Towards Smarter Vegetation Health Clustering: Insights from Fractal Dimension, NDVI, and LST Metrics Derived via Remote Sensing Landsat Dataset

Suhad A. Yousif[1], Venus W. Samawi[2(✉)], Nadia M. G. Al-Saidi[3], and Yeliz Karaca[4]

[1] Computer Science Department, Al-Nahrain University, Baghdad, Iraq
suhad.a.yousif@nahrainuniv.edu.iq
[2] Department of MIS/Smart Business, Isra University, Amman 11622, Jordan
venus.samawi@iu.edu.jo
[3] Department of Applied Sciences, University of Technology, Baghdad 10066, Iraq
nadia.m.ghanim@uotechnology.edu.iq
[4] University of Massachusetts Chan Medical School (UMASS), 55 Lake Avenue North, Worcester, MA 01655, USA
yeliz.karaca@ieee.org

Abstract. Vegetation plays a vital role in preserving the environment and mitigating the effects of climate change. Therefore, it is essential to identify areas of stress and non-vegetation factors to develop a plan to establish vegetation and combat desertification, which positively affects climate change and ensures a sustainable future. Remote sensing and satellite images have been used to identify vegetation areas and their density, which is a challenging problem. According to the literature, the Normalized Difference Vegetation Index (NDVI) is used in most studies to identify vegetation areas, while few utilize fractal dimensions and Land Surface Temperature (LST). Fractals are known for their effectiveness in identifying complex distribution patterns, such as vegetation areas in satellite images. This study will tackle the issue of categorizing the vegetation area into healthy, stressful, and non-vegetation regions based on satellite images. Two clustering methods are applied, namely K-means and threshold-based labelling, to categorize the vegetation area in the three aforementioned areas. The K-means and threshold-based labelling will be evaluated and compared based on the Silhouette score, Davies-Bouldin index (DBI), and Entropy metrics. The effect of selecting the proper features on clustering performance is also studied. The experimental results show that the threshold-based clustering method suits datasets with well-defined threshold boundaries between clusters. Conversely, the K-means clustering method provides flexibility for adaptive clustering with no pre-defined thresholds. Finally, it is worth noting that Fractal dimensions (FD) with threshold-based labelling achieve the best clustering results, making it crucial for detecting vegetation health and analyzing land surfaces.

Keywords: Clustering Vegetation Areas · K-means clustering · threshold-based Clustering · Normalized Difference Vegetation Index · Silhouette score · DBI · Entropy · Land Surface Temperature

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15886, pp. 70–87, 2026.
https://doi.org/10.1007/978-3-031-97576-9_5

1 Introduction

Vegetation is the lungs through which the Earth breathes. It limits global warming and plays a vital role in purifying the atmosphere of toxic gases and reducing climate change. Therefore, it is essential to focus on developing vegetation areas by specifically determining the distribution of the planet′s cover to improve regions with poor vegetation. Developing vegetation areas and promoting smart farms play an important role in building a sustainable future [1–4]. To develop vegetation areas, it is important to categorize vegetation regions into healthy, stressed, and non-vegetation based on remote sensing [5, 6] or satellite images. Satellite images are images of the Earth′s surface in different formats (e.g., multispectral, panchromatic, synthetic aperture radar images) [7].

One common method for vegetation detection is the NDVI [8]. Some studies combine NDVI, which uses the red and near-infrared (NIR) bands, with the enhanced vegetation index (EVI), which uses blue bands in addition to the red and near-infrared (NIR). Light detection and ranging (LiDAR) is an accurate method for obtaining three-dimensional (3D) surface information and is often used in vegetation studies. To extract the vegetation point set and describe the spatial morphological characteristics of different features in the LiDAR point cloud, 3D fractal dimensions [9] are used. The 3D fractal dimensions indicate the irregularity and roughness of land features [10–12]. Different types of features exhibit unique distribution and variation trends in their fractal dimensions. Therefore, fractals are used to improve detection accuracy in satellite images for vegetation detection. In 2023, Martinez and Labib [13] studied the NDVI values that respond to various vegetation types (trees and grass). They found a nonlinear relationship between the average NDVI and specific types. Additionally, increasing NDVI is relevant to varying vegetation changes. In their study, the assessment of NDVI is prioritized to improve urban health.

Satellite imagery is important in detecting land cover changes, especially in vegetation monitoring. This aspect has become increasingly important in scientific research because it addresses the challenges of stressed and uncultivated areas significantly affected by climate change. Therefore, mapping vegetation patterns is essential for tracking changes in vegetated areas and identifying non-vegetated, stressed, or healthy areas. Accurate vegetation detection is critical for various applications, including crop yield estimation, land cover and land use monitoring, urban growth assessment, and drought tracking [5]. Classification techniques, such as machine learning (ML) [14, 15], random forests (RF), support vector machine (SVM), gradient tree boosting (GTB), and classification and regression tree (CART) [16, 17], are used to classify satellite images into stressed, healthy, and non-vegetation areas. On the other hand, clustering techniques such as thresholding and K-means clustering are also used to categorize vegetation areas [6, 17, 18]. Recently, researchers have utilized Deep Learning (DL) to categorize vegetation areas [19]. In 2023, Kharat et al. [6] applied the RF classification and K-mean clustering methods to remote sensing data for recognizing land-cover types (vegetation, water, and soil). In their work, three metrics are used to evaluate the outcomes of K-means clustering: the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Based on the Silhouette Score metric, the achieved clustering performance was 46%. On the other hand, the RF classifier achieved 88% accuracy. Ayhan et al. [19] studied DL and traditional vegetation detection approaches using DeepLabV3 + a

custom CNN. The training and testing datasets were collected from different geographical locations, and the images are of different resolutions [20]. They introduced a new object-based method using NDVI and ML methods. Their technique was adopted for high-resolution airborne color images containing both RGB and NIR bands, as well as images containing only the RGB bands. Using two Kimisala test datasets, the researchers contrasted the performance of the DL methods with that of the object-based detection approach. All the result values confirm that NDVI-ML has better results than DL methods. Iryanti et al. [18] studied landslide risk in Garut using satellite-derived vegetation data, NDVI, and the K-means clustering method. They identify six landslide-prone areas, labelling NDVI grid codes 3 and 4 as potential zones due to sparse vegetation. Their analysis, based on data from NASA′s Terra and Aqua satellites, maps high-risk points with GIS visualization. The K-means clustering method is chosen for its effectiveness with the limited dataset. The silhouette coefficient is used to assess the quality of clustering results. The resulting clusters are divided into three groups, with cluster one having a silhouette coefficient close to 1, indicating robust clustering.

