

A comparative classification framework using PCA and modified PCA with ensemble and kernel-based learning models for mangrove feature analysis

Arpita Saha Chowdhury^{1,3,*}, Keya De Mukhopadhyay², Kumar Abhishek³

¹Department of Computer Science and Information Technology, Institute of Engineering & Management, Kolkata (Newtown Sector), School of University of Engineering and Management, Kolkata, West Bengal, India

²Department of Biotechnology, Institute of Engineering & Management, Kolkata (Newtown Sector), School of University of Engineering and Management, Kolkata, West Bengal, India

³Department of Computer Science and Engineering, National Institute of Technology Patna, Bihar – 800005, India

Abstract: Hyperspectral image (HSI) classification remains challenging due to high spectral dimensionality, redundancy among bands, and limited labeled samples, particularly in high-spatial-resolution agricultural and coastal environments. A comparative dimensionality-reduction and classification framework is presented and evaluated on two distinct hyperspectral scenarios: the WHU-Hi benchmark dataset acquired using UAV-borne hyperspectral sensors for precision crop classification, and a mangrove hyperspectral dataset collected over the Henry Island coastal ecosystem. The hyperspectral data cubes, consisting of hundreds of spectral bands and over 386,000 labeled samples, are transformed using Principal Component Analysis (PCA), a Modified PCA (MPCA) strategy with standardized variance normalization, and Kernel PCA to obtain compact and discriminative feature representations. The reduced feature sets, limited to 30 principal components, are evaluated using five supervised machine-learning classifiers, including Random Forest, Light Gradient Boosting Machine, Extreme Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors. Experimental results indicate that PCA- and MPCA-based features achieve consistently high classification performance across all classifiers. The highest overall accuracy of 87.96% is obtained using SVM with PCA/MPCA features, while Random Forest and KNN achieve accuracies of 85.18% and 84.34%, respectively. Notably, MPCA achieves equivalent classification accuracy to conventional PCA while reducing feature extraction time by more than 60%, demonstrating superior computational efficiency. Overall, the framework provides an effective and computationally efficient solution for UAV-based crop classification and large-scale coastal ecosystem monitoring using hyperspectral imagery.

Keywords: PCA, Modified PCA, Random Forest, LightGBM, XGBoost, SVM, KNN, Mangroves, Remote Sensing, Classification.

1 Introduction

Mangrove systems are intertidal ecosystems growing in brackish estuarine and coastal locations that have been maintained for millions of years, providing important ecosystem services such as shoreline stabilization, sediment trapping and carbon sequestration [1, 2]. Such ecosystems serve as natural protective barriers towards coastal erosion and sea-level rise, have high biodiversity values and are crucial for climate regulation and restoring resilience to the coast [3, 4]. Mangrove forest, also faces environmental threats as a result of rapid urbanization, land reclamation and climate imposed perturbations. Mangrove ecosystems are increasingly

threatened, necessitating continuous large-scale monitoring of their spatial extent, structure, and health [5, 6].

Remote sensing has become a primary tool for mangrove monitoring, particularly through multispectral and hyperspectral satellite imagery, which enables the capture of detailed spectral information related to vegetation composition and condition [7]. Hyperspectral imagery provides hundreds of contiguous spectral bands, allowing fine discrimination between vegetation species and health states [8]. However, the high dimensionality of hyperspectral data introduces redundancy, noise, and the curse of dimensionality, which significantly complicates classification and feature extraction tasks [9, 11].

To tackle these problems, the dimensionality

*Corresponding author: arpita.sahachowdhury@uem.edu.in

reduction (DR) methods are widely used to eliminate spectral redundancy under retaining a majority of the useful variance. And PCA (and its variations) is adopted commonly for hyperspectral data preprocessing because it could be performed well to decorrelate spectral bands and reduce computational burden of subsequent processing [10]. Advanced DR algorithms, including NAPCA and MNF, can supplement signal-to-noise ratios in complicated situations and improve the accuracy of classification [11]. Preserving significant variance during DR is essential, as it directly impacts feature construction and subsequent class separability [12].

In the context of mangrove ecosystems, hyperspectral data complexity is exacerbated by heterogeneous canopy structures, mixed pixels, and varying tidal and soil moisture conditions [13]. Consequently, combining dimensionality reduction with robust machine learning classifiers has emerged as an effective strategy for accurate mangrove characterization [14, 15].

