An Unsupervised-to-Supervised Framework for Vegetation Mapping Using Spectral Indices

Bishal Banik
Dept. of Computer Science & Engg.
Adamas University
Kolkata, India
bishal2001banik@gmail.com

Subhash Mukherjee
Dept. of Computer Science & Engg.
Assam University
Silchar, India
subhashmakaut@gmail.com

Somnath Mukhopadhyay
Dept. of Computer Science & Engg.
Assam University
Silchar, India
som.cse@live.com

Sunita Sarkar
Dept. of Computer Science & Engg.
Assam University
Silchar, India
sarkarsunita2601@gmail.com

Wangjam Niranjan Singh
Dept. of Computer Science & Engg.
Assam University
Silchar, India
niranwang@gmail.com

***Abstract*—Vegetation detection from satellite imagery is essential for ecological monitoring, agricultural planning, and land use analysis. Traditional supervised methods often demand extensive labeled datasets, which are time-consuming to prepare. This study presents a hybrid unsupervised-to-supervised framework for vegetation detection using unlabelled multispectral satellite imagery. Initially, vegetation pixels are labelled using NDVI thresholding, while water regions are excluded via NDWI-based masking. These selected pixels undergo dimensionality reduction using Principal Component Analysis (PCA) and are clustered using k-means++. The clustered pixels with the highest NDVI mean are labelled as vegetation, forming labels used to train a Random Forest classifier. The trained model then classifies the entire image, producing a refined vegetation map by reapplying NDWI to exclude riverine areas. Validation through Precision, Recall, F1 Score, Accuracy, AUC-ROC, Confusion matrix, Davies–Bouldin Score, and Dunn Index confirms effective clustering, and visual outputs highlight the framework's reliability. This method offers a scalable, label-free approach that provides an interpretable and efficient solution for vegetation mapping in complex landscapes.**

***Index Terms*—vegetation mapping, k-means++, Random Forest, NDVI, NDWI.**

I. Introduction

Vegetation detection and mapping play a vital role in environmental monitoring, agricultural assessment, land cover classification, and ecological conservation. Traditional vegetation mapping techniques used supervised learning methods that require large amounts of labelled data, which is not always available for remote regions. This challenge motivates the development of semi-supervised or hybrid methodologies that can leverage unlabelled data to generate reliable vegetation masks with minimal manual intervention. Vegetation indices such as the Normalized Difference Vegetation Index (NDVI) [1] is widely used to detect the photosynthetically active regions. However, particularly in diverse landscapes, the threshold-based NDVI segmentation may incorrectly classify non-vegetation areas with comparable spectral reflectance, like water bodies. In order to improve vegetation accuracy and eliminate such restrictions, the Normalised Difference Water Index (NDWI) is frequently used to isolate water features, such as rivers and wetlands.

This study proposes a hybrid unsupervised-supervised pipeline for vegetation detection using k-means++ clustering and a Random Forest (RF) classifier, applied on multispectral satellite imagery. The method first identifies high-confidence vegetation pixels by applying NDVI-based thresholding and removes water regions detected via NDWI. Then the pixels are reduced in dimensionality using Principal Component Analysis (PCA) and clustered using k-means++. The cluster with the highest NDVI mean is labeled as vegetation, creating a binary label set for training a supervised Random Forest model. The train set is applied to the image to produce a dense vegetation map.

This method enhances the unsupervised learning for initial label generation and supervised learning for spatial generalization. It eliminates the need for manually labelled ground truth while maintaining high classification accuracy. Furthermore, the Davies–Bouldin Score and Dunn Index are used to evaluate clustering quality. The accuracy, precision, recall, F1 Score, and Cohen's kappa score for classification are evaluated to validate the proposed method. The results are also supported by comprehensive visualizations demonstrating the effectiveness of the proposed framework.

The rest of the paper is organized as follows. Section

II presents related work in the field, followed by Section III where data description is discussed, and Section IV presents the proposed methodology of the paper. Then, the results and discussion are given in section V. Finally, we conclude this study in section VI.

II. Related Work

Decades of research have been focused on improving the accuracy and efficiency of vegetation detection from multispectral imagery, making it a fundamental topic in the field of remote sensing. Vegetation monitoring was revolutionized by the introduction of spectral indices in early landmark studies. Due to its sensitivity to chlorophyll content, the Normalized Difference Vegetation Index (NDVI), first presented by Tucker [2] and popularized by Rouse et al. work with ERTS data [3], is still one of the most widely used indicators. Because it compares red and near-infrared (NIR) reflectance, the NDVI is a very useful tool for identifying areas that are vegetation or non-vegetation. Using green and NIR bands to improve open water detection, McFeeters [4] proposed the Normalized Difference Water Index (NDWI) to improve class separability in complex environments.