Much research in the literature uses the NDVI to detect land cover, particularly vegetation areas, while few studies utilize fractal dimensions. Based on fractal theory, fractals serve as statistical measures of the irregularity and roughness of ground features. Fractal dimensions help to identify vegetation cover through its spatial morphology and offer valuable insights into the spatial complexity of ecosystems by focusing on complex distribution patterns. The main objective of this study is to detect vegetation cover areas by combining satellite images using the Natural Vegetation Index, fractal dimensions, and land surface temperature (LST) matrices. K-means clustering and threshold-based labelling are employed to analyze the Landsat remote sensing dataset and assess vegetation health. The analysis will categorize the vegetation areas into healthy, stressed, and non-vegetation. The performance of k-means and threshold-based labelling will be evaluated using the Silhouette Score, Davies-Bouldin Index, and Entropy. Based on these metrics, the effectiveness of the two clustering methods will be explored and compared. Additionally, selecting the appropriate feature or combination of features that influences the clustering of vegetation areas will be examined. In this research, K-means and thresholding methods are utilized as complements to each other since both of these methods have their useful aspects based on the treatment of vegetation species classification. Thresholding-based is suitable when interpretable, explicit conditions on the boundary are known for simplicity and interpretability, but K-means offers flexibility in overlapping or ambiguity in clusters. The joint use of these methods provides a solid approach to the classification of the health status of vegetation on satellite images.

The remainder of the paper is divided into three sections. Section 2 explains the study′s methodology, including features, collection methods, and evaluation methods. Section 3 describes the dataset used. Section 4 presents the experimental results and evaluation. Finally, we conclude with Sect. 5.

2 Methodology

This study aims to categorize the vegetation area into healthy, stressful, and non-vegetation regions. Consequently, we develop a model consisting of three stages. The first is the feature extraction stage, where three features are extracted from the remote-sensing images. The clustering stage is performed using K-means and threshold-based labelling. Finally, there is the evaluation stage. In this section, the stages of the developed model are described.

2.1 Feature Extraction

Extracting features from Landsat Surface Reflectance Images (TIF format) datasets involves calculating key features that provide insights into vegetation health, land surface complexity, and thermal characteristics. NDVI represents vegetation health, where higher values indicate dense vegetation. FD quantifies the complexity of spatial patterns in the landscape, which can reflect structural diversity. LST is calculated from the thermal infrared band and highlights surface temperature variations critical for ecological and agricultural assessments. Together, these features provide a comprehensive understanding of environmental dynamics, enabling efficient classification and analysis of land cover and vegetation health. In this section, some background about the three features extracted from remote-sensing images used in this work is presented, along with samples of these extracted features.

Fractal Dimension (FD): Any object in real life has a degree of complexity. Some have a low degree, and other objects have a high degree that the topological dimension cannot describe. This complexity motivates mathematicians to find other techniques (methods) to calculate it, which is known as the fractal dimension (FD). The fractal dimension offers very important tools for describing and analyzing complex objects. It is frequently employed to characterize structures that exhibit irregularities or self-similarity across various scales. Studies show that the fractal dimension may provide information about ecosystems′ spatial complexity by measuring vegetation distribution. Although there are several ways to calculate it, the box-counting method is the most straightforward and precise algorithm for estimating its value. It relates the number of boxes N(ϵ) needed to cover a shape with the side length of the box ϵ [12], so that FD is calculated using Eq. (1):

$$FD = lim_{\varepsilon \to 0} \frac{logN(\epsilon)}{log(1/\epsilon)} \tag{1}$$

To analyze vegetation, the satellite images are partitioned into grids of varying sizes to count the number of nonempty cells in each grid [10, 11, 21]. To characterize the vegetation area, the digital images are divided into N(ϵ) distinct, non-overlapping grids of size ϵ. The FD value is constructed by stepwise decreasing of ϵ, whose regression line slope is given as in Eq. (2):

$$FD \approx \frac{\Delta logN(\epsilon)}{\Delta log\epsilon} \tag{2}$$

In this work, we used the Fractal dimension to enhance the clustering method in vegetation analysis. It enhances the clustering accuracy by distinguishing between dense forests and fragmented landscapes. It also combines with machine learning models, such as K-means and threshold, to improve the classification of vegetation health states.

Normalized Difference Vegetation Index (NDVI): Remote sensing statistics are often used to assess the coverage, density, and health of vegetation in the Normalized Difference Vegetation Index (NDVI). The index uses vegetation′s spectrum reflectance properties, particularly in the red and near-infrared (NIR) bands. Depending on the cellular makeup of the leaves, vegetation and leafy plants absorb the reddest light (R) for photosynthesis and reflect a sizable amount of near-infrared (NIR) light. It is computed using Eq. (3) [3, 13]:

$$NDVI = \frac{NIR - R}{NIR + R} \tag{3}$$

where

- NIR refers to the reflectance in the near-infrared band,
- R is the reflectance in the red band.

