Hyperspectral Image clustering and local feature selection using Gini Impurity

Prashant Kumar Mali, Hitenkumar Motiyani, Quazi Sameed, Anand Mehta
Department of Civil Engineering
Institute of Infrastructure, Technology, Research And Management, Ahmedabad, Gujarat, 380026
prashant.mali.19c@iitram.ac.in, hitenkumar.motiyani.19c@iitram.ac.in, sameed.quazi.19c@iitram.ac.in, anandmehta@iitram.ac.in

***Abstract*—This study proposes a unique segmentation-based clustering algorithm that utilises *k*-means for segmentation, further uses a local feature selection technique to obtain the top bands for each cluster and deploys clustering on segmented hyperspectral imagery. The suggested methodology is a framework with several stages. *k*-means is initially utilized for image segmentation. From the obtained segments, significant segments are identified using Gini impurity. Finally, the cluster map is obtained by merging insignificant clusters with significant clusters. This step also makes use of novel local feature selection strategy. Three sets of hyperspectral images are used in investigations to evaluate the efficiency of the proposed methodology. For assessment, the criteria Normalized Mutual Information and Purity score are utilised. The investigation findings demonstrate that the proposed methodology outperforms the other segmentation methodologies that were compared. According to the results, using band selection and redundancy strategies significantly improves accuracy.**

***Index Terms*—Segmentation, Feature Selection, Hyperspectral Image, Gini Impurity, Clustering.**

I. Introduction

Hyperspectral Imaging (HSI) has received a lot of interest in processing recently. Since, HSIs may offer rich band information from many wavelengths, they are frequently used in a variety of research fields [1]. The reflectances of electromagnetic waves of various wavelengths are captured by HSIs, and each electromagnetic wave's reflectance is saved as a two dimensional image. Thus, an HSI is a data cube that includes a large number of two dimensional images [2]. Even while HSI applications have seen tremendous success, dealing with huge dimensional data to extract information remains a difficult task [3]. Hence, to tackle such crucial situation, techniques like cluster analysis, image segmentation and dimentionality reduction technique like principal component analysis are utilized [4].

Cluster analysis is one strategy for developing essential data analysis methods in the early stages [5]. It is intended to divide experimental data into a certain number of sets, each of which should contain components that are unique from the others and as comparable to one another as is practical. This implies that there is some degree of separation or similarity between the things that need to be classified [6]. The number of these classes may be predetermined or the result of restrictions placed on them.

The scope of this research is a new clustering technique that employs Gini Impurity. Also, a novel local band selection strategy with relevancy and redundancy concept is proposed. Here, local band selection means we are using reduced bandset for each cluster. While, the rest of the paper is structured as follows. Section II describes the related work of existing studies. Section III provides a description of the methodology. The datasets and functions used are listed in Section IV. The study's results are presented in Section V. Section VI serves as the paper's conclusion.

II. Related work

Some of the hyperspectral image clustering work is based on a well-liked algorithm called the Sequential Hierarchical method, which treats each entity as a separate cluster [7]. Two clusters are combined at each level of the clustering process, and the process is repeated until only one cluster, comprising the whole dataset, is left. However, partitioning techniques result in discrete, non-overlapping groups [8]. Due to the fact that only a single data division is formed, the technique is frequently referred to as nonhierarchical clustering [9].

In general terms, dimensionality of HSI can be reduced using feature extraction or feature selection (also known as band selection) techniques. In order to extract features, the original HSI is projected into a lower dimensional space, which results in the creation of a smaller data set [10]. To accurately represent the original data set, some discriminative bands are chosen while selecting bands. Feature extraction often generated better outcomes in tests. The smaller data sets are easier to understand because of band selection, which can preserve the physical information of the original data [11].