In order to suppress built-up area interference, Xu et al. later modified the NDWI by substituting shortwave infrared (SWIR) for near-infrared (NIR) [5]. Based on the NIR and SWIR1 bands, Gao's version of NDWI is proposed to directly estimate the water content of vegetation from satellite imagery [6]. According to Jackson et al. [7], these indices are used to map vegetation and water features as well as pre-filter input data for machine learning pipelines, which reduce noise and enhances class separability.

Principal Component Analysis (PCA) is a popular preprocessing step to address the high dimensionality present in multispectral and hyperspectral data. Dimensionality reduction improves computational efficiency and feature selection for classification tasks, as demonstrated by Maćkiewicz et al. [8] and later applied by Ali, UA Md Ehsan and Hossain, [9] in the field of hyperspectral image classification.

The computational simplicity of clustering algorithms, particularly KMeans and its enhanced variant k-means++, has made them popular for unsupervised classification tasks. Arthur and Vassilvitskii [10] proposed k-means++ to improve initialization, which significantly reduces the likelihood of poor local minima. Goud et al. [11] achieved more dependable class separation by applying an improved KMeans variant to hyperspectral data. Ng et al. [12] provided additional evidence for the efficacy of clustering in spectral data.

One of the most popular supervised classification methods in remote sensing is Random Forest (RF), a tree-based ensemble classifier. The usefulness of RF in land cover classification was initially emphasized by Pal [13] because of its resistance to overfitting and low parameter tuning. Belgiu and Drăguţ [14], who emphasized the technological versatility in various data sets and land cover types, provided a comprehensive analysis of the uses of RF in remote sensing. The relevance of RF in vegetation mapping was further demonstrated by Kulkarni et al. [15], who demonstrated how RF could predict forest attributes using multispectral and ancillary data. Because they can use both labeled and unlabeled data, hybrid pipelines that combine supervised classification and unsupervised segmentation are becoming more and more popular. In order to improve classification under class imbalance, Tzu-Tsung Wong et al. [16] proposed a hybrid model that integrates RF and clustering. Xizhen et al. [17] presented a semi-supervised hyperspectral classification framework that significantly improved spatial coherence by utilizing super-pixels and Markov Random Fields. In situations where there is little ground truth, semi-supervised learning has also proven successful. Ziru Yu et al. used Generative Adversarial Networks (GANs) [18] to improve the accuracy of hyperspectral classification and add to the limited training data. Cho et al. [19] made a pseudo-label refinement framework that deals with label noise in places with few resources. In contrast, Willian Paraguassu Amorim et al. [20] presented a semi-supervised CNN model utilizing consistency regularization. To close the gap between labeled and unlabeled data, Zhu et al. [21] also used collaborative representation. These developments have been contextualized within deep learning-based remote sensing by broad surveys like Ma et al. [22] and Hu et al. [23]. The combination of machine learning and spectral indices for better vegetation classification is further supported by comparative studies. To categorize vegetation in mixed landscapes, Sheykhmousa et al. [24] used support vector machine versus Random Forest for remote sensing image classification. The efficacy of decision-tree-based classification in operational mapping was validated by Petropoulos et al. [25] through the use of Landsat data and spectral indices. This was expanded upon by Emmanuel Paradis [26], who used unsupervised classification methods, like KMeans, for large-scale analysis of spectral imaging data.

Despite significant progress in vegetation detection using supervised and unsupervised learning approaches, existing methods still exhibit critical limitations. Supervised classifiers like Random Forests or CNNs require extensive labeled datasets, which are often unavailable, particularly in ecologically diverse or remote regions. On the other hand, methods that do not need supervision, like K-Means clustering, do not have a clear meaning and often have trouble separating classes cor-

rectly, especially when pixels are mixed in frequency. Researchers have suggested hybrid and semi-supervised frameworks, but many of these methods depend on domain-specific priors or deep learning models that need much computing power and manual tuning. Moreover, few studies systematically integrate spectral indices like NDVI and NDWI with machine learning in a scalable, interpretable pipeline for vegetation mapping. There is a noticeable gap in lightweight, label-free frameworks that can combine index-based filtering, clustering, and classification to get high accuracy without much ground truth data. This study fills in the gaps by suggesting a mixed unsupervised-to-supervised framework that uses k-means++ and Random Forest, with NDVI/NDWI-driven vegetation and water masking as guides.