Its value ranges from -1 to $+1$, such that.

- If NDVI > 0.5: it indicates healthy and dense vegetation.
- If $0.2 <$ NDVI ≤ 0.5: it indicates sparse or stressed vegetation.
- If NDVI ≤ 0.2: it indicates that it signifies bare soil, water bodies, or non-vegetated areas.

The NDVI can be used for:

1. *Vegetable Health Assessment:* The NDVI has a direct correlation with plant productivity and chlorophyll content. Strong photosynthetic activity is indicated by higher NDVI values, while stress or degradation is indicated by lower values.
2. *Classification of Land Cover:* By differentiating between vegetated and non-vegetated areas, NDVI makes mapping land use and land cover easier.
3. *Temporal Analysis:* NDVI helps with studies of deforestation, desertification, and crop production prediction by allowing the tracking of seasonal and long-term vegetation changes.

The NDVI is calculated based on reflectance data from satellite sensors like Landsat or Sentinel-2. For instance:

- Bands 4 (Red) and 5 (NIR) are used by Landsat 8.
- Bands 4 (Red) and 8 (NIR) are used by Sentinel-2.

Typically, NDVI values are represented as heat maps or classified for aggregation into groups for clustering or statistical analysis. Correlating NDVI with other parameters like FD or LST improves the accuracy of vegetation classification. It is a strong

index to study green vegetation on a large scale; due to its sensitivity to atmospheric content, soil background, and mixed pixels, which may require any sort of corrections or complementary indices.

Land Surface Temperature (LST): The temperature of the Earth′s surface is represented by the Land Surface Temperature (LST), which is calculated from thermal infrared (TIR) data acquired by remote sensing. Understanding vegetation stress, urban heat islands, land-atmosphere interactions, and climate research all depend on LST. According to Planck′s Law, LST is computed from the thermal radiation that the surface emits. Equation (4) provides the brightness $\boldsymbol{L_\lambda}$ at a given wavelength λ [22].

$$L_\lambda = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}-1}} \tag{4}$$

where:

- h is Planck′s constant $\left(6.626 \times 10^{-34}\,\mathrm{J}\cdot\,\mathrm{s}\right)$,
- c is the speed of light $\left(3 \times 10^8\,\mathrm{m/s}\right)$,
- k is Boltzmann′s constant $\left(1.381 \times 10^{-\Upsilon} \downarrow \mathrm{K}\right)$,
- T is the temperature (K).

The top-of-atmosphere (TOA) radiance, which satellite sensors collect for LST estimation, needs to be adjusted for emissivity and atmospheric influences. The LST is calculated through the following three steps:

1. *Conversion to Spectral Radiance:* The calibration constants of the sensor are used to calculate the TOA spectral radiance L *as in* Eq. (5):

$$L = \frac{L_{\max} - L_{\min}}{Q_{\text{calmax}} - Q_{\text{calmin}}}(Q_{\text{cal}} - Q_{\text{calmin}}) + L_{\min} \tag{5}$$

where:
Q_{cal} is the digital number,
$L_{\max}, L_{\min}$ are sensor-specific radiance limits,
$Q_{\text{calmax}}, Q_{\text{calmin}}$ are calibration values.

2. *Conversion to Brightness Temperature:* The inverse Planck function is used to determine the brightness temperature T_b as in Eq. (6).

$$T_b = \frac{K_2}{\ln\left(\frac{K_1}{L} + 1\right)} \tag{6}$$

where: K_1, K_2 The calibration constants are specific to the sensor, L is the spectral radiance.

3. *Surface Emissivity Correction:* Surface emissivity ϵ, is a gauge of a surface′s ability to emit heat radiation (considered while calculating LST (see Eq. (7)).

$$LST = \frac{T_b}{1 + \left(\frac{\lambda_b}{hc}\right)\ln(\epsilon)} \tag{7}$$

where ϵ depends on land cover type (e.g., vegetation, soil, water).

The LST can be utilized for climate monitoring by monitoring long-term and seasonal temperature patterns, vegetation analysis by identifying thermal stress in plants, and urban heat island detection by identifying temperature anomalies in urban regions. A sample of the Landsat Surface Reflectance Dataset after feature extraction is illustrated in Table 1. Table 2 shows the preprocessing of the extracted Features.

Table 1. Sample of Landsat Surface Reflectance Dataset after Feature Extraction

File	NDVI	Fractal Dimension	LST
10000.tif	0.044197019	1.935009224	147.8502568
10001.tif	0.030440027	1.758902982	147.8973872
10002.tif	−0.11129284	1.817798437	147.8878181
10003.tif	−0.02735228	1.928295604	147.8711278
10004.tif	0.029509578	1.863981182	147.8735925
10005.tif	−0.03912205	1.913426392	147.8767818
10006.tif	−0.02609288	1.99312674	147.8746662
10007.tif	−0.01163762	1.948725655	147.8967401
10008.tif	0.021041805	1.951532602	147.896794
10009.tif	−0.03101563	1.910427572	147.8917216
10010.tif	−0.07040081	1.939831155	147.8970637
10011.tif	0.09750851	1.929277217	NAN
10012.tif	−0.1307911	1.942162104	147.8854531
10013.tif	0.003484572	1.869693652	147.8575551

2.2 Stage Two: Clustering of Vegetation Area

Threshold-Based Labeling Method: Using predetermined domain-specific thresholds, the threshold-based labelling is a clustering technique that groups data points into meaningful classes according to the attributes of individual features. This method starts by examining the values of a particular feature, like the NDVI, LST, or FD, and then clustering the data according to the mean feature values.