In many image processing methods, segmentation is an essential component. It is possible to identify regions of interest and objects in the scene from the segmentation findings, which is super beneficial for the subsequent image analysis or annotation [12]. Some of the recent work which have carried out clustering of Hyperspectral images are: [7], [8], [9], [10], [11], [12], [13], [14].

III. Proposed Methodology

The proposed methodology is described in the following manner: performing the segmentation on a hyperspectral image using *k*-means in order to acquire new segments. The next step involves using Gini Impurity to identify significant segments. Considering significant segments with bands having minimal

redundancy from the local band selection technique in the following stage, combine the remaining segments with the significant ones according to minimum distance from significant segments. As, HSI has high dimentionality or large number of bands [15]. Lot of these bands are correlated. In other words, lots of band have redundant information. Due to this we assume that deploying band selection strategy may result in improvement in accuracy. Visual depiction of the suggested methodology is shown in Fig. 1.

Gini Impurity (GI) is used in methodology which basically is a measurement of the likelihood that, if an element were randomly labelled, it would be correctly identified. Formally, GI is formulated in equation as:

$$\text{Gini}(g) = \sum_{i=1}^{J} p_i(1 - p_i) \tag{1}$$

where, g is the Gini Impurity, p_i is the i^{th} pixel of cluster and J is the total number pixels in a cluster.

In this work, Gini Impurity is used to recognize significant segments by arranging all segments in ascending order, as here our assumption is that the lower Gini Impurity implies that for a segment most of the pixel values are similar. Further, Gini Impurity is utilized to identify the top bands in each cluster by arranging bands with greater Gini Impurity in descending order as here we assume that high value of Gini Impurity means more variability in the bands which translates to more information content.

A. *Image Segmentation*

The technique of dividing a digital image into various homogeneous sub regions is known as Image Segmentation [6]. Image Segmentation's intention is to group pixels together into significant image regions, or regions that correlate to high affinity and specificity, objects, or natural portions of objects [16]. The procedure for segmenting an image is shown by utilising *k*-means to arrange the dataset into a matrix. While, *k*-means is a clustering based algorithm [17]. This method seeks the optimizing result by looking for *k* divisions that meet a particular requirement [18]. First, select a few pixels to serve as the initial cluster centroid; next, group the remaining pixels to their cluster centroids in accordance with the requirement of the minimum distance; finally, we will obtain the initial classification; however, if the classification is unreasonable, we will modify it, iterating repeatedly until we obtain a satisfactory classification [19]. Additionally, after performing segmentation through *k*-means, new segments are created which are organised via Gini Impurity and further, these segments are involved in band selection technique.

B. *Local Band Selection (Local Feature Selection)*

A hyperspectral image has a very high spectral resolution due to the number of bands it contains [1]. As a result, it's critical to obtain the bands in a way that preserves as much information as possible. The band selection method measures the importance of each spectral band according to some informative measurements such as Gini Impurity which uses a ranking criteria to select the top-ranked bands in a sorted sequence containing more relevant and less redundant bands [10]. Moreover, redundancy is produced by bands with comparable gini indices. So, a distance-weighted parameter score is introduced to prevent this. The proposed band selection strategy is inspired by the methodology discussed in [11] and [20].

Formally, score is formulated in equation as:

$$delta = \max_{GI_j > GI_i} d \tag{2}$$

$$\text{score} = (delta) * (GI_i) \tag{3}$$

where, *delta* is a function that stores maximum distance obtained by calculating all the distances between a test band and all the bands having gini Index higher than that of test band, d is distance between mean vectors of all pixels in a cluster for particular segment to remaining bands of the segment, *GI* is the gini Index for corresponding bands, i and j are band indexes.

In this study, score incorporates both relevancy and redundancy criteria, *GI* is used to identify the relevant bands and *delta* is used to reduce redundancy. The algorithm for proposed methodology is depicted below

Algorithm 1 Segmentation with local band selection

Input: Hyperspectral Image.
Output: Clustered Image.

