frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Original Research
PUBLISHED 20 May 2025
DOI 10.3389/fbioe.2025.1430222

OPEN ACCESS

EDITED BY
Di Wu,
Southwest University, China

REVIEWED BY
Dulani Meedeniya,
University of Moratuwa, Sri Lanka
Qibin He,
University of Chinese Academy of Sciences, China

*CORRESPONDENCE
Ahmad Jalal,
ahmadjalal@mail.au.edu.pk
Hui Liu,
hui.liu@uni-bremen.de

RECEIVED 09 May 2024
ACCEPTED 15 April 2025
PUBLISHED 20 May 2025

CITATION
Naseer A, Almudawi N, Aljuaid H, Alazeb A, AlQahtani Y, Algarni A, Jalal A and Liu H (2025) Multi-modal remote sensory learning for multi-objects over autonomous devices. *Front. Bioeng. Biotechnol.* 13:1430222. doi: 10.3389/fbioe.2025.1430222

Multi-modal remote sensory learning for multi-objects over autonomous devices

Aysha Naseer[1], Naif Almudawi[2], Hanan Aljuaid[3], Abdulwahab Alazeb[2], Yahay AlQahtani[4], Asaad Algarni[5], Ahmad Jalal[1,6]* and Hui Liu[7,8,9]*

[1]Department of Computer Science, Air University, Islamabad, Pakistan, [2]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, [3]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran, Saudi Arabia, [4]Department of Informatics and Computer Systems, King Khalid University, Abha, Saudi Arabia, [5]Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia, [6]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, South Korea, [7]Guodian Nanjing Automation Co., Ltd., Nanjing, China, [8]Jiangsu Key Laboratory of Intelligent Medical Image Computing, School of Future Technology, Nanjing University of Information Science and technology, Nanjing, China, [9]Cognitive Systems Lab, University of Bremen, Bremen, Germany

Introduction: There has been an increasing focus on object segmentation within remote sensing images in recent years due to advancements in remote sensing technology and the growing significance of these images in both military and civilian realms. In these situations, it is critical to accurately and quickly identify a wide variety of objects. In many computer vision applications, scene recognition in aerial-based remote sensing imagery presents a common issue.

Method: However, several challenging elements make this work especially difficult: (i) Different objects have different pixel densities; (ii) objects are not evenly distributed in remote sensing images; (iii) objects can appear differently depending on viewing angle and lighting conditions; and (iv) there are fluctuations in the number of objects, even the same type, in remote sensing images. Using a synergistic combination of Markov Random Field (MRF) for accurate labeling and Alex Net model for robust scene recognition, this work presents a novel method for the identification of remote sensing objects. During the labeling step, the use of MRF guarantees precise spatial contextual modeling, which improves comprehension of intricate interactions between nearby aerial objects. By simultaneously using deep learning model, the incorporation of Alex Net in the following classification phase enhances the model's capacity to identify complex patterns in aerial images and adapt to a variety of object attributes.

Results: Experiments show that our method performs better than others in terms of classification accuracy and generalization, indicating its efficacy analysis on benchmark datasets such as UC Merced Land Use and AID.

Discussion: Several performance measures were calculated to assess the efficacy of the suggested technique, including accuracy, precision, recall, error, and F1-Score. The assessment findings show a remarkable recognition rate of around 97.90% and 98.90%, on the AID and the UC Merced Land datasets, respectively.

KEYWORDS

multi-modal, remote sensing, multi-objects, autonomous devices deep learning, computer vision, posterior probability, likelihood estimation, scene analysis

1 Introduction

The evolution during the last several years of remote sensing (RS) technologies, in terms of platforms, sensors, and information support, has led to the significant increase in availability of EO data for geospatial analysis. Semantic segmentation is a key task in RS that has implications for a diverse area of applications, such as land use classification (Ahmad et al., 2020; Alazeb et al., 2023), changes detection (Islam et al., 2016; Saha et al., 2019), and environment surveillance (Srivastava et al., 2012; Jalal et al., 2021). However, these fine segmentation algorithms are not practically achievable due to the unavailability of large-scale labeled data and more critically, the quality of the labeled data is poor due to coverage restrictions of the satellite sensors particularly over different geographical terrains and various types of land. Conventionally every pixel in satellite imagery corresponds to a wide geographical region (Naseer et al., 2024a); therefore, pixel-level annotation is not only time consuming but also financially expensive and requires high levels of specialization (Manfreda et al., 2018; Khan et al., 2020).