Table 2. Preprocessing of the Extracted Features

File	NDVI	Fractal Dimension	LST
10000.tif	0.496425	0.876778	0.039139
10001.tif	0.489106	0.796982	0.894325
10002.tif	0.413696	0.823668	0.720693
10003.tif	0.458357	0.873736	0.417845
10004.tif	0.48861	0.844594	0.462568
10005.tif	0.452095	0.866999	0.520438
10006.tif	0.459027	0.903112	0.482049
10007.tif	0.466718	0.882993	0.882583
10008.tif	0.484105	0.884265	0.883562
10009.tif	0.456408	0.86564	0.791523
10010.tif	0.435452	0.878963	0.888454
10011.tif	0.52479	0.874181	0.600507
10012.tif	0.403321	0.880019	0.677779

This method applied several steps:

1. *Feature-Based Clustering*: The input for clustering is a feature of interest (such as FD, LST, or NDVI). The feature data is divided using the K-means algorithm into a predetermined number of clusters, k, each of which contains data points with comparable feature values.
2. *Cluster Analysis:* The mean value of the feature within each cluster is calculated once the clusters have been created. This analysis reveals the underlying properties of the data by showing the central tendency of the feature within each cluster.
3. Threshold Assignment: The application of previously predetermined thresholds and domain expertise to label each cluster of data points with a label that has meaning and context.

   For example:

   - The NDVI clusters are labelled as:
     - *Healthy vegetation* if the mean NDVI > 0.5.
     - *Stressed vegetation* if $0.2 <$ mean NDVI ≤ 0.5.
     - *Non-vegetated areas* if the mean NDVI ≤ 0.2.
   - Based on LST, clusters are labelled based on temperature ranges:
     - *High Temperature* for mean values > 30 °C.
     - *Moderate Temperature* for values between 15 °C and 30 °C.
     - *Low Temperature* for values ≤ 15 °C.

- Based on Fractal Dimension, clusters are labelled as:
  - *Complex Structures* if the FD value exceeds 1.5.
  - *Simple Structures,* otherwise.

4. *Label Assignment:* Each feature′s threshold values are mapped to its corresponding cluster, producing useful categorical labels that explain the physical or biological significance of the feature.

K-Means Clustering: Works perfectly well on datasets that have well-defined thresholds unique to the domain for any feature. Despite its simplicity, threshold-based labelling provides a realistic way to classify land surface elements, such as the status of vegetation, using predetermined thresholds and inspecting characteristics of clusters. Its main strength is described in ***Algorithm_one,*** which shows the transparency and the potential to provide perceptual labels that correspond to real-world phenomena that are indispensable for vegetation monitoring and ecological studies [23].

Algorithm_One: K-means Clustering with Meaningful Labeling
Input:

- Dataset D with features $F = \{NDVI, FD, LST\}$
- Number of clusters $k = 3$
- Predefined thresholds for each feature

Output:

- Dataset D' with cluster labels and meaningful descriptive labels.

Algorithm Steps:

1. Initialization: The feature values V_f is used to cluster the satellite images n the feature $f \in F$.
2. Clustering Process: The K-means algorithm is applied to split V_f into three clusters:
   a) Randomly initialize k centroids.
   b) Each data point in V_f is assigned to the nearest centroid using Euclidean distance.
   c) The centroid is recalculated as the mean of all data points in each cluster.
   d) Repeat steps 2.b and 2.c until centroids stabilize or maximum iterations are reached.
   e) Compute Cluster Statistics: For each cluster C_i ($i = 1,2,3$):
   f) Calculate the mean value μ_i of feature f in C_i.
3. Assign descriptive labels: Use μ_i and predefined thresholds to assign a descriptive label L_i to each cluster C_i:
   If f = NDVI:
   - $\mu_i > 0.5 \rightarrow L_i =$ "Healthy"
   - $0.2 < \mu_i \leq 0.5 \rightarrow L_i =$ "Stressful"
   - $\mu_i \leq 0.2 \rightarrow L_i =$ "Non-Vegetated"

If f = LST:

- $\mu_i > 30 \rightarrow L_i =$ "High Temperature"
- $15 < \mu_i \leq 30 \rightarrow L_i =$ "Moderate Temperature"
- $\mu_i \leq 15 \rightarrow L_i =$ "Low Temperature"

If f = Fractal Dimension:

- $\mu_i > 1.5 \rightarrow L_i =$ "Complex Structures"
- $\mu_i \leq 1.5 \rightarrow L_i =$ "Simple Structures"

4. Update Dataset: Add two columns to D for feature f:
   $Cluster_f$: Cluster labels (numeric).
   $Label_f$: Descriptive labels (textual).
5. Repeat steps 1–4 for each feature $f \in F$.
6. Output: The updated dataset D' where the cluster labels and descriptive labels are described for each feature.

This algorithm provides descriptive labels for each data point in the dataset, mapping the cluster assignments to intelligible categories. These indices are added in additional columns to be saved as labelled datasets, after which some statistical analyses are implemented to provide information like mean feature values within each cluster. These statistics are shown in a summary table that connects cluster means to the descriptive labels.

Environmental characteristics like vegetation health, structural complexity, and thermal characteristics are efficiently grouped into interpretable clusters by this K-means algorithm implementation. The method improves the dataset′s usability for vegetation monitoring and land surface analysis by utilizing thresholds that are based on domain knowledge, offering researchers and practitioners useful information.

2.3 Stage Three: Evaluation Metrics

Silhouette Score: A statistical metric used to evaluate the quality of clustering by analyzing how accurately data points are allocated to their appropriate clusters. It measures how similar an object is to its cluster (cohesion) compared to other clusters (separation). The Silhouette score does not require previously classified data. This measure captures the clustering′s variance, separation, and compactness [24]. The formula of the Silhouette score is illustrated in Eq. (8).

$$\text{Sil} = \frac{b - a}{\text{Maximum}(b, a)} \tag{8}$$

where a is the average distance within the same cluster, and b is the average distance to all other clusters. Better clustering is indicated by a value close to 1. Conversely, the worst-case scenario is represented by a value of -1, and the overlapping clusters are indicated by a value of 0.