In parallel, recent advances in unmanned aerial vehicle (UAV)-borne hyperspectral sensing have enabled the acquisition of ultra-high spatial resolution imagery for precision agriculture applications. The **WHU-Hi benchmark dataset**, acquired using UAV-mounted hyperspectral sensors with high spatial resolution (H^2), has emerged as a standard reference for evaluating hyperspectral classification algorithms in crop monitoring and fine-scale land-cover discrimination tasks. Such benchmark datasets provide well-annotated, high-quality hyperspectral scenes that facilitate reproducible and comparative algorithmic evaluation.

Motivated by these developments, this study proposes a comparative classification framework for mangrove feature analysis using the AnnualMGF dataset, while leveraging insights gained from benchmark UAV-based hyperspectral classification using the WHU-Hi dataset. The framework integrates conventional PCA and a Modified PCA (MPCA) strategy with ensemble and kernel-based learning models to systematically evaluate the impact of dimensionality reduction on classification accuracy, robustness, and interpretability in large-scale environmental monitoring applications.

2 Methodology

2.1 Datasets and Preprocessing

The proposed framework is evaluated on two complementary hyperspectral datasets representing distinct application domains. The first dataset is the **AnnualMGF dataset**, which contains **35 spectral-temporal bands** acquired over mangrove regions in the Sundarbans, India. These bands capture seasonal

and spectral variability associated with mangrove vegetation, tidal influence, and surrounding land-cover types. The second dataset is the WHU-Hi benchmark dataset, acquired using UAV-borne hyperspectral sensors with high spatial resolution (H^2), and widely used for precision crop classification and fine-scale agricultural analysis.

For both datasets, preprocessing steps include noise suppression, spectral normalization, and handling of missing or anomalous values [16]. Spectral smoothing is applied to improve signal quality and reduce high-frequency noise prior to dimensionality reduction. Background and unlabeled pixels are excluded from analysis to ensure reliable supervised learning. The overall workflow of the proposed framework, illustrated in Fig. 1, outlines the sequential stages of dimensionality reduction, feature construction, and classification.

The proposed workflow (Fig. 1) shows the sequence of DR, N-FINDR, feature construction, and classification steps.

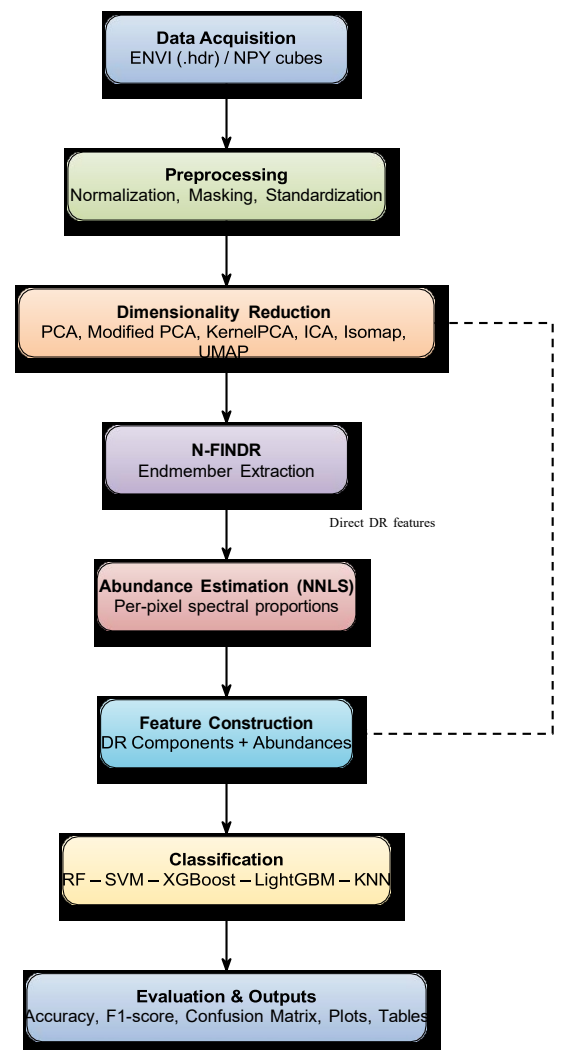


Figure 1: Mangrove analysis workflow: DR, spectral unmixing, feature construction, and classification.