To this end, there is a trend among scholars trying to employ semi-supervised methods which can minimize the amount of pixel-wise labeling, while using the gray images to extract information (Guo et al., 2020; Liu et al., 2017). Nevertheless, prevalent techniques entail biases in the procedure where the labeled data is selected at random, which in turn results in the massive formation of skewed models and innate inferior performance (Han et al., 2017; Muhammad et al., 2018). Second, segmentation in the remote sensing has problems that are closely related to characteristics of images such as different scales of the objects are caused shooting angles, inequality of counts of various objects, a large number of small objects in the aerial images and etc. Problems of truncation and occlusion make recognition even more challenging (Martin, 2011). Developing on previous attempts at employing remote sensing, computational methods from the past relied on statistical and rule-based methodologies, for instance, the decision tree algorithms or unsupervised clustering; modern deep learning techniques are scalable and more vigorous.

Earlier, classification based on remote sensing data used simple statistical and rule based techniques including decision tree algorithms and unsupervised clustering methods to classify land cover and land use. For example, the research in (Jayanetti et al., 2017) showed how land cover data can be combined with vGI such as Foursquare to land use. However, such methods have limitations associated with requiring the definition of certain thresholds and assumptions which reduces accuracy for complex terrain. Taking inspiration from these methods, deep learning-based methods have just lately been developed as superior and more scalable ways to handle remote sensing jobs. For instance, a model for categorizing suitable land for agriculture based on geographic mapping using deep learning was proposed by (Meedeniya et al., 2020), Deep Learning for Sustainable Agriculture, which greatly increased the forecast accuracy for paddy fields. In the same way, U-net and Fully Convolutional Networks (FCN) were used by (Mahakalanda et al., 2022), Deep learning-based prediction for rubber plantations, to identify the land use of rubber plantations. This process took very little time and achieved a very high accuracy of 94.13%, increasing its applicability in crop monitoring.

The study tackles a number of significant research issues in the fields of image segmentation and remote sensing, such as:

- How can object segmentation in remote sensing images be improved to enhance accuracy and efficiency in identifying multiple objects?
- What role do contextual and spectral-spatial features play in improving scene classification?
- How can AlexNet, when integrated with effective segmentation methods, improve the recognition of complex scenes in remote sensing imagery?
- What is the comparative performance of the proposed hybrid system against existing state-of-the-art methods on benchmark datasets?

In this article, the authors introduce an approach for dealing with the key issues in remote sensing scene recognition. While having several limitations, our approach mitigates those issues as follows: First, it minimizes interference and retains crucial aspects of appearance; second, it addresses the noise aspect and relevant spectral—spatial dependencies. Key contributions of this work include:

- Robust segmentation methods: A time comparison of MRF, FCM, and DBSCAN clustering with MRF as the best performer within the time-space complexity.
- Feature integration for scene recognition: Application of both spectral and spatial information along with the Haralick texture higher-order statistical measures in order to improve the segmentation coherency and correctness.
- Deep learning integration: Use of AlexNet for making use of segmented data in order to enhance the kind of recognition of the complicated scenes with much more details and precision.
- Comprehensive evaluation: Comparison with other methods to recognize the proposed hybrid system using newly established benchmarks AID or UCM, with accuracy rates of 97.90% and 98.90% respectively.

The subsections that follow in this paper follow this format. In Section 2, the body of extant literature is examined in detail. A detailed description of our suggested approach, including segmentation, labeling, feature extraction, and their combination, is provided in Section 3. In Section 4, a thorough analysis of the datasets used, the experimental design used, and the resulting results are discussed. Finally, Section 5 presents the results of the research we did.

2 Literature review

We reviewed the literature in a variety of disciplines, including object classification, segmentation, labeling, and scene classification, in order to analyze the complexity of aerial and remote sensing images. This helped us create the right dynamics and metrics for our strategy.

2.1 Multi-object segmentation

In remote sensing image processing, segmentation is a crucial activity that attempts to divide an image into similar regions.