Davies-Bouldin Index (DBI): The DBI is a commonly used metric to assess clustering performance. It measures the compactness and separation of clusters, with lower DBI values indicating better clustering outcomes. The average similarity of each cluster to the most comparable cluster is evaluated by DBI [25]. Equation (9) represents the DBI formula:

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \tag{9}$$

where:

k: Total number of clusters,
S_i: Average intra-cluster distance for cluster$i$,
S_j: Average intra-cluster distance for cluster$j$,
M_{ij}: Inter-cluster distance between clusters i and$j$.

Entropy: A statistical metric that is used to express how diverse or unpredictable the label distribution is in each sample. In clustering, the entropy measures the homogeneity of the space (high entropy means no clustering, low entropy means high clustering) [26]. It reflects the unpredictability of assignments and assesses how equally the data points are divided among various labels or clusters in clustering or classification. The mathematical definition of entropy is represented in Eq. (10):

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{10}$$

where:

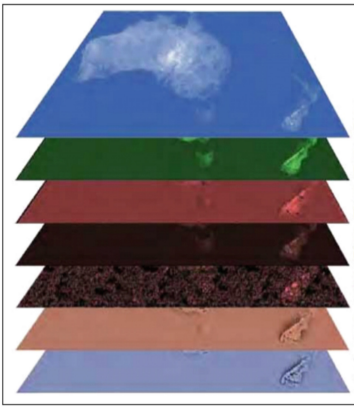
- H: Entropy, which refers to the diversity in the distribution of clusters,
- k: The number of clusters,
- p_i: percentage of cluster i data points. It is computed as:$p_i = \frac{n_i}{N}$, where n_i represent the number of points in the cluster$i$, and N is the total number of data points.

3 Dataset Description

In this work, we adopted the Landsat Surface Reflectance dataset [27]. The dataset consists of 17,148 atmospherically corrected satellite images. Each frame (or: scene) covers an area of 185 km × 185 km (about 186 km × 186 km at the equator) of the Earth - in all possible global directions, including Africa, South America, North America, Europe, Asia, Indonesia, and parts of Antarctica (during daylight). The data used is a collection of spectral bands surface reflectance images from the Landsat satellite in raster format (pixel by pixel). These bands are designed to record in particular sets of wavelengths and are important for vegetation and land surface studies. The main spectral bands that have been used in the Landsat missions are shown in Fig. 1. The purpose of each band is:

- Blue (0.45–0.51 μm), detects shallow water penetration in coastal surveillance.

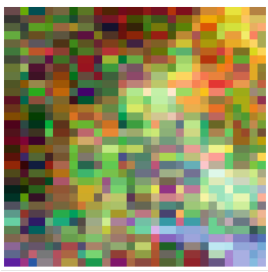
- Green (0.53–0.59 μm): sensitive to greenness of vegetation.
- Red (0.63–0.69 μm): (used in NDVI) to indicate vegetation health.
- Near-infrared (NIR) (0.85–0.88 μm): important for NDVI/EVI indices and biomass calculation.
- SWIR1/SWIR2 (1.56–2.29 μm): moisture and vegetation stress coverage.
- TIR (10.6–11.2 μm): developed to derive Land Surface Temperature (LST).



Purpose/ Utility
Penetrates shallow water, coastal observations
Detects water bodies, soil, and vegetation
Sensitive to vegetation greenness.
Used in NDVI, reflects vegetation health.
Critical for vegetation indices like NDVI/EVI.
Detects moisture content in soil and vegetation.
Sensitive to changes in soil and vegetation moisture
Land Surface Temperature (LST) calculations

Fig. 1. Band Designations for Landsat Satellites

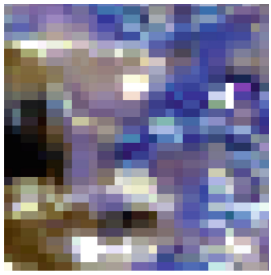
Computed bands allow us to calculate vegetation indices (e.g., NDVI), fractal dimensions, and LST, parameters important to our method for estimating vegetation health. Figure 2 provides an example of raster images (10000.tif, 10001.tif, and 10002.tif) from the dataset. These pictures illustrate how land cover, reflectivity level, and density of vegetation change over the different spectral bands, which may give a hint at the complex nature of the input data for unsupervised clustering.



10000.tif Image:



10001.tif Image



10002.tif Image

Fig. 2. Sample from Landsat_Surface_Reflectance Dataset

4 Experimental Results and Analysis

The results of this study provide a comparative analysis of the two clustering techniques (threshold-based labelling and K-means clustering) after applying them to the Landsat Surface Reflectance dataset. The methods were evaluated based on their ability to cluster vegetation health and land surface characteristics using NDVI, FD, and LST features. To evaluate and compare the clustering performance of both methods (threshold-based labelling and K-means clustering), three measures are utilized, including the Silhouette score (shown in Eq. 8), the Davies-Bouldin index (DBI) using Eq. (9), and the Entropy metrics using Eq. (10). In this section, the key results are accompanied by data visualizations to illustrate the clustering performance and the assigned labels.

4.1 Threshold-Based Labeling Method Description

For threshold-based Labeling, the results illustrated in Fig. 3 show the clustering performance across all features. First, NDVI, the Silhouette Score, and the entropy metrics show relatively high performance, indicating a good balance of cluster cohesion and separation. The DBI is comparatively lower, suggesting relatively well-defined clusters. For FD, the Silhouette Score shows the best clustering performance. However, the DBI is significantly low, indicating superior clustering with minimal clustering overlap. Lastly, the entropy metric conveys good cluster labeling. For LST, the Silhouette Score indicates low performance compared with NDVI and FD methods. To sum up, the FD feature shows the best clustering performance overall with the highest Silhouette Score, the lowest DBI, and the entropy measures. LST gives the most informative threshold label (highest Entropy), and NDVI provides balanced metrics, but overall, NDVI does not outperform the other features under any single metric. The analysis demonstrates that some features are stronger than others, and thus, one should choose features based on the desired metrics in clustering applications.