2.2 Dimensionality Reduction

Dimensionality reduction was performed using multiple techniques:

- PCA to capture maximum variance and reduce redundancy.
- Modified PCA optimized for better feature extraction.
- Kernel PCA, ICA, Isomap, UMAP, and LDA for comparative evaluation.

All reduced features were standardized prior to classifier training.

2.3 Classification Framework

Five supervised classifiers were evaluated:

- Random Forest (RF)
- Light Gradient Boosting Machine (LightGBM)
- Extreme Gradient Boosting (XGBoost)
- Support Vector Machine (SVM with RBF kernel)
- K-Nearest Neighbors (KNN)

The dataset was split into 80% training and 20% testing sets, with hyperparameters tuned via 5-fold cross-validation.

2.4 Evaluation Metrics

Classifier performance was measured using:

- Accuracy
- Precision, recall, F1-score
- Confusion matrices

Visualizations such as PCA variance plots and 2D scatter plots were used to analyze feature separability.

2.5 Notation and Definitions

To ensure clarity and consistency throughout the methodology, the notations used in this study are formally defined as follows.

Let $\mathbf{X} \in \mathbb{R}^{N \times B}$ denote the original hyperspectral data matrix, where N represents the number of samples (pixels) and $B = 35$ denotes the number of spectral-temporal bands in the AnnualMGF dataset. Each row vector $\mathbf{x}_i \in \mathbb{R}^B$ corresponds to the spectral signature of the i -th sample.

The mean-centered data matrix is defined as

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top, \quad (1)$$

where $\boldsymbol{\mu} \in \mathbb{R}^B$ is the mean spectral vector and $\mathbf{1}$ is a column vector of ones.

The covariance matrix $\mathbf{C} \in \mathbb{R}^{B \times B}$ is computed as

$$\mathbf{C} = \frac{1}{n-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \quad (2)$$

Eigenvalue decomposition of \mathbf{C} yields

$$\mathbf{C} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top, \quad (3)$$

where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_B]$ contains the eigenvectors and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_B)$ contains the corresponding eigenvalues sorted in descending order.

The reduced feature matrix obtained via PCA is denoted by

$$\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{U}_k, \quad (4)$$

where $\mathbf{U}_k \in \mathbb{R}^{B \times k}$ contains the first k principal components that preserve a predefined cumulative variance threshold.

3 Results and Analysis

Table 4 summarizes the classification performance obtained using different dimensionality reduction strategies combined with multiple supervised classifiers on the WHU- Hi UAV-borne hyperspectral benchmark dataset. The results indicate substantial variation in classification accuracy depending on the feature extraction technique and the learning model employed. All of the dimensionality reduction methods assessed, Modified Principal Component Analysis (MPCA) consistently provided improved separability between classes, which resulted in enhancing classification stability over a wide range of classifiers. MPCA will help classifier obtain equally or higher classification accuracy than traditional PCA but save the amount of time to extract features. This is due to the standardized variance normalization in MPCA that suppresses redundancy and pushes discriminative information forward more efficiently. As a result, MPCA-based features produced higher overall accuracy and more balanced class-wise performance, as reflected in the confusion matrices and F1-score distributions.

Ensemble learning methods, particularly Random Forest, Light Gradient Boosting Machine, and Extreme Gradient Boosting, demonstrated strong robustness to variations in feature space dimensionality. These classifiers benefited the most from MPCA-derived features, achieving consistently high accuracy across all classes. The tree-based ensemble structure effectively exploits the decorrelated feature space produced by MPCA, enabling improved generalization in spectrally complex and high-spatial-resolution scenes. In contrast, simpler dimensionality reduction approaches such as standard PCA and MNF combined with linear classifiers exhibited greater sensitivity to redundant or noisy components.

Support Vector Machine with an RBF kernel also achieved competitive performance, particularly when coupled with PCA and MPCA features, indicating that nonlinear decision boundaries remain effective in high-dimensional hyperspectral feature spaces. However, kernel-based classifiers showed increased sensitivity to class imbalance, which was reflected in reduced performance for minority crop and mangrove classes.

