data," *Remote sensing of environment*, vol. 24, no. 3, pp. 459-479, 1988.
- [16] G. Chander and B. Markham, "Revised Landsat-5 TM radiometric calibration procedures and postcalibration dynamic ranges," *IEEE Transactions on geoscience and remote sensing*, vol. 41, no. 11, pp. 2674-2677, 2003.
- [17] H. Tamiminia, B. Salehi, M. Mahdianpari, L. Quackenbush, S. Adeli, and B. Brisco, "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review," *ISPRS journal of photogrammetry and remote sensing*, vol. 164, pp. 152-170, 2020.
- [18] Q. Wu, "geemap: A Python package for interactive mapping with Google Earth Engine," *Journal of Open Source Software*, vol. 5, no. 51, p. 2305, 2020.
- [19] C. Nguyen, V. Sagan, M. Maimaitiyiming, M. Maimaitijiang, S. Bhadra, and M. T. Kwasniewski, "Early detection of plant viral disease using hyperspectral imaging and deep learning," *Sensors*, vol. 21, no. 3, p. 742, 2021.
- [20] A. Gyawali, H. Adhikari, M. Aalto, and T. Ranta, "From simple linear regression to machine learning methods: Canopy cover modelling of a young forest using planet data," *Ecological Informatics*, vol. 82, p. 102706, 2024.
- [21] M. Maimaitijiang, V. Sagan, S. Bhadra, C. Nguyen, T. Mockler, and N. Shakoar, "A fully automated and fast approach for canopy cover estimation using super high-resolution remote

- sensing imagery," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 219-226, 2021.
- [22] G. R. Morgan, D. Zlotnick, L. North, C. Smith, and L. Stevenson, "Deep Learning for Urban Tree Canopy Coverage Analysis: A Comparison and Case Study," *Geomatics*, vol. 4, no. 4, pp. 412-432, 2024.
- [23] M. Claverie *et al.*, "The Harmonized Landsat and Sentinel-2 surface reflectance data set," *Remote sensing of environment*, vol. 219, pp. 145-161, 2018.
- [24] K. Rohith, V. P. S. H. V. Vudatha, T. K. Rathna, and G. J. Lakshmi, "Remote sensing-based vegetation types classification using landsat and change analysis using setinel-2a in google earth engine," 2023.
- [25] O. Smith and H. Cho, "An Open-Source Canopy Classification System Using Machine-Learning Techniques Within a Python Framework," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 46, pp. 175-182, 2021.
- [26] V. Nasiri, A. A. Darvishsefat, H. Arefi, V. C. Griess, S. M. M. Sadeghi, and S. A. Borz, "Modeling forest canopy cover: a synergistic use of Sentinel-2, aerial photogrammetry data, and machine learning," *Remote Sensing*, vol. 14, no. 6, p. 1453, 2022.
- [27] R. Saini, "Integrating vegetation indices and spectral features for vegetation mapping from multispectral satellite imagery using AdaBoost and random forest machine learning classifiers," *Geomatics and Environmental Engineering*, vol. 17, no. 1, pp. 57-74, 2023.
- [28] U. V. Pawar, "Rainfall distribution and trends over the semi-arid Marathwada region of Maharashtra, India," *Arabian Journal of Geosciences*, vol. 15, no. 23, p. 1738, 2022.
- [29] R. V. Hajare, "STUDY OF FRUIT FARMING IN SOUTH MAHARASHTRA PLATEAU," 2023.
- [30] S. o. India. "Digital Products – Online Maps Portal." Survey of India. https://onlinemaps.surveyofindia.gov.in/Digital_Product_Show.aspx (accessed).
- [31] Y. Guo *et al.*, "Dark-object subtraction atmosphere correction for water body information extraction in Zhuhai-1 hyperspectral imagery," *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 27, no. 2, pp. 382-391, 2024.
- [32] P.-O. Olsson *et al.*, "Radiometric correction of multispectral uas images: Evaluating the accuracy of the parrot sequoia camera and sunshine sensor," *Remote Sensing*, vol. 13, no. 4, p. 577, 2021.
- [33] R. Richter, "Atmospheric correction methods for optical remote sensing imagery of land," in *Advances in Environmental Remote Sensing: Remote Sensing Applications*: Taylor Francis, 2011, pp. 161-172.
- [34] P. S. Chavez, "Image-based atmospheric corrections-revisited and improved," *Photogrammetric engineering and remote sensing*, vol. 62, no. 9, pp. 1025-1035, 1996.
- [35] T. Nguyen, "Optimal ground control points for geometric correction using genetic algorithm with global accuracy," *European Journal of Remote Sensing*, vol. 48, no. 1, pp. 101-120, 2015.