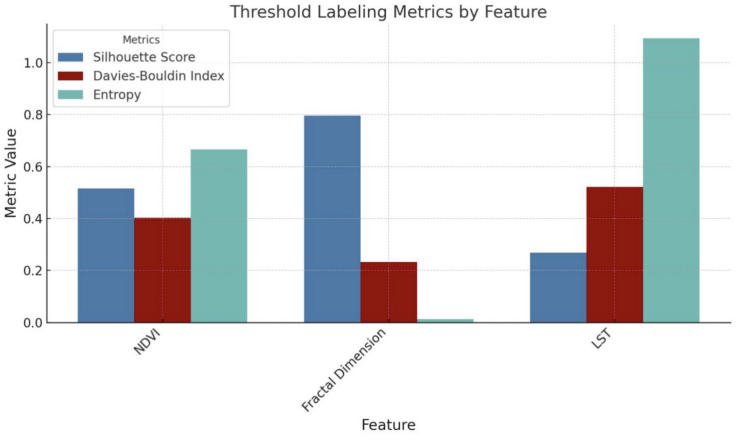


Fig. 3. Threshold Labeling Method: comparison

4.2 K-Means Clustering with Interpretable Clusters

The results illustrated in Fig. 4 show the clustering performance across all features using the K-means clustering. The clustering results of LST in terms of the Silhouette Score measure relatively low, indicating that the similarity of an object to its cluster (cohesion) compared to other clusters (separation) is poor. On the other hand, the DBI measures are relatively high, indicating high overlap and ill-defined cluster boundaries. Although the entropy is relatively low across features, it is considered medium, indicating that the elements within a cluster show average similarity. This confirms that LST has a low capacity in the cluster with a moderate ability to distinguish a clear cluster with cohesion.

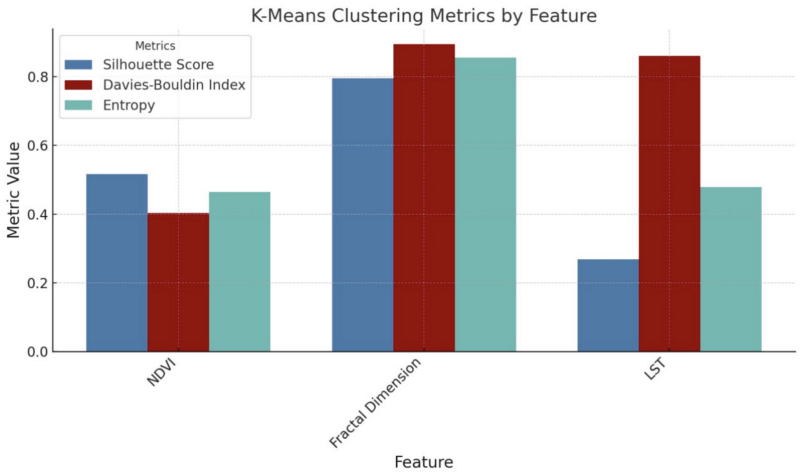


Fig. 4. Results of K-means Labels Across Features

The Silhouette Score for the FD is higher when compared to NDVI and LST (i.e., the clusters are more compact and separated for the FD). On the other hand, the DBI features the highest score among all and indicates high overlap and ill-defined cluster boundaries. The second fact is that the entropy is of high value, indicating a high changeability and diversity in cluster assignments [19]. These findings suggest that although FD is superior in cluster cohesion compared to NDVI, it has obvious difficulty in cluster separation. Thus, it cannot be recommended in the current study. The results for clustering are optimal for NDVI. It yields the best Silhouette Score overall features, indicating the best inter- and intra-cluster distance. The DBI is low, indicating well-separated and compact clusters. Entropy is moderate, which means there is a good amount of diversity of labels without over-mixing. The results showed that NDVI performed best with clear, cohesive, well-separated, meaningful clusters when K-means clustering is applied.

The comparison (K-means versus threshold-based labels) of clustering metrics, Silhouette Score, DBI, and entropy assesses both the cohesion and separation of clusters and the diversity of labels, as depicted in Fig. 5. This comparison highlights the importance of an appropriate selection of the features to be clustered and the right method

fitting to the executed clustering. To sum up, K-means and threshold-based label methods showed the same behavior in terms of compactness and cohesion (Silhouette Score). However, threshold-based methods outperformed K-means regarding cluster separation (DBI) and informativeness (Entropy) for NDVI and LST. The metric should be chosen based on the clustering goal. If your priority is information gain maximization, threshold-based labels are better. However, if you need well-defined clusters, perhaps K-means is the most appropriate one. To sum up, for clustering vegetation health into three categories, K-means clustering with NDVI and threshold-based labels with FD as features would probably perform best regarding compactness and definition of clusters, which corresponds with the goals of this study.