III. Data Collection

This study focuses on the Silchar region in the Barak River basin in southern Assam, Northeast India. The area is a representative site for testing vegetation segmentation techniques due to its rich heritage, vibrant culture, and diverse landscape of vegetation, water bodies, and built-up areas. The following geographic coordinates bound the specific area of interest:

- 24.92051°N, 92.69716°E
- 24.91885°N, 92.89368°E
- 24.74158°N, 92.89336°E
- 24.74058°N, 92.69831°E

The study area is shown in Fig. 1, generated using QGIS software with Google Maps base layer.

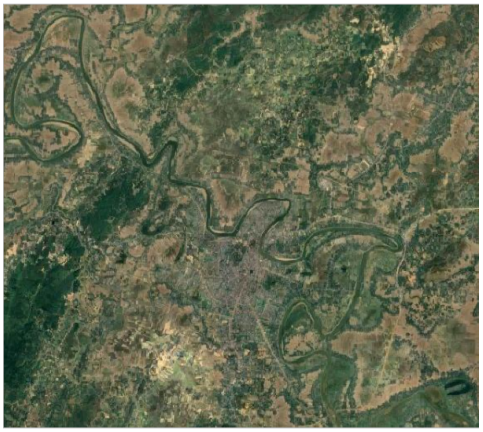

Fig. 1. Geographic extent of the study area in Silchar, Barak Valley.

IV. Methodology

In this section, we have discussed an unsupervised-to-supervised method for vegetation detection using multispectral satellite images. This section includes subsections of satellite data acquisition, vegetation and water index calculation, selection of vegetation area, clustering and dimensionality reduction, and Random Forest-based classification.

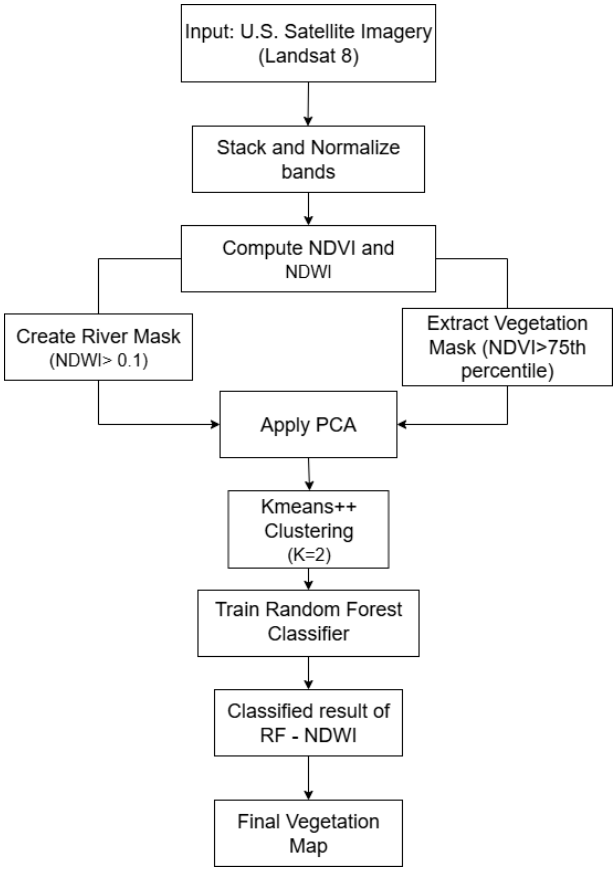

Fig. 2. Proposed methodology using k-means++ and Random Forest

The methodology shown in Fig. 2, outlines a hybrid approach for vegetation detection using multispectral satellite data. The process involves normalizing spectral bands and calculating NDVI and NDWI to identify vegetation and water. PCA is used to reduce dimensions, and k-means++ is used for clustering. Random Forest classifier is used on these labels to predict vegetation, excluding river areas.

Algorithm 1 describes the vegetation detection process in detail. Six spectral bands are normalized to determine vegetation and water regions, and NDVI and NDWI are calculated. Vegetation pixels, high-NDVI, and non-river pixels are chosen, and PCA is used to reduce them. Clustering is done by k-means++, and vegetation is identified using the highest mean NDVI. These pseudo-labels are used to train a Random Forest classifier, which is then applied to the entire image. NDWI is used to refine the final vegetation mask by removing river areas.

