INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

A Novel Algorithm for Efficient Hyperspectral Image Preprocessing

Ranjana Gore[1]*, **Pramod Bhalerao**[2], **Shweta Yadhav**[1], **Abhishek Das**[1]

1 CSE, MIT Art, Design and Technology University, Pune, Maharashtra, India
2 MGM's JNEC, Chhatrapati Sambhajinagar, Maharashtra, India

OPEN ACCESS

Received: 28/03/2025

Accepted: 28/06/2025

Published: 21/07/2025

Citation: Gore R, Bhalerao P, Yadhav S, Das A (2025) A Novel Algorithm for Efficient Hyperspectral Image Preprocessing. Indian Journal of Science and Technology 18(27): 2221-2226. https://doi.org/10.17485/IJST/v18i27.577

* **Corresponding author**.

goreranjana123@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The main objective of this work is to develop an efficient preprocessing algorithm for classifying image bands of Hyperspectral sensor. GUI facilitate preprocessing for major group of researchers in this domain. **Methods:** During image acquisition using a Hyperion sensor, certain bands may not be captured properly, these are bad bands. These bad bands pose a significant hurdle in spectral analysis. A novel algorithm for identifying and classifying bad bands has been designed using statistical analysis of image bands. The machine learning classifiers the Hybrid K-nearest neighbor and Clustering techniques are used to develop the model and improve the accuracy of the spectral classification. **Findings:** The proposed algorithm effectively identified and classified bad bands. The Hybrid K-nearest neighbor classifier delivered 100% accuracy whereas the decision tree classifier shows 98% accuracy. The developed graphical interface significantly improved usability and efficiency in hyperspectral image preprocessing. The proposed work identified two significant thresholds 0.5 and 10.5 for bad band identification. **Novelty:** This work gives new benchmark for hyperspectral image preprocessing with highest accuracy.

Keywords: Bad Band Identification; Hyperion; Hyperspectral; Preprocessing; Remote Sensing

1 Introduction

Remote sensing is the efficient technology which uses electromagnetic energy for surface analysis. Remote sensing images are broadly classified into two types of remote sensing images viz. Multispectral and hyperspectral. The hyperspectral images are distinguished by their inclusion of hundreds of bands, proving it beneficial for precise mineral mapping and identification. The hyperspectral images have been used for mineral mapping, LULC mapping [(1)] and classification [(2)], lithological discrimination [(3)] in various studies and Mineral mapping [(4)]. Hyperion is one of the Earth Observing-1 satellites. Hyperion is a hyperspectral sensor on board an EO-1 satellite with 242 spectral bands having a spectral range of 400nm to 2500nm. The hyperspectral images need to be preprocessed for bad band removal, radiometric correction, geometric correction, noise removal, and dimensionality reduction to obtain accurate results in spectral analysis.

Hyperspectral remote sensing plays an important role in mineral mapping and identification activities. It can be used to identify minerals with features of reflectance spectra. Minerals and rocks display analytic spectral characteristics throughout the electromagnetic spectrum to map the mineral abundance. The hyperspectral remote sensing can finely identify the minerals as compared to the multispectral remote sensing, also it is possible to construct the spectrum for each and every pixel from the hyperspectral image. This spectrum can be used for target discrimination, classification and detection

The hyperspectral images of Hyperion sensor have 242 bands. From these bands some have information, and some may lack useful information. This happens due to the nonfunctioning of CCD arrays of sensors during image acquisition. Generally, the bad bands containing no information are bands 1 to 7, bands 58 to 76, and bands 225 to 242. Bands 122 to 132, bands 165 to 182, and bands 221 to 224 are also bad as they contain water vapor contents, but this varies. The bad bands may differ for every image as it depends upon the scanning time and the location of acquisition. These noninformative bands hampers the quality of image leading to difficulties in further processing and spectral analysis of these images. It also reduces the accuracy of spectral classification and affects spectral analysis significantly. So, it is utmost required to identify the bad bands and skip those from further analysis.