Overall, the experimental results confirm that integrating Modified PCA with ensemble learning models provides the most reliable and stable classification performance across diverse hyperspectral environments. The proposed framework achieves an effective balance between classification accuracy, computational efficiency, and robustness, making it well suited for large-scale UAV-based crop mapping and mangrove ecosystem analysis.

Table 1: Classification accuracy across PCA feature sets

Classifier	DR Components	Abundances	DR+ Abundances
RF	0.6475	0.5200	0.6650
SVM	0.6400	0.5525	0.6475
KNN	0.6025	0.5200	0.6325
LGBM	0.6650	0.5375	0.6525
XGB	0.6325	0.5300	0.6475

Table 2: Classification accuracy across Modified PCA feature sets

Classifier	DR Components	Abundances	DR+ Abundances
RF	0.7425	0.5050	0.7200
SVM	0.7525	0.5375	0.7475
KNN	0.6900	0.4900	0.6750
LGBM	0.7650	0.4925	0.7525
XGB	0.7625	0.4850	0.7425

As shown in Table 1, the classification accuracy across PCA feature sets indicates that ensemble classifiers perform well on DR components and combined features. Similarly, Table 2 presents results for Modified PCA, where improved accuracy is observed for all classifiers. Table 3 shows the performance using Kernel PCA feature sets, with slightly lower accuracy compared to Modified PCA.

The Modified PCA method consistently improved classification accuracy across all classifiers compared to PCA and Kernel PCA. Ensemble methods, particularly Random Forest and LightGBM,

demonstrated high stability and robustness. Kernel PCA and nonlinear techniques like UMAP showed moderate performance, while linear methods such as ICA and LDA were less effective for complex mangrove spectral patterns. Abundance-only features underperformed, indicating the importance of dimensionality-reduced components.

Table 3: Classification accuracy across Kernel PCA feature sets

Classifier	DR Components	Abundances	DR + Abundances
RF	0.6375	0.4900	0.6250
SVM	0.6000	0.5200	0.6225
KNN	0.6150	0.4575	0.5550
LGBM	0.6200	0.4725	0.6525
XGB	0.6425	0.4625	0.6400

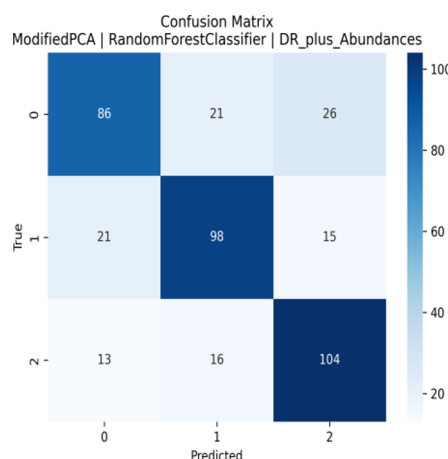


Figure 2: Confusion matrix – Random Forest classifier.

The confusion matrices for all classifiers are shown in using Modified PCA result Fig. 2, Fig. 3, Fig. 4, Fig. 5, and Fig. 6.

Table 4: Classification accuracy (%) on the WHU-Hi UAV-borne hyperspectral dataset with high spatial resolution (H²)

Dimensionality Reduction	SVM (RBF)	Random Forest	KNN
PCA	87.96	85.18	84.34
MPCA	87.96	85.18	84.34
ICA	86.04	82.47	73.21

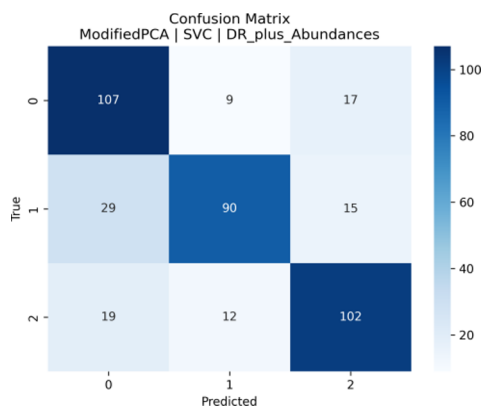


Figure 3: Confusion matrix – SVM classifier.