- Segmentation using *k*-means clustering. Convert cluster map to segmentation map. It will contain number of segments (group of pixels).
- Now, each segment is considered as a cluster.
- Arrange all clusters in ascending order of their Gini Impurity.
- Choose top segments which contain number of pixels greater than five as we assume cluster less than five do not have much informative content.
- Similarly for each segment compute Gini Impurity and *delta* for all bands. A higher Gini Impurity signifies variability in a band, hence more significant.
- Calculate *delta* by detecting bands whose gini is greater than j^{th} band and by finding max distance among band whose gini is greater than j^{th} band.
- Evaluate score and arrange segments in descending order of score parameter.
- Now assess the pairwise distance between top Significant and remaining segments considering identified bands for each significant clusters.
- Propagate labels to clusters from significant clusters according to minimum distance.
- Now rearrange these segments to form a cluster map.

IV. Experimental Setup

The stated approaches was programmed using Python through a personal machine with sixteen gigabytes of Random

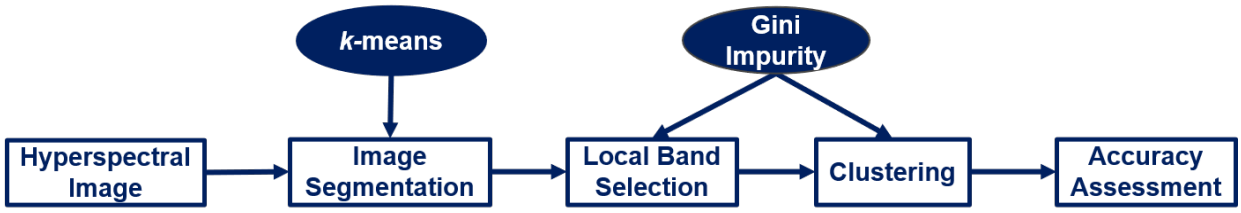

Fig. 1: Proposed Methodology

Access Memory (3200 MHz) and Intel i5 10th Generation processor (4.30 GHz Max Turbo Frequency), with a four gigabytes of dedicated graphics memory of NVIDIA GTX 1650, operated on windows 11 [21].

This approach was evaluated using various hyperspectral data, and it was then compared at various phases of the proposed methodology. Firstly using *k*-means for image segmentation to generate new segments, then utilising the local band selection method while taking into account significant segments with bands having minimal redundancy, combine the remaining segments with the significant segments according to the minimum distance from significant segments.

Nevertheless, a few additional approaches were used for comparison, as illustrated. Principal Component Analysis (PCA) with *k*-means was implemented where the information were normalised as well as to ascertain their associated covariance was discovered. Moreover, Simple Linear Iterative Clustering (SLIC) [22] is performed which suggests the similarity clustering of pixel space position and pixel colour feature to achieve a segmented image. Firstly, segments are created in SLIC and then, gathered together by defining gini index of each segment. Further, according to increasing gini index, each segment is arranged, as the segment having least gini value shows more information. Likely, best segments are selected and rest of them are accumulated with the selected ones by obtaining mean of individual segments and calculating distances between all of them, respectively. Moreover, the proposed approach is also compared with two other hyperspectral clustering approaches, including nearest neighbour concept, H2NMF [23] and CFSFDP [24]. Similar to the initial phase of the proposed method, segmentation comes before CFSFDP.

A. *Data Set*

The 204-band Salinas dataset was captured over the Salinas Valley in California using a 224-band Airborne Visible Infrared Imaging Spectrometer sensor. It distinguishes itself with a high spatial resolution of 3.7-m pixels. There are 512 lines and 217 samples in the covered region. Salinas ground truth includes 16 distinct forms of land use and cover [25].