A. Satellite Data Acquisition

The multispectral raster data contained a total of six spectral bands: Blue, Green, Red, Near-Infrared [NIR],

Algorithm 1 Vegetation Detection via k-means++ and Random Forest

Input: Multispectral bands: Blue, Green, Red, NIR, SWIR1, SWIR2

Output: Binary vegetation mask M_{final}

1: Stack and reshape input bands into a 2D feature matrix X
2: Normalize X
3: Compute NDVI and NDWI for all pixels
4: Define river mask: NDWI > 0.1
5: Set NDVI threshold T_{NDVI} as 75th percentile
6: Select vegetation pixels where NDVI $> T_{\text{NDVI}}$ and not in river mask
7: Extract corresponding feature vectors X_{veg}
8: Reduce X_{veg} to 4 dimensions using PCA
9: Apply k-means++ clustering with $k = 2$
10: Identify cluster with highest mean NDVI as vegetation
11: Assign binary labels: 1 for vegetation, 0 for non-vegetation
12: Train a Random Forest classifier on X_{veg} and pseudo-labels
13: Predict vegetation class for all pixels in X
14: Reshape predictions into 2D map M_{rf}
15: Remove river pixels from M_{rf} using NDWI mask
16: Output final mask M_{final}

Shortwave Infrared 1 [SWIR1], and Shortwave Infrared 2 [SWIR2]. Each band is loaded and converted into a float array using the Rasterio library. A 6-dimensional feature vector used to represent each pixel in the 3D array, then it is converted into a 2D feature matrix by stacking the spectral bands and normalizing the input features. The satellite data are collected from the U.S. Geological Survey [27], and data from Landsat 8 OLI with cloud cover of 5%, and the bands used are Blue, Green, Red, NIR, SWIR1, and SWIR2.

B. Vegetation and Water Index Calculation

To enhance the spectral separability of land cover classes, two widely used spectral indices, such as NDVI [28] and NDWI [29], are computed and given in equations 1 and 2.

$$\text{NDVI} = \frac{(NIR - Red)}{(NIR + Red)} \tag{1}$$

NDVI values were clipped between $[-1, 1]$ to ensure numerical stability and used to highlight vegetation pixels.

$$\text{NDWI} = \frac{(Green - NIR)}{(Green + NIR)} \tag{2}$$

The pixels with NDWI > 0.1 are heuristically identified as water regions, and subsequently excluded from vegetation index to prevent spectral overlap with water regions.

C. Selection of Vegetation

Pixels having high NDVI values above the 75^{th} percentile are chosen to create a threshold-based mask, calculated thresholds at the 60^{th}, 70^{th}, and 80^{th} percentiles, and it is constantly changing in vegetation classification and accuracy. All pixels that overlapped with the river mask derived from NDWI results are eliminated to refine the output, with the NDWI threshold between 0.0 and 0.2, observing its effect on river mask performance.

D. Clustering and Dimensionality Reduction

The vegetation pixels used in Principal Component Analysis (PCA) are kept, with the top four principal components, in order to minimize computational complexity and noise in high-dimensional spectral data. The k-means++ algorithm is used to perform clustering on vegetation pixels extracted using NDVI thresholding. The k-means++ algorithm is then used to cluster these reduced features into k=2 clusters. We have also experimented with k = 2, 3, and 4 and found that k = 2 yielded the best Dunn Index and DB Score for distinguishing vegetation vs. non-vegetation pixels. The mean NDVI value of each cluster is calculated in order to determine which of the two clusters represents true vegetation. While the other cluster is categorized as background or noise, the one with the higher average NDVI is designated as vegetation. In accordance with this, a binary label mask (0 = non-vegetation, 1 = vegetation) is made.

Algorithm 2 uses the k-means++ initialization strategy to choose the initial cluster centres. This strategy enhances clustering performance by distributing initial centroids according to data distribution.

E. Random Forest Based Classification

The proposed methodology involves the supervised classification of vegetation and non-vegetation pixels using a Random Forest (RF) classifier. k-means++ clustering on PCA-reduced vegetation pixels is used to create the training labels; the cluster with the highest NDVI is used as vegetation. The RF model is trained and tested using an 80:20 split [30] of these pseudo-labeled samples. As explained in Algorithm 3, the trained classifier is then used to predict vegetation labels across the entire dataset.