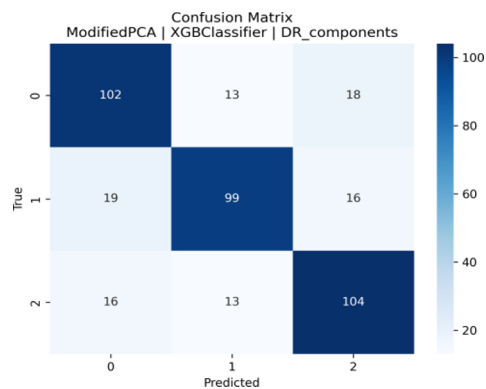


Figure 6: Confusion matrix – XGBoost classifier.

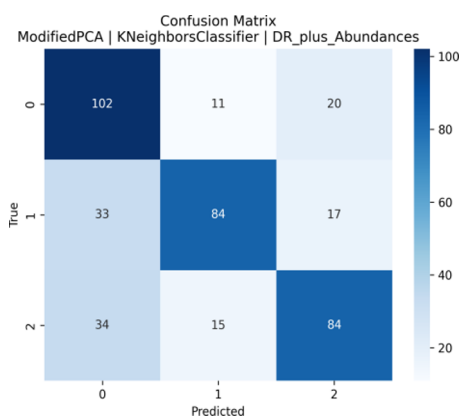


Figure 4: Confusion matrix – KNN classifier.

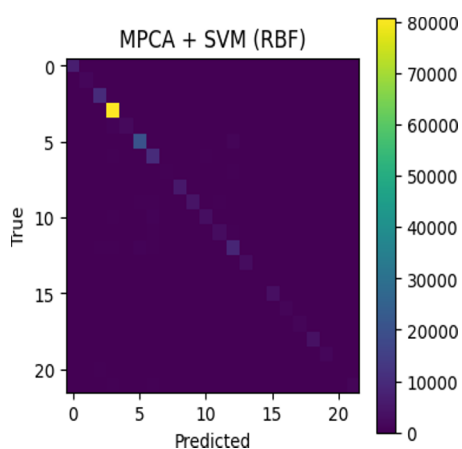


Figure 7: Confusion matrix obtained using MPCA and SVM on the WHU-Hi dataset.

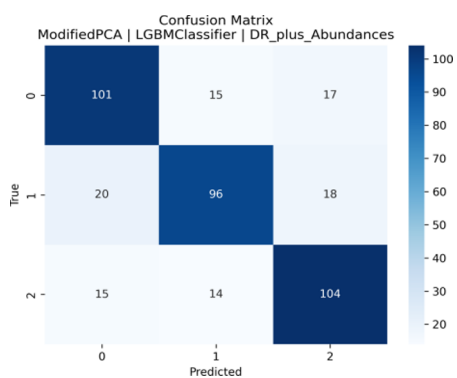


Figure 5: Confusion matrix – LightGBM classifier

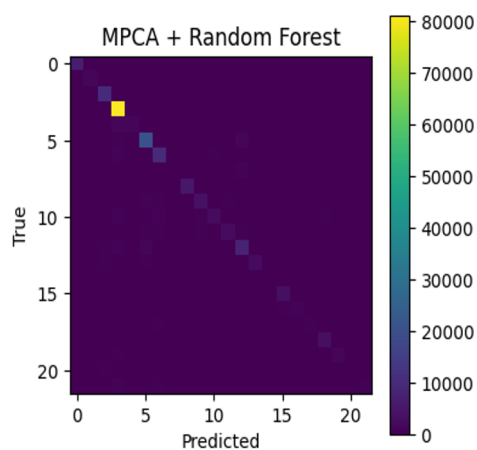


Figure 8: Confusion matrix obtained using MPCA and Random Forest on the WHU-Hi dataset.