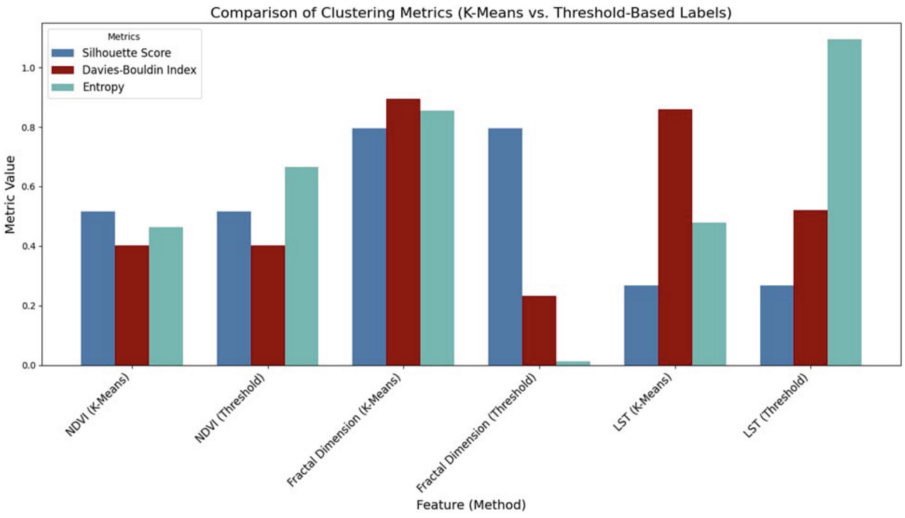
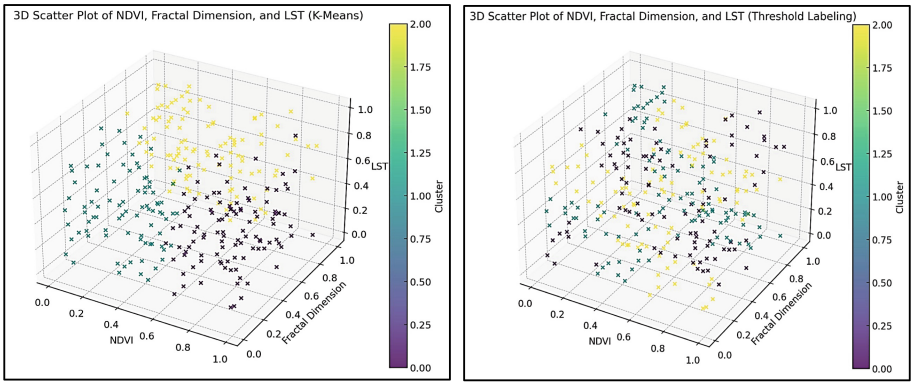


Fig. 5. Comparison of results between the two methods (K-means and threshold-based labels)

Figure 6 displays two scatter plots representing the clustering results of K-means and threshold-based labels by combining the features: NDVI, FD, and LST. The data points are color-coded based on their assigned clusters, with the color bar indicating cluster labels.

In terms of cluster boundaries, the clusters appear to be more cohesive and relatively separated in the feature space illustrated in Fig. 6(a), which represents K-means clustering. On the other hand, the threshold labeling clustering method shown in Fig. 6(b) depicts less cohesive clustering, and there is more overlap between clusters. Although the K-means method mathematically generates optimal clusters, the resulting cluster boundaries are less intuitive because they do not directly correspond to feature thresholds. Threshold-based clustering is easier to interpret, as cluster membership is explicitly attached to the feature ranges (e.g., NDVI < 0.3).

K-means clustering is well-formed and ensures reasonably separated clusters by defining boundaries that are updated based on variance in data. Therefore, it is a suitable method for identifying accurate vegetation health. It is also not explicit clustering



(a)Clustering with K-means **(b)Clustering with threshold-based labels**

Fig. 6. Scatter plots of clustering results

because we are not minimizing inter-cluster but intra-cluster variance. On the other hand, threshold labelling is logical and easy to interpret, associating clusters with given ranges of features, but it is also not flexible and may lead to overlaps if the distribution of data is far from the thresholds. K-means grouping, on the one hand, is more appropriate for depth, accuracy, and precision, partly to cover the cons of threshold, on the contrary, where simplicity and replicability are far more important than precision.

5 Conclusion

This study is concerned with identifying vegetation health by clustering satellite images using NDVI, FD, and LST as features. We applied K-means and threshold-based labeling to Landsat remote sensing data to classify vegetation into healthy, stressed, and non-vegetation categories. Our findings reveal that threshold-based labeling creates cohesive and well-separated clusters for FD, as shown by high Silhouette and low DBI scores along with low entropy. In contrast, K-means clustering exhibits greater label diversity for NDVI and LST than FD due to their better entropy values. Both clustering methods yield similar Silhouette Scores for the three features (NDVI, FD, and LST), indicating effective cluster cohesion and separation. These results highlight the strengths of each method: threshold-based labeling is ideal for features with defined thresholds, while K-means allow for adaptability in cases without such limits. FD with Threshold-based labels demonstrates superior clustering results, offering clear and interpretable outcomes, which are crucial for evaluating vegetation health and analyzing land surfaces. This analysis offers important insights for researchers and practitioners in vegetation health monitoring and environmental data analysis, helping them choose the most suitable clustering technique based on specific datasets and objectives. In future work, since FD proves its effectiveness in detecting vegetation, we will investigate the performance of FractalNet to classify the vegetation in satellite images in terms of accuracy. We also aim to compare the performance of FractalNet with convolutional neural networks (CNNs) in terms of accuracy and processing speed.

Disclosure of Interests. The authors have no competing interests to declare in this study.