V. Result and Discussion

In this section, we have implemented the proposed methodology and generated extensive results to validate it. The final vegetation map is accurately

Algorithm 2 k-means++ Clustering Algorithm

Input: Data points $X = \{x_1, x_2, \ldots, x_n\}$, number of clusters k

Output: Initial cluster centers for k-means

Choose the first center c_1 uniformly at random from X
for $i = 2$ to k **do**
  for each $x \in X$ **do**
    Compute $D(x)^2$, the squared distance from x to the nearest center already chosen
  end for
  Choose the next center $c_i = x'$ from X with probability $\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$
end for
Proceed with the standard k-means algorithm using the k initialized centers
repeat
  Assign each point to the nearest center
  Recompute centers as the mean of assigned points
until convergence

Algorithm 3 Random Forest Classification using k-means++

Input: Feature matrix X_{veg}, pseudo-labels y_{veg}

Output: Trained RF model and predicted vegetation labels y_{full}

Split $(X_{\text{veg}}, y_{\text{veg}})$ into training and testing sets (80:20)
Train Random Forest classifier on the training set
Evaluate performance on the test set (e.g., accuracy, precision, recall)
Apply the trained model to the full feature matrix X to predict y_{full}
return Predicted labels y_{full}

detected while effectively excluded the water regions. Visual validation using NDVI and NDWI confirmed the reliability of the proposed methodology in distinguishing vegetation from non-vegetation. The source code for the proposed methodology is available at https://github.com/somcse/An-Unsupervised-to-Supervised-Framework-for-Vegetation-Mapping-Using-Spectral-Indices.

A. Clustering Performance Evaluation

Two internal validation metrics are used to evaluate the quality of unsupervised clustering on NDVI-rich vegetation pixels using k-means++, such as Dunn Index and Davies–Bouldin (DB) Score [31]. The DB score for this experiment is 0.8466, better clustering is indicated by a lower value of DB score. The Dunn Index [32] calculates the ratio of the maximum intra-cluster distance to the minimum inter-cluster distance. The Dunn Index for this study is 0.3142, better clustering is indicated by a higher Dunn score. Table I shows the result of the above scores.

TABLE I
NORMAL RANGES FOR CLUSTERING EVALUATION METRICS

Metric	Ideal Range	Interpretation	Score
Davies-Bouldin Score (DB)	0 to 1 (↓ better)	Lower values indicate better clustering; 0 is optimal.	0.8466
Dunn Index	more than 0.3 (↑ better)	Higher values indicate compact and well-separated clusters.	0.3142

Fig. 3 shows a binary river mask derived from the NDWI, where pixels with NDWI > 0.1 are classified as water (value = 1). This thresholding effectively highlights river networks.

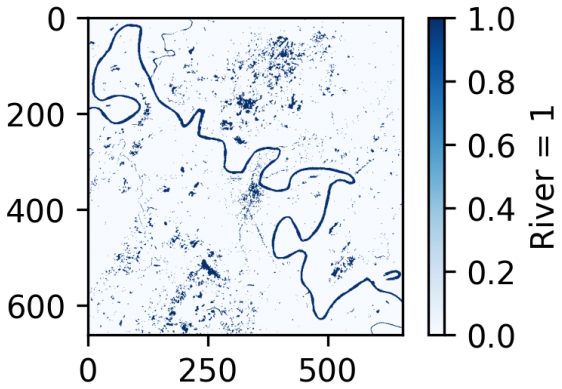

Fig. 3. River mask generated using NDWI with a threshold > 0.1.

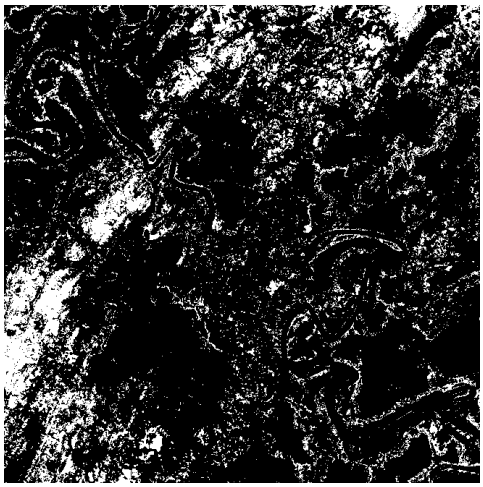

Fig. 4. Binary clustering of vegetation pixels using k-means++.

Fig. 4 presents a binary vegetation map using the k-means++ clustering algorithm, where pixels labeled 1 represent vegetation areas. This unsupervised approach effectively distinguishes vegetation from other land cover types based on spectral patterns.