Semantic segmentation specifically aims to allocate every homogenous region to a unique geographical object category, such as cities, farms, or woods. Semantic segmentation has been approached through a variety of approaches in recent decades, including Markov random field (MRF) models (Zhang et al., 2018). Level sets (Ball and Bruce et al., 2007). Clustering (Ma and Yang, 2009; Liu et al., 2015) and deep learning. Traditionally, early techniques such as clustering anticipated that pixels between distinct objects would display distinguishing traits, whereas pixels representing the same object would have identical characteristics. This method works well for low-to medium-resolution images (Zhao et al., 2019), but is ineffective for high-resolution (HSR) remote sensing images. Within an object in a HSR image, individual pixels may have distinct looks, but certain pixels from different objects may have identical (Längkvist et al., 2016) properties. Optimizing a multi-kernel method designed for semantic segmentation in high spatial resolution remote sensing imagery using an advanced Markov Random Field model effectively improves the robust classification of resilient objects in remote sensing images (He, 2024) in his work proposes the Grouping Prompt Tuning Framework (GoPT) based on semantic grouping for multi-modal image segmentation. This results in the original few-shot learning method with only one percent trainable parameters, and each new prompt tuner method brings state-of-the-art performance across multiple multi-modal segmentation tasks. This work reveals the possibility that efficient training of the foundation models implements early learning to address the multi-modal perception issues of weak transfer and scarce labeled data. To address domain gaps in remotely sensed segmentation tasks, authors (Wang et al., 2024) this work introduces Dynamic Loss Correction (DLC), a novel approach. Therefore, in order to apply machines to cross-domain scenarios, DLC adaptively adjusts loss functions to help establish the correspondence between related feature distributions across domains. This method also improves the model precision in terms of segments by stabilizing ephemeral alignment between features of different sources of data thus recommended for use where data of different types used.

2.2 Multi-object recognition

There are several challenges in the field of object classification for researchers. These difficulties include issues like localizing objects (Jain and Anto, 2022; Wang et al., 2022), recognizing and analyzing object connections (Sumbul et al., 2019) recovering hidden features, and classifying objects to produce desired outcomes. The widely accepted bag-of-words (Wang et al., 2017) technique has been the prevalent and effective framework for the classifying and recognizing (Naseer et al., 2024b) of objects in modern times. The bag-of-features approach has been the subject of several remarkable investigations (Ghabrani et al., 2023) present a new approach to classify land cover using spatial information derived from statistical properties of complicated CP and QP SAR data. They use super pixels to represent local spatial relationships and a built graph to express global dependencies. In order to estimate the land cover categorization image, labels are propagated from labeled to unlabeled super-pixels.

Furthermore, in a different study (Ahmed and Jalal et al., 2024), presented a unique illustration method designed for certain object classes. The characteristics of each image category were initially defined by a Gaussian mixture model (GMM). They constructed representations for comparison using the Euclidean distances between the pictures and these GMM models. Class-specific characteristics and visual components might be used to convey an image owing to the concatenation of these representations across all classes. A useful method for using multi-object categorization to identify indoor-outdoor situations is described (Ahmed and Jalal et al., 2024). Entails the process using two different ways to segment imagery, after which multiple kernel learning (MKL) will be employed to classify objects. To improve the classification, this procedure combines area-specific signatures with local descriptors. An approach was presented by (Ansith and Bini, 2022) classified land usage in high-resolution remote sensing images using a modified GAN architecture based on an encoder. The suggested technique feeds a latent vector into the generator after it has been generated by an encoder. Images from high-resolution remote sensing datasets are fed into the encoder. Support Vector Machines (SVM) are recommended by (Sangeetham and Sanam, 2023) as a method for identifying high-resolution images. Principal Component Analysis (PCA) is used to extract features from the pictures prior to classification. Support vector machines are then used to classify the feature vectors that are produced. Detecting an object's contours and motion. However, there is more possibility for incorporating various remote sensing visualization approaches into image processing because to the present high spatial resolution of remote sensing (RS) images and the decreased difference between RS and natural images (He et al., 2024) in their work, Orientation-Aware Multi-Modal Learning for Road Intersection Identification and Mapping, is centered on the association of orientation-learning to map road intersections using multi-modal data such as LiDAR and imagery. This framework merges spatial and geometric orientation characteristics in the fusion of multi-modal data in order to increase the accuracy of the maps. These methods highlight the need to develop frameworks which directly incorporate modality-specific spatial and geometric characteristics to enhance real-world application performance.