The other datasets used in the tests are from Pavia Center and Pavia University. During a flight campaign over the northern Italian city of Pavia, the ROSIS (Reflective Optics System Imaging Spectrometer) sensor obtained these two images. Pavia University has 103 spectral information and nine classes, compared to Pavia Center's 102 spectral and nine classes, and its subset's eight. The original datasets from Pavia Centre and Pavia University are 1096 × 1096 and 610 × 610 pixels respectively, and are reduced into 400 × 400 pixels and 610 × 340 pixel datasets. Both datasets have 1.3 metres of geometric resolution.

B. *Accuracy Assessment*

The functions Normalized Mutual Information (NMI) and Purity are used to evaluate the Accuracy of Clustering. [26]. Here, the normalised mutual information function evaluates the degree of agreement between the two assignments. The NMI has a range of 0 to 1. If the class is perfectly generated by the clustering, NMI = 1; otherwise, NMI = 0, indicating that the class membership is random.

Let $\mathbb{C}$ be the set of classes produced from the ground reference data, and let Ω be the set of Clusters derived by the algorithm. The Mutual Information, $\mathrm{MI}(\Omega, \mathbb{C})$ calculated as follows:

$$\mathrm{MI}(\Omega, \mathbb{C}) = \sum_k \sum_j p(\omega_k \cap c_j) \log_2 \frac{p(\omega_k \cap c_j)}{p(\omega_k) \cdot p(c_j)} \tag{4}$$

where $p(\omega_k)$, $p(c_j)$ and $p(\omega_k \cap c_j)$ are probabilities that a randomly chosen pixel belongs to the cluster ω_k, class c_j and cluster ω_k as well as class c_j , respectively. Then the NMI can be obtained as follows:

$$\mathrm{NMI}(\Omega, \mathbb{C}) = \frac{\mathrm{MI}(\Omega, \mathbb{C})}{\mathrm{mean}(\mathrm{H}(\Omega), \mathrm{H}(\mathbb{C}))} \tag{5}$$

where $\mathrm{H}(\Omega) = -\sum_k p(\omega_k) \log_2 p(\omega_k)$ and $\mathrm{H}(\mathbb{C}) = -\sum_j p(c_j) \log_2 p(c_j)$ are the entropies of Ω and $\mathbb{C}$, respectively.

While, purity is the extent to which clusters belong to a single class. It falls inside the range of [0, 1] and is a real number. When purity rises, clustering performance will get better.

$$\mathrm{Purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j| \tag{6}$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is the set of clusters, $\mathbb{C} = \{c_1, c_2, \ldots, c_j\}$ is the set of classes and N = total number of test pixels.

V. Results And Discussion

Image segmentation and band selection techniques are used to classify hyperspectral images while taking into account relevance and redundancy. First, we have compared the result of the proposed method i^{th} clustering techniques. Then, we

have carried out following experiments: k_seg, k_seg + LBS_R and proposed.

Table I shows accuracy values corresponding to various parameters. Nonetheless, apart from proposed method, approaches such as local band selection technique with redundancy is conveyed. Let say that, kM stands for k-means clustering while LBS_R portrays band selection technique with Redundancy factor not included. However, k_seg represents number of segments, k_cl shows number of significant clusters and L demonstrates number of top bands taken into consideration.

Moreover, PCA + kM is used, using a fraction of principal components (PC) chosen to ensure that PCs can explain atleast 99% of the variance in the dataset. In PCA + SLIC + kM best segment values are 60, 180, and 170 for Salinas, Pavia Center, and Pavia University's dataset.

TABLE II: Salinas accuracy values.

	k_seg	k_cl	L	NMI	Purity
k_seg	5	28	All bands	0.7921	0.6738
k_seg + LBS_R	7	18	12	0.7932	0.6426
proposed	6	26	12	0.8130	0.7011

Many initialised k_seg values are implemented in the range of 2 to 15 while executing k-means segmentation-based clustering for the Salinas dataset. Moreover, various acceptable NMI values are generated, and the maximum NMI value is indicated. The best number of regions for Salinas was discovered to be six (for proposed) as listed in Table II.