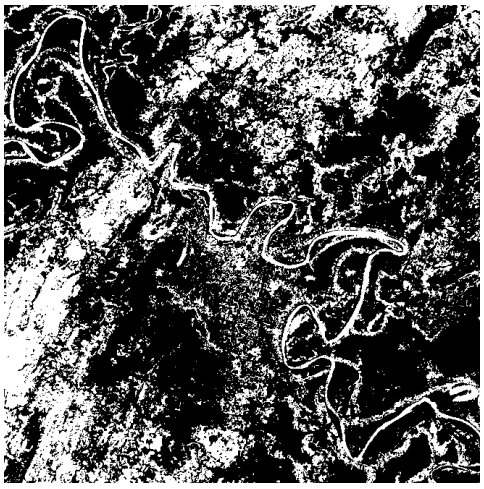

Fig. 5. Vegetation classification map generated using Random Forest.

Fig. 5 shows a vegetation classification map using a Random Forest (RF) classifier. This supervised approach uses spectral features to accurately distinguish vegetation from non-vegetation regions.

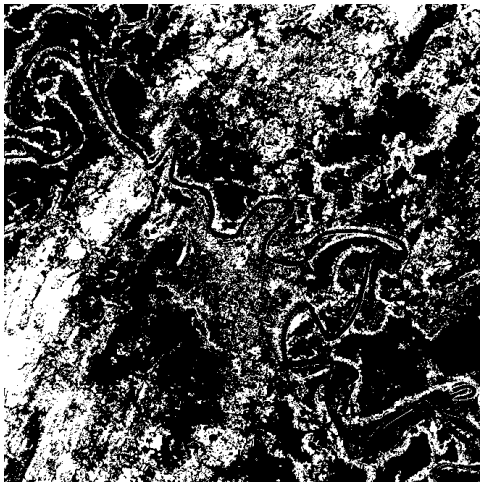

Fig. 6. Final vegetation map, generated after excluding river regions using NDWI-based water masking.

Fig. 6 illustrates the final vegetation map after masking out river regions using NDWI-based water detection.

B. Comparative Analysis of Classification Models

Table II shows results about NDVI Thresholding, Decision Trees, Random Forest (RF), and the proposed k-means++ and RF approach, based on accuracy, F1 score, and ROC-AUC. NDVI Thresholding has 73.2% accuracy and a relatively low F1 score of 0.512. Decision Trees have 78.4% accuracy and an F1 score of 0.592, indicating moderate performance. k-means++ clustering with a Random Forest classifier having accuracy (86.6%), F1 score (0.699), and ROC-AUC (0.921).

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCE ACROSS METHODS

Method	**Accuracy (%)**	**F1 Score**	**ROC-AUC**
NDVI Thresholding	73.2	0.512	0.812
Decision Trees	78.4	0.592	0.834
Supervised RF (random)	80.5	0.624	0.874
Ours (k-means++ + RF)	**86.6**	**0.699**	**0.921**

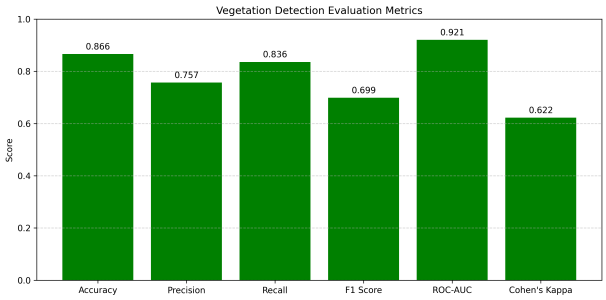

Fig. 7. Validation scores bar plot.

Fig. 7 shows the evaluation metrics, including accuracy, precision, recall, F1 score, ROC-AUC, and Cohen's Kappa, which give an analysis of the performance of the classifier. The plot indicates that a significant percentage of pixels are correctly classified, with an overall accuracy of 0.866. Although precision remained low at 0.757, exposing some false positives, the recall score of 0.836 demonstrated the sensitivity in identifying every vegetation pixel. The mean of precision and recall is reflected in the F1 score of 0.699, which is a reasonable trade-off. Further, the RF ability to differentiate between areas with and without vegetation is demonstrated by its ROC-AUC score of 0.921.