3 Materials and methods

3.1 System methodology

The basic process of this model begins with the identification and classification of objects seen in images obtained via remote sensing. For image segmentation, it makes use of techniques like FCM, MRF, and DBSCAN clustering. The advantages of MRF in terms of timing efficiency and segmentation accuracy led our team to select it. Our dedication to optimizing processing efficiency while guaranteeing precise object identification in remote sensing imagery is shown in this choice. Using feature extraction, properties including texture, spectral features, and SSF are extracted from the labeled objects. The model can produce more reliable and accurate scene recognition results by fusing AlexNet's feature learning skills with the contextual information from MRF. In order to offer a graphic depiction of our system's complex

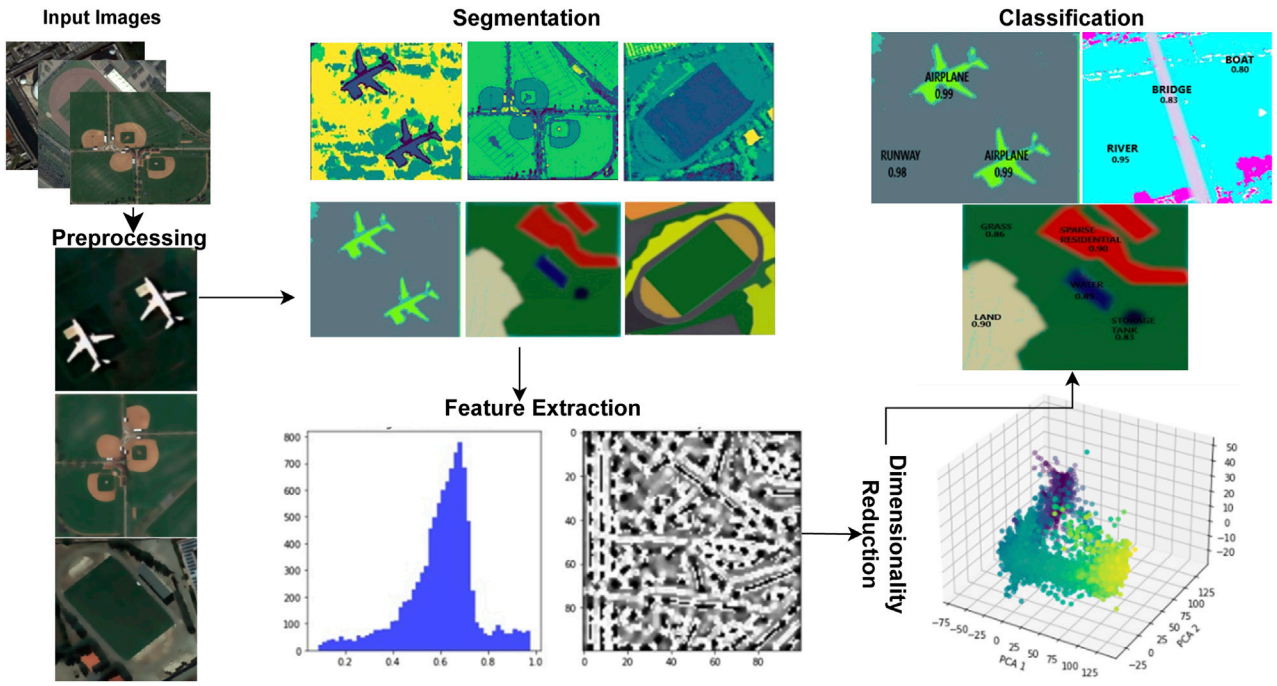

FIGURE 1
The hierarchical view of the suggested model over remote sensory images.

FIGURE 2
Outcomes of filtered images using Bilateral filter over some images of both datasets.

hierarchical structure, Figure 1 describes the hierarchical perspective that encompasses the complex elements and features of our OSC model.