Similarly, a set of k_cl values is used for clustering, and L values are changed for band selection. Starting with k_cl = 10 to 30 and L changed from 10 to 100, corresponding NMI values are seen, and the greatest NMI value is taken into consideration. NMI is calculated using the best values of k_cl and L, and it is then examined to determine that it has greatly increased from 79.21% to 81.3%. Accuracy is improved by 0.14% and 2.57%, respectively, using the band selection technique with the redundancy removal score on bands.

TABLE III: Pavia Center's accuracy values.

	k_seg	k_cl	L	NMI	Purity
k_seg	5	17	All bands	0.7388	0.8641
k_seg + LBS_R	3	19	13	0.7538	0.8208
proposed	5	18	10	0.7976	0.9001

Many initialised k_seg values are implemented in the range of 2 to 15 while executing k-means segmentation-based clustering for the Pavia Center dataset. Moreover, various acceptable NMI values are generated, and the maximum NMI value is indicated. The best number of regions for Pavia Center was discovered to be five (for proposed) as listed in Table III.

Similarly, a set of k_cl values is used for clustering, and L values are changed for band selection. Starting with k_cl = 8 to 20 and L changed from 10 to 100, corresponding NMI values are seen, and the greatest NMI value is taken into consideration. NMI is calculated using the best values of k_cl and L, and it is then examined to determine that it has greatly increased from 73.88% to 79.76%. Accuracy is improved by 1.99% and 7.32% respectively, using the band selection technique with the redundancy removal criteria included.

TABLE IV: Pavia University's Accuracy values.

	k_seg	k_cl	L	NMI	Purity
k_seg	3	17	All bands	0.5042	0.7087
k_seg + LBS_R	3	19	15	0.5242	0.7009
proposed	3	19	18	0.5303	0.7141

Many initialised k_seg values are implemented in the range of 2 to 15 while executing k-means segmentation-based clustering for the Pavia University dataset. Moreover, various acceptable NMI values are generated, and the maximum NMI value is indicated. The best number of regions for Pavia University was discovered to be three (for proposed) as listed in Table IV.

Similarly, a set of k_cl values is used for clustering, and L values are changed for band selection. Starting with k_cl = 8 to 20 and L changed from 10 to 100, corresponding NMI values are seen, and the greatest NMI value is taken into consideration. NMI is calculated using the best values of k_cl and L, and it is then examined to determine that it has greatly increased from 50.42% to 53.03%. Accuracy is improved by 3.82% and 4.92% respectively, using the band selection technique with inclusion of the redundancy removal strategy.

Fig. 2 displays a distinction between the original image and the classified images from the Salinas dataset. The proposed methodology outperformed the compared segmentation-based clustering methodologies, as shown in Table I. Also, it is noted that the higher accuracy Scores are generated from proposed strategy.

VI. CONCLUSION

In this study, a new clustering strategy is proposed for hyperspectral image. First image segmentation is carried out then the obtained segments are clustered on the basis of Gini Impurity. Further a local band selection is also proposed which is utilized in the last stage of the clustering strategy. Local band selection takes into account both relevancy and redundancy criteria while identifying more suitable bands. The proposed strategy was then compared with other clustering strategies. From the experiments it was noted that utilization of local band selection technique resulted in higher accuracy and the main constraint that was faced during experimentation was identification of suitable parameters for image segmentation, clustering and feature selection (band selection). Overall, it

TABLE I: Assessment of Salinas, Pavia Center, and Pavia University's accuracy using several methods.