A confusion matrix, Fig. 8, is plotted as a heat map to visualize the classification results and the efficacy of the suggested vegetation identification algorithm. The matrix shows how many pixels are correctly and incorrectly classified in the vegetation and non-vegetation classes. With 67,456 true positives and 11762 false negatives, an 83.6% recall is achieved. The 58,189 false positives, which show that some non-vegetation pixels were incorrectly classified as vegetation, has a minor effect on the precision. This study highlights areas that need improvement, such as reducing vegetation, while confirming the outstanding recall and overall performance.

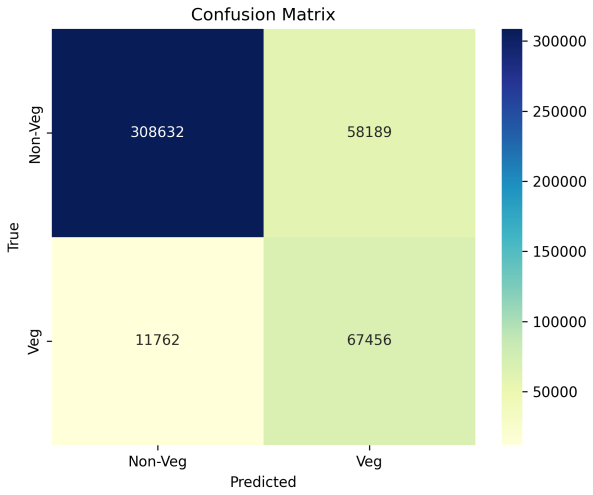

Fig. 8. Confusion matrix showing predicted vs. true labels for vegetation classification.

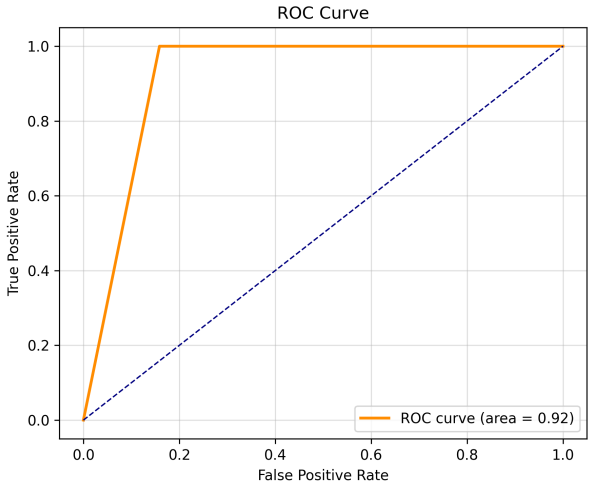

Fig. 9. ROC curve illustrating classification performance.

A Receiver Operating Characteristic (ROC), curve as in Fig. 9 is generated to show the graphical representation of vegetation classification. The ROC curve shows the trade-off between the true positive rate and false positive rate at various threshold values. The Random Forest shows a high rise towards the top-left corner, indicating strong sensitivity with few false alarms. The ability to differentiate between vegetation and non-vegetation classes is shown by the Area Under the Curve (AUC), which is 0.9207. The high AUC value shows that the unsupervised-to-supervised hybrid technique is flexible and effective at identifying vegetation regions from multispectral satellite images.

VI. Conclusion and Future Scope

This study proposes an efficient hybrid approach for vegetation detection by combining unsupervised k-means++ clustering with Random Forest classification. NDVI and NDWI indices are used to distinguish between vegetation and water regions in order to enhance discrimination near river bodies. PCA is used to reduce the dimensionality of the multispectral imagery, and k-means++ is applied for clustering. Also, the clustering quality is validated by validation metrics DB and Dunn. The results are used to train the RF model, and the classification results accuracy is validated by several validation metrics such as accuracy, precision, F1 score, recall, kappa score, ROC-AUC, and confusion matrix.

The proposed framework suggests a robust and label-free method for vegetation detection. In the future, we will add different machine learning and deep learning models for better clustering and classification.

Acknowledgment

This work was supported by the sponsored project under the **Department of Science and Technology (DST), Govt. of India - Tribal Sub-Plan (TSP)** scheme of IIT Bhilai Innovation and Technology Foundation (IBITF), **Sanction Number:** IBITF/Note/TSP/SanctionLetter/2024-25/0125.

References

[1] A. Essaadia, A. Abdellah, A. Ahmed, F. Abdelouahed, and E. Kamal, "The normalized difference vegetation index (ndvi) of the zat valley, marrakech: comparison and dynamics," *Heliyon*, vol. 8, no. 12, p. e12204, 2022.

[2] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sensing of Environment*, vol. 8, no. 2, pp. 127–150, 1979.