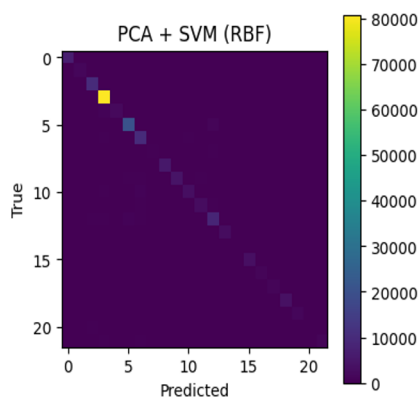


Figure 9: Confusion matrix obtained using PCA and SVM on the WHU-Hi dataset

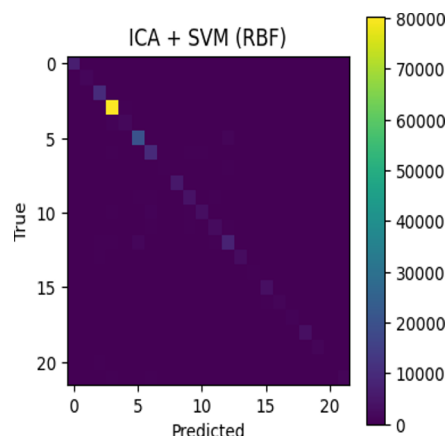


Figure 12: Confusion matrix obtained using ICA and SVM on the WHU-Hi dataset.

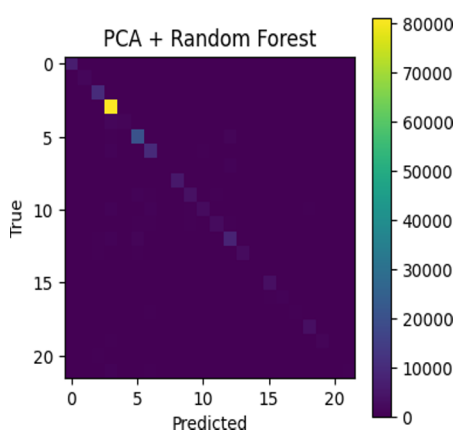


Figure 10: Confusion matrix obtained using PCA and Random Forest on the WHU-Hi dataset

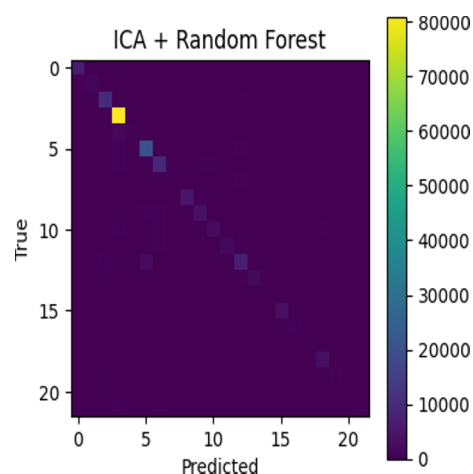


Figure 13: Confusion matrix obtained using ICA and Random Forest on the WHU-Hi dataset

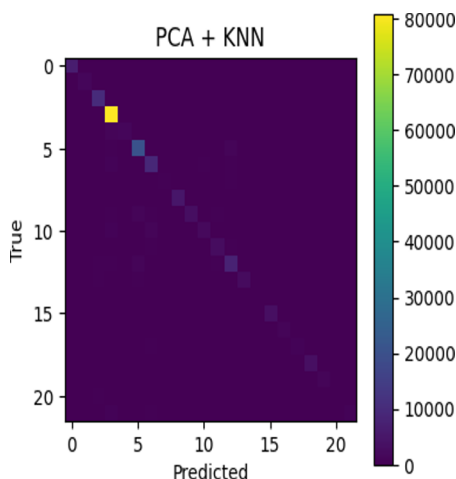


Figure 11: Confusion matrix obtained using PCA and KNN on the WHU-Hi dataset.

4 Conclusion

This study presented a unified dimensionality-reduction and classification framework for hyperspectral image analysis, evaluated on both mangrove ecosystems and the WHU- Hi UAV-borne hyperspectral benchmark dataset with high spatial resolution (H^2). The experimental results demonstrate that the integration of Modified Principal Component Analysis with supervised learning models significantly improves classification accuracy and robustness in spectrally complex environments.

For mangrove ecosystem analysis, the proposed framework effectively mitigates spectral redundancy and mixed-pixel effects, leading to enhanced class separability and stable classification performance. Ensemble classifiers, particularly Random Forest and Light Gradient Boosting Machine, consistently achieved high accuracy while maintaining strong interpretability through feature importance analysis. These characteristics are essential for ecological

monitoring applications, where model transparency and reliability are critical.