Gore worked on hyperspectral data which handled for bad bands [5] which improved the accuracy of spectral classification and mineral identification. Kowkabi and et al developed a statistical preprocessing algorithm [6] which used AVIRIS Indian Pines and HYDICE urban datasets. In other approach, a novel graph-regularized tensor ring decomposition algorithm was developed for hyperspectral images. It demonstrated improved reconstruction quality with spectral retention [7]. The researchers used MA-FE method on three public remote sensing datasets and proved its effectiveness with accuracy of 99%, 95% and 93% [8] in remote sensing image scene classification respectively. But still, this method needs to be improved for the generalization of large datasets. Kumar and et. al. [9] developed a band selection using dual partitioning where informative bands are identified based on the ranking. The CNN techniques used, and this framework is used for Salinas, Indian Pines and KSC datasets where overall accuracy is 99%,99% and 97% respectively. Cozolino et al. [10] presented the overview of different preprocessing techniques implemented for hyperspectral imaging data. The techniques discussed lack integration into a user-friendly platform and less accessible to non-experts.

The current work proposes a method that combines bad band removal, band selection and usability enhancements into a single, novel and integrated framework which automates the bad band identification and spectral classification. The proposed methodology is rigorously designed to accurately identify and select the informative bands for further analysis. The proposed algorithm gives a straightforward, threshold-based classification which simplifies the preprocessing of hyperspectral images without compromising the accuracy. Its integration into a graphical interface enhanced usability by making it accessible to the broader range of researchers.

2 Methodology

The process begins with the retrieval of hyperspectral data acquired by the Hyperion sensor, which is obtained from the USGS website https://glovis.usgs.gov. Our proposed methodology entails an in-depth examination of the individual image bands, total 2420 images, aiming to create a dedicated preprocessing software module. This module's main purpose is to effectively identify and flag any bad bands within the dataset, ensuring the quality and accuracy. The diagram for methodology is shown in Figure 1.

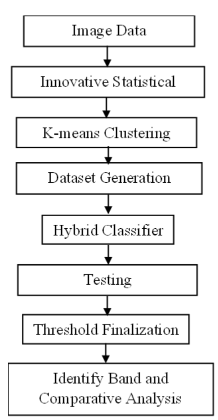

Fig 1. Methodology

> There are 242 spectral bands ranging from 400 to 2500nm. The product is L1R with dimensions 256 X 242 X 3128. The study Area for which images are considered is the Lonar Crater of district Buldhana, Maharashtra. The Latitude and Longitude are 19°58'33.1"N,76°30'45.3"E respectively. The date of acquisition of satellite data was 10-02-2016, 16-02-2017.

Statistical analysis of image bands has been performed extensively for the construction of a dataset from the image dataset. Mean and standard deviation are used for analyzing the quality of the images. So, the relevant features are extracted from the remotely sensed data, analyzed, and optimized for further analysis using machine learning. K-means Clustering can help to identify the patterns or similarities within the different band images of the dataset. To identify the bad bands after statistical analysis K-means clustering is used and clusters are formed as informative bands cluster and non-informative bands cluster. Three clusters are formed for representing types of bands. The resultant dataset is obtained as a representative class from each cluster. The labels assigned to the data are non-illuminated bands, water vapor absorption bands, and Informative bands.

A dataset is generated based on the statistical analysis and clustering results. In dataset generation, a table containing statistical data for each image band (10 images) is constructed. Now this table is used to form clusters and identify the label for each band. Now the labeled dataset is ready. This dataset is used for training the classifier. The chosen classifier, K-Nearest Neighbors (KNN), is utilized in training the model. The Nearest neighbor parameter is set to 3 and default leaf size is selected as 30. The metric used is Mankowski with parameter p=2. Hybrid KNN is implemented with voting method. The Hard Voting is selected for the class with the most votes. Soft Voting is selected in case of average predicted class probability which selects the class with highest probability. The soft voting is applied if both classifiers output probability estimates.

Testing is performed to evaluate the classifier's accuracy. The test size is selected 25% with random state 100. At last, Threshold values are to be finalized for the identification of bad bands and informative bands. The finalized thresholds will be used in developing a GUI for bad band identification as one of the preprocessing steps. Result and performance analysis: Band types are identified based on the threshold values. The performance of developed algorithm is checked with leave one out method.