Dataset	Salinas			Pavia Center			Pavia University		
	k_cl	NMI	Purity	k_cl	NMI	Purity	k_cl	NMI	Purity
kM	16	0.7742	0.6734	8	0.7694	0.7961	9	0.5330	0.6696
PCA + kM	16	0.7656	0.6785	7	0.7750	0.8048	8	0.5663	0.7150
k_seg	28	0.7921	0.6738	17	0.7388	0.8641	17	0.5042	0.7087
k_seg + LBS_R	18	0.7932	0.6426	19	0.7538	0.8208	19	0.5242	0.7009
PCA + SLIC + kM	16	0.7185	0.6112	8	0.6743	0.8060	9	0.4375	0.6582
H2NMF [24]	26	0.6372	0.5877	8	0.7596	0.8592	9	0.4641	0.6261
CFSFDP [23]	26	0.6960	0.6703	14	0.7591	0.8219	9	0.5772	0.6973
proposed	26	0.8130	0.7011	18	0.7976	0.9001	19	0.5303	0.7141

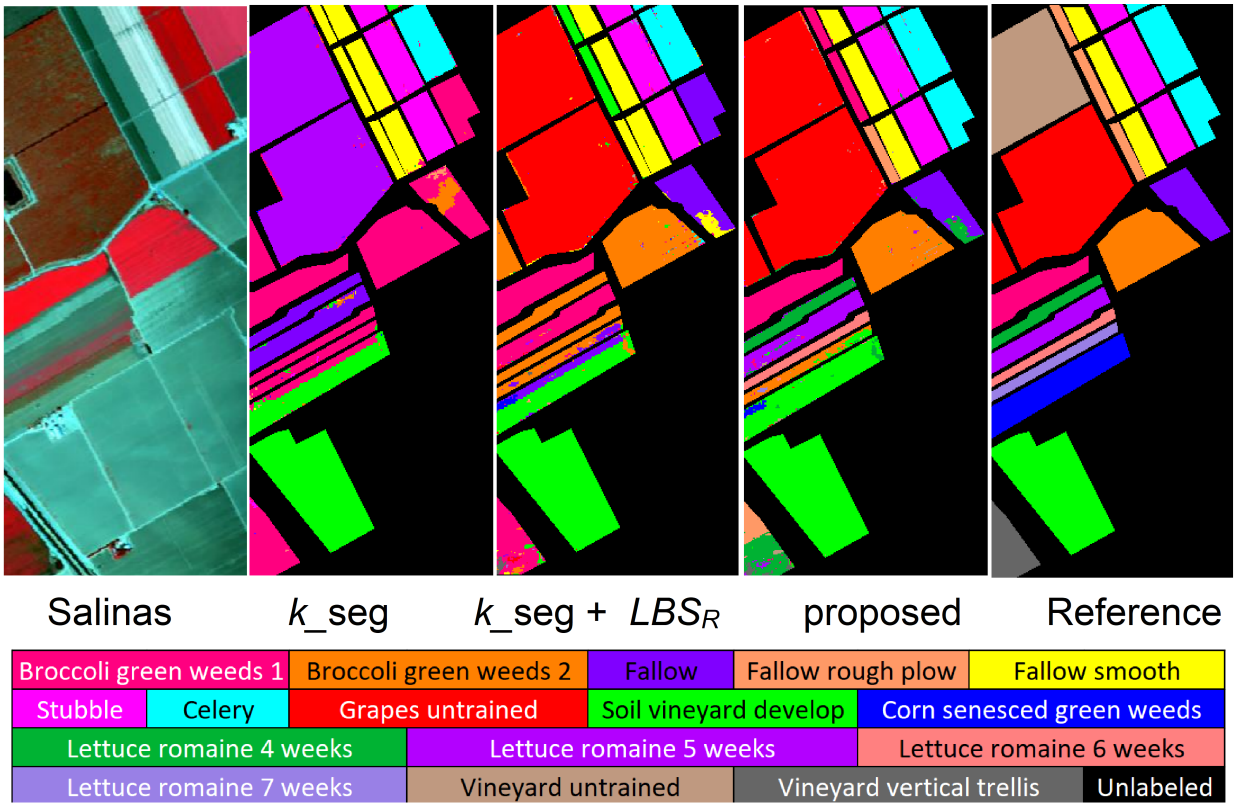

Fig. 2: Salinas Dataset's clustering maps.

can be concluded that the proposed strategy has significantly outperformed other compared clustering strategies.