3.2 Noise removal

The bilateral filter is a well-liked image processing method that may be used for a variety of operations, such as edge preservation, noise reduction and smoothing. Whtaen pursuing filtering, the bilateral filter (Tripathi and Mukhopadhay, 2012) considers both the spatial distance and the intensity difference between pixels. This dual-domain method enables it to blur an image while maintaining crucial edges and precise details. The weights in the weighted average of neighboring pixels computed by the bilateral filter rely on both the spatial and intensity distances. The weighted average for each pixel may be written mathematically (Hu et al., 2023) as given in Equations 1-3:

$$I(x,y) = \frac{1}{W(x,y)} \sum_{p \in N(x,y)}^{1} I(p).G_s\left(\|p-(x,y)\|\right)G_r\left(\|I(p)-I(x,y)\|\right) \quad (1)$$

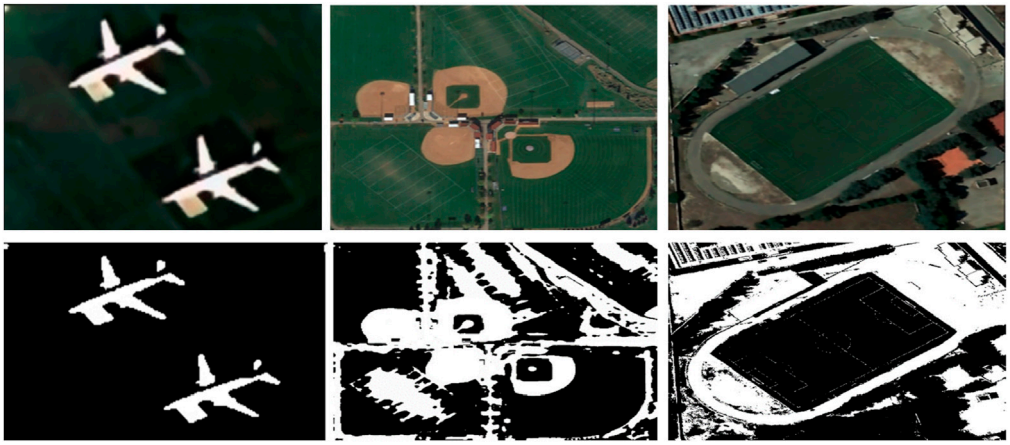

FIGURE 3
Fuzzy C-mean segmentation over some images from the UCM Dataset (row 1) represents the filtered images (row 2) demonstrates the segmented images.

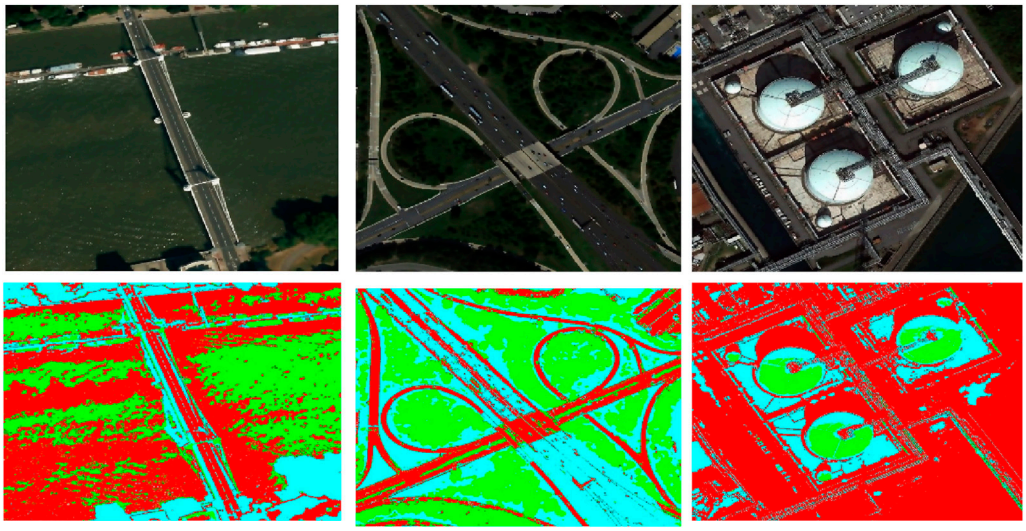

FIGURE 4
Density based clusters over some images from the AID Dataset (row 1) represents the filtered images (row 2) demonstrates the segmented images.

where I (x, y) is the intensity value of the pixel being filtered, the normalization factor W (x, y) makes sure the total weight adds up to 1. N (x, y) represents neighboring pixels around (x, y).

$$G_s\|p-(x,y)\| = exp\left(-\frac{\|p-(x,y)\|^2}{2\delta_s^2}\right) \quad (2)$$

$$G_r\|I(p)-I(x,y)\| = exp\left(-\frac{\|I(p)-I(x,y)\|^2}{2\delta_r^2}\right) \quad (3)$$

where "p" represents the neighboring pixel location, δ_s spatial standard deviation, and δ_r range standard deviation. The filtered result is shown in Figure 2.