Algorithm: Classification of Bands
Input: An image X for band classification.
Output: Classification result - Type of band.
Algorithm:

- For all 242 bands, X_{ij} represents the pixel value at the ith row and jth band.
  Calculate statistical Features
- For j=1 to 242
  Calculate statistical features F_j
- Apply the clustering algorithm to the feature vectors F_j for each band j.
- Let, C_j be the clusters formed.
- Analyze the clusters C_j and devise a labeled dataset Lj where each pixel is labeled as L_{ij}, indicating the cluster to which it belongs.
- Train the classifier over the Lij Dataset and evaluate it over the testing dataset with 75-25 split. Set minkowski distance p=2 , k=3
- Set Voting= 'soft' / 'hard' for majority vote
- From the testing results thresholds T1, and T2 are finalized
  T1=0.5 and T2=10.5
- Compare with thresholds to classify.
- For each band j, Compare the classification threshold Tj with features Fj:
  $$\begin{cases} \text{“}Bad\ Band\text{”}, & if\ Fj <= T1 \\ \text{“}Bad\ Band\ due\ to\ Water\ Vapors & if\ T1 < Fj <= T2 \\ \text{“}Informative\ Band\text{”}, & otherwise \end{cases}$$
- Classification Result:
  For j=1 to 242
  output Class_j

3 Results and Discussion

3.1 Unsupervised Clustering

The K-means algorithm is iteratively applied ten times to the data produced from statistical analysis, for obtaining training dataset. The outcome reveals sum of squared error 15.38. The initial centroids Cluster 0: (226, 0.426) and Cluster 1: (179, 0.718135), are randomly selected as starting points. The result of K-means clustering indicates that 155 bands are grouped together with 64% in Cluster 0, and 87 bands are clustered with 36% in Cluster 1.

The density-based clustering algorithm is employed with 10 iterations on the same data to enhance the results. It gives 15.387 within-cluster sum of squared error. Randomly selected initial starting points are Cluster 0: (226, 0.426) and Cluster 1: (179, 0.718135). Global replacement of missing values with mean/mode values is conducted. Cluster 0 comprises 158 instances, constituting 65%, while Cluster 1 consists of 84 instances, accounting for 35%. The log-likelihood obtained is -11.46518, reflecting the formation of two clusters, one containing 158 bands and the other with 84 bands.

3.2 Supervised Classification

The process of generating the optimized training dataset is concluded by incorporating the outcomes of both statistical analysis and clustering results. The dataset is partitioned into a training set (75%) and a testing set (25%). The dataset labeled previously is used for training using K-nearest neighbor (KNN) classifier using Minkowski distance metric, 2 with k=3 and the leaf size of 30. The soft or hard voting is selected depending on the class probabilities. This hybrid KNN approach improved the efficiency of classification. The resulting accuracy is an impressive 97%. After thorough analysis, training, and testing of hybrid KNN classifier, the thresholds for identifying different bands are ultimately finalized. After finalizing the thresholds, the system performance is tested using leave-one-out validation for KNN and decision tree classifier. The accuracy of KNN algorithm is 100% and for decision tree 98%.

3.3 Finalizing the Thresholds

If the threshold is less than or equal to 0.5 then that image band will be identified as a bad band due to the non-illumination. If the threshold is between 0.5 and 10.5 then the image band is identified as a bad band due to the water vapor absorption. Otherwise, if the threshold is greater than 10.5 then that band is informative and contains useful information for further processing. Table 1 shows the thresholds identified for each type of band.

Table 1. Thresholds For Band Identification

Band Type	Threshold for Band
Bad Bands due to non-Illumination	T <=0.5
Bad Bands due to Water Vapor Absorption	0.5< T <=10.5
Informative Bands	T >10.5

3.4 Software/ GUI for Bad Band Identification

The software module developed for the identification of band types is shown in Figure 2. Band 27 is an informative band and the output of band 126 is the bad band as it contains water vapors.