VII. ACKNOWLEDGEMENT

The Hyperspectral Images used in this study were provided by Prof. P. Gamba and Prof. David Landgrebe, for which the authors are grateful.

REFERENCES

[1] A. Femenias and S. Marín, "Hyperspectral imaging, chapter 15," in *Electromagnetic Technologies in Food Science*. John Wiley Sons, Ltd, 2021, pp. 363–390.

[2] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Sparsity-based clustering for large hyperspectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 410–10 424, 2021.

[3] M. Mateen, J. Wen, D. Nasrullah, and M. Azeem Akbar, "The role of hyperspectral imaging: A literature review," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, 09 2018.

[4] A. Mehta and O. Dikshit, "Comparative study on projected clustering

methods for hyperspectral imagery classification," *Geocarto International*, vol. 31, pp. 296–307, 05 2015.

[5] A. F. Alkarkhi and W. A. Alqaraghuli, "Cluster analysis, chapter 11," in *Easy Statistics for Food Science with R*, A. F. Alkarkhi and W. A. Alqaraghuli, Eds. Academic Press, 2019, pp. 177–186.

[6] A. Mehta and O. Dikshit, "Projected clustering of hyperspectral imagery using region merging," *Remote Sensing Letters*, vol. 7, no. 8, pp. 721–730, 2016.

[7] A. Mehta and S. Pasari, "Hyperspectral image clustering using nearest neighbor," in *2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS)*, 2021, pp. 194–197.

[8] S. L. Polk and J. M. Murphy, "Multiscale clustering of hyperspectral images through spectral-spatial diffusion geometry," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 4688–4691.

[9] C. Hinojosa, E. Vera, and H. Arguello, "A fast and accurate similarity-constrained subspace clustering algorithm for hyperspectral image," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 773–10 783, 2021.

[10] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag, Berlin, Heidelberg, 5th Edition, 2013, pp. 403–446.

[11] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1242072

[12] A. Mehta and O. Dikshit, "Segmentation-based projected clustering of hyperspectral images using mutual nearest neighbour," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 12, 2017, pp. 5237–5244.

[13] Q. Cao, D. Mishra, J. Wang, S. Wang, H. Hurbon, and M. Y. Berezin, "Hskl: A machine learning framework for hyperspectral image analysis," in *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2021, pp. 1–5.

[14] H. Motiyani, P. K. Mali, and A. Mehta, "Hyperspectral image segmentation, feature reduction and clustering using k-means," in *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2022, pp. 389–393.

[15] K. Cui and R. J. Plemmons, "Unsupervised classification of aviris-ng hyperspectral images," in *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2021, pp. 1–5.

[16] H. Chokshi and A. Agarwal, "Image segmentation," in *Advanced Sensing in Image Processing and IoT 1st Edition*, 02 2022, pp. 43–62.

[17] Y. Li and H. Wu, "A clustering method based on k-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 12 2012.

[18] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[19] A. K. Jain, M. N. Murty, and P. J. Flynn., *Data Clustering: A Review*, 1999, pp. 264–323.

[20] A. Mehta and O. Dikshit, "Segmentation-based clustering of hyperspectral images using local band selection," *Journal of Applied Remote Sensing*, vol. 11, no. 1, pp. 015 028–015 028, March 2017.

[21] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 6 2014. [Online]. Available: https://doi.org/10.7717/peerj.453

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[23] N. Gillis, D. Kuang, and H. Park, "Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2066–2078, 2014.

[24] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[25] M Graña, MA Veganzons, B Ayerdi, "Hyperspectral remote sensing scenes," accessed: 2021-12-26. [Online]. Available: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[26] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, cambridge Books.