- [36] J. D. Stamford, S. Violet-Chabrand, I. Cameron, and T. Lawson, "Development of an accurate low cost NDVI imaging system for assessing plant health," *Plant Methods*, vol. 19, no. 1, p. 9, 2023.
- [37] P. Lemenkova and O. Debeir, "Multispectral satellite image analysis for computing vegetation indices by R in the Khartoum Region of Sudan, Northeast Africa," *Journal of imaging*, vol. 9, no. 5, p. 98, 2023.
- [38] J. Liu, J. Fan, C. Yang, F. Xu, and X. Zhang, "Novel vegetation indices for estimating photosynthetic and non-photosynthetic fractional vegetation cover from Sentinel data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 109, p. 102793, 2022.
- [39] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao, "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing," *Journal of Forestry Research*, vol. 32, no. 1, pp. 1-6, 2021.
- [40] F. A. Hadi, W. A. Ahmad, and F. H. Mahmood, "Efficiency of Thematic Mapper Data for Detecting the Changes of Land Use-Land Cover in Mosul Lake and Surrounding Area Northern Iraq," *Iraqi Journal of Science*, pp. 662-676, 2019.
- [41] M. Mazzarino and J. T. Finn, "An NDVI analysis of vegetation trends in an Andean watershed," *Wetlands Ecology and Management*, vol. 24, pp. 623-640, 2016.
- [42] A. Taufik, S. S. S. Ahmad, and A. Ahmad, "Classification of Landsat 8 satellite data using NDVI thresholds," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, no. 4, pp. 37-40, 2016.

- [43] R. Khalil, M. S. Khan, Y. Hasan, N. Nacer, and S. Khan, "Supervised ndvi composite thresholding for arid region vegetation mapping," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14420-14427, 2024.
- [44] R. Wahlang and S. Chaturvedi, "Relationship Between Above-Ground Biomass and Different Vegetation Indices of Forests of Ri-Bhoi District, Meghalaya, India," *Int. J. Eng. Tech. Res.*, vol. 9, 2020.
- [45] J. A. Sobrino and N. Raissouni, "Toward remote sensing methods for land cover dynamic monitoring: Application to Morocco," *International journal of remote sensing*, vol. 21, no. 2, pp. 353-366, 2000.
- [46] M. Hansen, S. Stehman, M. Piao, and Z. Hongyan, "Temporal Forest Change Detection and Forest Health Assessment using Remote Sensing," in *IOP Conf. Ser.: Earth Environ. Sci.*, 2017, vol. 19, no. 012017, pp. 1755-1315.
- [47] W. Song, X. Mu, G. Ruan, Z. Gao, L. Li, and G. Yan, "Estimating fractional vegetation cover and the vegetation index of bare soil and highly dense vegetation with a physically based method," *International journal of applied earth observation and geoinformation*, vol. 58, pp. 168-176, 2017.
- [48] S. Huang and F. Siegert, "Land cover classification optimized to detect areas at risk of desertification in North China based on SPOT VEGETATION imagery," *Journal of Arid Environments*, vol. 67, no. 2, pp. 308-327, 2006.
- [49] Y. Jiao and P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," *Quantitative Biology*, vol. 4, pp. 320-330, 2016.
- [50] A. Tharwat, "Classification assessment methods," *Applied computing and informatics*, vol. 17, no. 1, pp. 168-192, 2021.
- [51] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.
- [52] R. Diallo, C. Edalo, and O. O. Awe, "Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score," in *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*: Springer, 2024, pp. 283-312.
- [53] G. M. Foody, "Sample size determination for image classification accuracy assessment and comparison," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273-5291, 2009.
- [54] G. Thoonen, K. Hufkens, J. V. Borre, T. Spanhove, and P. Scheunders, "Accuracy assessment of contextual classification results for vegetation mapping," *International journal of applied earth observation and geoinformation*, vol. 15, pp. 7-15, 2012.
- [55] J. Liang, "Confusion matrix: Machine learning," *POGIL Activity Clearinghouse*, vol. 3, no. 4, 2022.
- [56] P. Verma, A. Raghubanshi, P. K. Srivastava, and A. Raghubanshi, "Appraisal of kappa-based metrics and disagreement indices of accuracy assessment for parametric and nonparametric techniques used in LULC classification and change detection," *Modeling Earth Systems and Environment*, vol. 6, pp. 1045-1059, 2020.
- [57] T. Wan, H. Jun, W. Pan, and H. Hua, "Kappa coefficient: a popular measure of rater agreement," *Shanghai archives of psychiatry*, vol. 27, no. 1, p. 62, 2015.