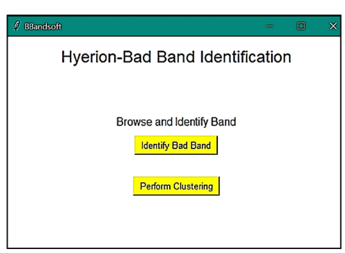

Fig 2. Software Module

The result of perform clustering is as shown in Figure 3.

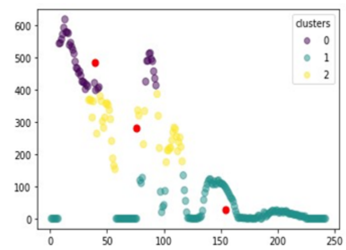

Fig 3. Output of k-means Clustering

The three clusters 0,1 and 2 corresponding to different band types are formed and the corresponding cluster centroids are marked with red points. After analysis of all the images the different types of bands are finalized. These bands are listed in Table 2 with the type of band identified. This table presents the final results of our analysis, providing a universally applicable selection of informative bands from the Hyperion sensor. These selected bands are confirmed to be useful for further spectral analysis across various applications.

Table 2. Output Class of Band Types

Non-illuminated Bad Bands	Bad Bands due to Water Absorption	Informative Bands
1 to 7	121 -133, 143	8 to 57, 77 to 120
58 to 76	165- 182, 184-187	134 to 164
225 to 242	219- 224	183,188,189, 190-218

The proposed methodology on average identified 156 bands as informative which eliminated more bad bands. The comparative analysis using Lonar crater dataset has been performed. Overall, ENVI selected on an average of 16% bad bands for further processing with an accuracy of 87% while the proposed algorithm gave 98 to 100% of accuracy. Also, on average ENVI selected 196 informative bands while the proposed methodology selected 156 bands. The graph of comparison with ENVI software are as shown in Figure 4 presenting improved accuracy. The Lonar crater images dataset is preprocessed with proposed GUI, considering informative bands and ENVI which excludes predefined bands, the result of mineral identification of both is comparable. The significant minerals are identified with good score using the proposed algorithm. Table 3 shows the performance comparison of proposed work with the existing one.

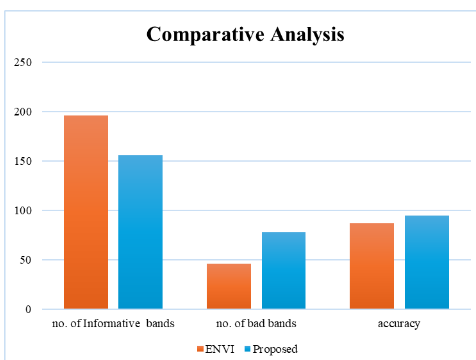

Fig 4. Comparative Analysis with ENVI

This surpasses the results obtained by Ji et al.[11], Kumar, et al.[9], indicating a more robust approach to bad band identification.Moreover, the cost-effectiveness of this tool addresses a significant barrier in hyperspectral image analysis—the high expense of existing software solutions. By providing a reliable and affordable alternative, this method democratizes access to advanced preprocessing techniques.

Table 3. Performance Comparison with Exiting Work

Method	Accuracy (%)	Bad Band Handling	Key Advantage
Proposed Method	98-100%	Effective band identification based on threshold values	Highly accurate and computationally efficient
ENVI Software (Commercial)	87%	Higher false positives in bad band identification	Costly and requires licensing
Ji et al. (2019) – Matched Filter Weights [11]	92%	Detects noisy bands but has limitations in spectral variability	Requires extensive training data
Kumar et al. (2022) [9] – CNN with Dual Partitioning	99%	High computational cost for real-time applications	High accuracy but requires deep learning expertise

4 Conclusion

For hyperspectral images, preprocessing is most important step before spectral analysis. The current work has proposed a methodology which refines band selection by selecting more bad bands than previous techniques and provides most informative bands which improves the accuracy of spectral analysis. The clustering techniques were applied to decide the threshold values that can be used to classify and identify the image bands. The band types are Bad band due to non-illumination, bad band due to water vapor absorption, and informative bands. The classifier is trained on this dataset and then the testing is performed to evaluate the accuracy of the classifier. The thresholds are finalized, and the GUI is developed in Python. Hybrid KNN classifier gave an accuracy of 100% with outperforming accuracy 98% by Decision tree classifier. The proposed work identified two significant thresholds 0.5 and 10.5 for bad band identification.