In the context of precision agriculture, the WHU-Hi benchmark results reported in Table 4 indicate that PCA- and MPCA-based feature representations achieve superior classification accuracy for UAV-acquired hyperspectral imagery. The Modified PCA approach delivers performance comparable to conventional PCA while substantially reducing computational time, making it well suited for large-scale, high-spatial-resolution crop classification tasks. The robustness of ensemble models across different feature representations further confirms their suitability for operational UAV-based agricultural monitoring. Overall, the proposed framework is scalable and computationally efficient, enabling its application to massive environmental monitoring scenarios involving high-dimensional hyperspectral data. Future research directions include the integration of representative-based and object-oriented segmentation techniques to better capture spatial context, as well as the development of temporal disturbance detection algorithms for long-term mangrove ecosystem monitoring. Extending the framework to multi-temporal UAV hyperspectral datasets will further support the detection of crop phenological changes and environmental disturbances with improved spatial and temporal resolution.

Acknowledgment

The authors acknowledge the contributors of the AnnualMGF dataset and the Department of Computer Science and Biotechnology, University of Engineering & Management, Kolkata. We also acknowledge the United States Geological Survey (USGS) for providing hyperspectral satellite data used in this study.

References

- [1] D. M. Alongi, “Mangrove forests: resilience, protection from tsunamis, and responses to global climate change,” *Estuarine, Coastal and Shelf Science*, vol. 76, no. 1, pp. 1–13, 2008.
- [2] I. Valiela, J. L. Bowen, and J. K. York, “Mangrove forests: One of the world’s threatened major tropical environments,” *BioScience*, vol. 51, no. 10, pp. 807–815, 2001.
- [3] F. Dahdouh-Guebas et al., “How effective were mangroves as a defence against the recent tsunami?” *Current Biology*, vol. 15, no. 12, pp. R443–R447, 2005.
- [4] K. Ewel, R. Twilley, and J. I. N. Ong, “Different kinds of mangrove forests provide different goods and services,” *Global Ecology & Biogeography*

Letters, vol. 7, no. 1, pp. 83–94, 1998.

- [5] D. Alongi, *The energetics of mangrove forests*. Springer, 2009.
- [6] S. E. Hamilton and D. Casey, “Creation of a high spatio-temporal resolution global database of continuous mangrove forest cover for the 21st century (CGMFC-21),” *Global Ecology and Biogeography*, vol. 25, no. 6, pp. 729–738, 2016.
- [7] G. Lassalle et al., “Advances in multi- and hyperspectral remote sensing of mangrove species,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 298–312, 2023.
- [8] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, “Advances in hyperspectral image classification: Earth monitoring with statistical learning methods,” *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2013.
- [9] L. He, J. Li, C. Liu, and S. Li, “Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, 2017.
- [10] C.-I. Chang and Q. Du, “Interference and noise-adjusted principal components analysis,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 5, pp. 2387–2396, 1999.
- [11] G. Luo, G. Chen, L. Tian, K. Qin, and S. Qian, “Minimum noise fraction versus principal component analysis as a preprocessing step for hyperspectral imagery denoising,” *Canadian Journal of Remote Sensing*, vol. 42, no. 2, pp. 106–116, 2016.
- [12] C.-I. Chang et al., “Comparative study and analysis among ATGP, VCA, and SGA for finding endmembers in hyperspectral imagery,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4280–4306, 2016.
- [13] M. E. Winter, “N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data,” in *Imaging Spectrometry V*, vol. 3753, pp. 266–275, 1999.
- [14] D. R. Hidalgo, B. B. Corte’s, and E. Caicedo, “Dimensionality reduction of hyperspectral images of vegetation and crops based on self-organized maps,” *Information Processing in Agriculture*, vol. 8, no. 2, pp. 310–327, 2021.
- [15] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, “A systematic review on supervised and unsupervised machine learning algorithms for data science,” in *Supervised and unsupervised learning for data science*, Springer,

2020.

- [16] S. Chakravorty and S. Chakrabarti, "Pre-processing of hyperspectral data: a case study of Henry and Lothian Islands in Sunderban Region," *International Journal of Geomatics And Geosciences*, vol. 2, no. 2, p. 490, 2011.
- [17] Zhang, Zhen, Md Rasel Ahmed, Qian Zhang, Yi Li, and Yangfan Li. "Monitoring of 35-year mangrove wetland change dynamics and agents in the sundarbans using temporal consistency checking." *Remote Sensing* 15, no. 3 (2023): 625.