References

1. Hannah_Ospina.: Ensuring a sustainable future through plant health. NL Netherlands (2024). https://nlplatform.com/articles/ensuring-sustainable-future-through-plant-health#:~:text=Healthy%20plants%20play%20a%20crucial,and%20secure%20a%20sus-tainable%20future
2. Heeb, L., Jenner, E.: Climate-smart pest management: implementation guidance for policymakers and investors. (FAO) Practice Brief Climate-Smart Agriculture (2017). https://openknowledge.fao.org/server/api/core/bitstreams/fd66f08d-ea43-4bb8-8fa0-bda26d694840/content
3. Mehmood, K., et al.: Analyzing vegetation health dynamics across seasons and regions through NDVI and climatic variables. Sci. Rep. **14**(1), 11775 (2024). https://doi.org/10.1038/s41598-024-62464-7
4. Samawi, V.W: SMCSIS: An IoT-based secure multi-crop irrigation system for smart farming. Int. J. Innovative Comput. Inf. Control (IJICIC) **17**(4) (2021). http://www.ijicic.org/ijicic-170408.pdf
5. Gu, Z., Zeng, M.: The use of artificial intelligence and satellite remote sensing in land cover change detection: review and perspectives. Sustainability **16**(1), 274 (2024). https://doi.org/10.3390/su16010274
6. Kharat, V., Khatdeo, S., Kothe, H., Kshirsagar, R., Dixit, M., Balan, M.S.: Land cover clustering and classification of satellite images. In: First International Conference on Data Science and Advanced Computing (ICDSAC 2023), ITM Web Conf. 56. EDP Sciences (2023). https://doi.org/10.1051/itmconf/20235602004
7. Thwal, N.S., Ishikawa, T., Watanabe, H.: Land cover classification and change detection analysis of multispectral satellite images using machine learning. Image Sig. Process. Remote Sens. XXV **111551O**, 11155 (2019). https://doi.org/10.1117/12.2532988
8. Khodadadi, N., et al.: Predicting normalized difference vegetation index using a deep attention network with bidirectional GRU: a hybrid parametric optimization approach. Int. J. Data Sci. Anal. (2024). https://doi.org/10.1007/s41060-0
9. Yang, H., Chen, W., Qian, T., Shen, D., Wang, J.: The extraction of vegetation points from LiDAR using 3D fractal dimension analyses. Remote Sens. **7**(8), 10815–10831 (2015). https://doi.org/10.3390/rs7081081
10. Yousif, S.A., Mohammed, A.J., Al-Saidi, N.M.G.: Texture images analysis using fractal extracted attributes. Int. J. Innovative Comput. Inf. Control **16**(4) (2020)
11. Yousif, S.A., Abdul-Wahed, H. ., Al-Saidi, N.M.: Extracting a new fractal and semi-variance attributes for texture images. In: AIP Conference Proceedings, vol. 2183, no. 1. AIP Publishing (2019)
12. Al-Azawi, R.J., Al-Saidi, N., Jalab, H.A., Ibrahim, R., Baleanu, D.: Image splicing detection based on texture features with fractal entropy (2021)
13. Martinez, A. D., Labib, S.: Demystifying normalized difference vegetation index (NDVI) for greenness exposure assessments and policy interventions in urban greening (2023). Environ. Res. **220** (2023). https://doi.org/10.1016/j.envres.2022.115155
14. Jehad, R., Yousif, S.A.: Fake news classification using random forest and decision tree (j48). Al-Nahrain J. Sci. **23**(4), 49–55 (2020)
15. Yousif, S.A., Samawi, V.W., Al-Saidi, N.M.: Automatic machine learning classification algorithms for stability detection of smart grid. In: 2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI), pp. 34–39. IEEE (2022)

16. Kwan, C., Gribben, D., Ayhan, B., Li, J., Bernabe, S.: An accurate vegetation and non-vegetation differentiation approach based on land cover classification. Remote Sens. **12**(23), 3880 (2020). https://doi.org/10.3390/rs12233880
17. Ouchra, H., Belangour, A., Erraissi, A.: Comparison of machine learning methods for satellite image classification: a case study of casablanca using landsat imagery and google earth engine. J. Environ. Earth Sci. **5**(2), 118–134 (2023). https://doi.org/10.30564/jees.v5i2.5928
18. Iryanti, M., Nurjanah, R., Waslaluddin, Arifin, M.: Landslide mapping using K-means cluster by NDVI data in Garut, West Java, Indonesia. In: The 1st International Seminar on Physics and Its Application, Phys.: Conference Series 2900 012020. IOP (2024). https://doi.org/10.1088/1742-6596/2900/1/012020
19. Ayhan, B., et al.: Vegetation detection using deep learning and conventional methods. Remote Sens. **12**(15) (2020). https://doi.org/10.3390/rs12152502
20. Interior, U.D.: USGS science for a changing world. Landsat 8 acquired from the U.S (2023). https://www.usgs.gov/landsat-missions/landsat-8
21. Al-Saidi, N.M., Abdul-Wahed, H.Y.: Classification of remote sensing images via fractal discriptores. In: 2018 International Conference on Advance of Sustainable Engineering and its Application (ICASEA), pp. 99–104. IEEE (2018)
22. Li, Z.L., et al.: Satellite remote sensing of global land surface temperature: definition, methods, products, and applications. Rev. Geophys. **61**(1), e2022RG000777 (2023)
23. Yousif, S.A., Al-Dulaimy, A.: Clustering cloud workload traces to improve the performance of cloud data centers. In: Proceedings of the World Congress on Engineering, vol. 1, pp. 7−10 (2017)
24. Abdulhameed, T.Z., Yousif, S.A., Samawi, V.W., Al-Shaikhli, H.I.: SS-DBSCAN: Semi-supervised density-based spatial clustering of applications with noise for meaningful clustering in diverse density data. IEEE Access (2024)
25. Ros, F., Riad, R., Guillaume, S.: PDBI: a partitioning Davies-Bouldin index for clustering evaluation. Neurocomputing **528**, 178–199 (2023). https://doi.org/10.1016/j.neucom.2023.01.043
26. Joel, B.: Swartz, an entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbours. Health Place **4**(1), 67–77 (1998). https://doi.org/10.1016/S1353-8292(97)00026-9
27. Landsat_Surface_Reflectance Dataset. https://www.kaggle.com/datasets/reymas-ter/hwsd-landsat-processed