[3] J. Rouse Jr, R. Haas, J. Schell, and D. Deering, "Monitoring vegetation systems in the great plains with erts. third earth resources technology satellite-1 symposium: Volume 1; technical presentations, section b, sc freden, ep mercanti, and ma becker, eds., nasa special publ," *InConference Proceedings, Document ID: 19740022592*, pp. 1–9, 1974.

[4] S. K. McFeeters, "The use of the normalized difference water index (ndwi) in the delineation of open water features," *International journal of remote sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[5] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *International Journal of Remote Sensing*, vol. 27, no. 14, pp. 3025–3033, Jul. 2006.

[6] B.-C. Gao, "Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment*, vol. 58, no. 3, pp. 257–266, 1996.

[7] T. J. Jackson, D. Chen, M. Cosh, F. Li, M. Anderson, C. Walthall, P. Doriaswamy, and E. R. Hunt, "Vegetation water content mapping using landsat data derived normalized difference water index for corn and soybeans," *Remote Sensing of Environment*, vol. 92, no. 4, pp. 475–482, 2004.

[8] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers and Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[9] U. M. E. Ali, M. A. Hossain, and M. R. Islam, "Analysis of pca based feature extraction methods for classification of hyperspectral image," in *2019 2nd international conference on innovation in engineering and technology (ICIET)*. IEEE, 2019, pp. 1–6.

[10] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, ser. SODA '07. USA: Society for Industrial and Applied Mathematics, 2007, p. 1027–1035.

[11] O. S. C. Goud, T. H. Sarma, and C. S. Bindu, "Improved k-means clustering algorithm for band selection in hyperspectral images," in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 2023, pp. 1–6.

[12] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[13] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[14] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24–31, 2016.

[15] A. D. Kulkarni and B. Lowe, "Random forest algorithm for land cover classification," *International Journal on Recent and Innovation Trends in Computing and Communication*, 2016.

[16] T.-T. Wong, N.-Y. Yang, and G.-H. Chen, "Hybrid classification algorithms based on instance filtering," *Information Sciences*, vol. 520, pp. 445–455, 2020.

[17] X. Han, Z. Jiang, Y. Liu, J. Zhao, Q. Sun, and J. Liu, "Semi-supervised hyperspectral image classification based on multiscale spectral–spatial graph attention network," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[18] Z. Yu and W. Cui, "Robust hyperspectral image classification using generative adversarial networks," *Information Sciences*, vol. 666, p. 120452, 2024.

[19] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7298–7308.

[20] W. P. Amorim, E. C. Tetila, H. Pistori, and J. P. Papa, "Semi-supervised learning with convolutional neural networks for uav images automatic recognition," *Computers and Electronics in Agriculture*, vol. 164, p. 104932, 2019.

[21] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *MLG Cambrideg*, pp. CMU–CALD–02–107, 07 2003.

[22] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.

[23] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high- resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[24] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308–6325, 2020.

[25] G. P. Petropoulos, C. Kontoes, and I. Keramitsoglou, "Burnt area delineation from a uni-temporal perspective based on landsat tm imagery classification using support vector machines," *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 1, pp. 70–80, 2011.

[26] E. Paradis, "Probabilistic unsupervised classification for large-scale analysis of spectral imaging data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 107, p. 102675, 2022.

[27] U.S. Geological Survey, "Earthexplorer," n.d., accessed: 2025-05-08. [Online]. Available: https://earthexplorer.usgs.gov/

[28] G. M. Gandhi, S. Parthiban, N. Thummalu, and A. Christy, "Ndvi: Vegetation change detection using remote sensing and gis – a case study of vellore district," *Procedia Computer Science*, vol. 57, pp. 1199–1210, 2015, 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).

[29] S. P. Pattanayak and S. K. Diwakar, "Seasonal comparative study of ndvi, ndbi and ndwi of hyderabad city (telangana) based on liss-iii image using remote sensing and dip," *Khoj: An International Peer Reviewed Journal of Geography*, vol. 5, no. 1, pp. 78–86, 2018.

[30] K. Wang, Z. Ai, W. Zhao, Q. Fu, and A. Zhou, "A hybrid model for predicting low oxygen in the return air corner of shallow coal seams using random forests and genetic algorithm," *Applied Sciences*, vol. 13, no. 4, 2023.

[31] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[32] A. Bhadana and M. Singh, "Fusion of k-means algorithm with dunn's index for improved clustering," in *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, 2017, pp. 1–5.