The proposed method exhibited consistent results for different datasets without retraining or fine-tuning. As compared to traditional methods that rely on spectral signatures or atmospheric corrections, the proposed algorithm introduces a threshold-based approach which makes preprocessing faster and more efficient. The GUI developed for the proposed algorithm provides a user friendly, simple and cost-effective alternative to expensive software like ENVI.

For future work, destriping process can be improved to overcome the drawback of the time intensive preprocessing task for existing solutions. Also, the generalization of the proposed algorithm can be done by extending it for PRISMA, AVIRIS, and EnMAP. It will increase the usability of the GUI/software developed if it will be integrated with plugin for existing remote sensing tools like ENVI, QGIS, or ERDAS Imagine.

References

1) Dapke P, Quadri SA, Nagare SM, Bandal SB, Baheti MR. A Comparative Analysis Machine Learning Techniques for LULC Classification Using Landsat-8 Satellite Imagery. *International Journal of Engineering and Geosciences*. 2025;10(1):84–92. https://doi.org/10.26833/ijeg.1503104.
2) Gore R, Mishra A, Deshmukh R, Randive P, Abbasov IB. LULC-Analysis of Land-use with the help of Unsupervised Classification. *Izvestiya SFedU Engineering Sciences*. 2020;64(2):184–192. https://doi.org/10.18522/2311-3103-2020-3-184-192.
3) Singh S, Martha TR, Das IC, Sreenivas K, Chauhan P. Comparison of Airborne and Spaceborne Hyperspectral Data for Lithological Discrimination Within the Eastern Dharwar Craton. *J Indian Soc Remote Sens*. 2025. https://doi.org/10.1007/s12524-025-02172-5.
4) Vignesh KM, Kiran Y. Comparative analysis of mineral mapping for hyperspectral and multispectral imagery. *Arabian Journal of Geosciences*. 2020;13:1–12. https://doi.org/10.1007/s12517-020-5148-8.
5) Gore R, Mishra A, Deshmukh R. Exploring the Mineralogy at Lonar Crater with Hyperspectral Remote Sensing. *Journal Geological Society India*. 2021;97(3):261–266. https://doi.org/10.1007/s12594-021-1676-4.
6) Kowkabi F, Keshavarz A, Kumar L. A Novel Statistical Preprocessing Approach for Hyperspectral Image Unmixing. 2021;p. 190–193. https://doi.org/10.1109/InGARSS51564.2021.9792139.
7) Sun S, Bao W, Qu K, Feng W, Zhang X, Ma X. Hyperspectral Image Super-Resolution Algorithm Based on Graph Regular Tensor Ring Decomposition. *Remote Sensing*. 2023;15(20):1–27. https://doi.org/10.3390/rs15204983.
8) Liu C, Dai H, Wang S, Chen J. Remote Sensing Image Scene Classification Based on Multidimensional Attention and Feature Enhancement. *IAENG International Journal of Computer Science*. 2023;50(4).
9) Kumar BLNP, Vaddi R, Manoharan P, Agilandeeswari L, Sangeetha V. A new band selection framework for hyperspectral remote sensing image classification. *Scientific Reports*. 2024;14(1):1–21. https://doi.org/10.1038/s41598-024-83118-8.
10) Cozzolino D, Williams PJ, Hoffman LC. An overview of pre-processing methods available for hyperspectral imaging applications. *Microchemical Journal*. 2023;193:1–5. https://doi.org/10.1016/j.microc.2023.109129.
11) Ji L, Wang L, Geng X. An automatic bad band pre-removal method for hyperspectral imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2019;12(12):4985–4994. https://doi.org/10.1109/JSTARS.2019.2944930.