3.3 Semantic segmentation

In order to simplify image representation for analysis, segmentation is dividing an image into homogenous and significant parts. The goal is to produce regions with comparable visual features, such as color or texture. In contrast, labeling reveals the meaning or class of each segment by assigning semantic labels to each one that was produced from segmentation. Segmented Result by all utilized techniques are shown in Table 1.

3.3.1 Fuzzy C-mean segmentation

This section explains the Fuzzy C-Mean (FCM) segmentation process. It begins by using pixels as data points to detect similar

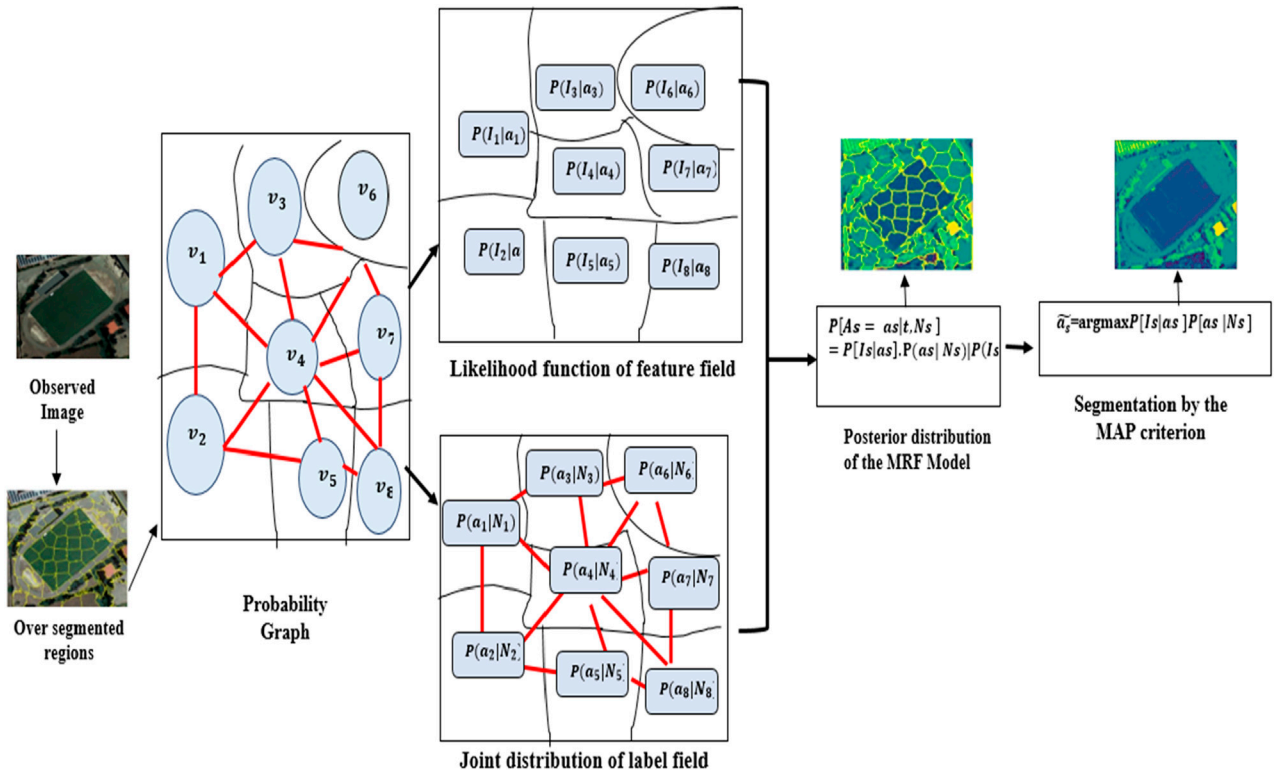

FIGURE 5
Illustration of the MRF model over image segmentation.

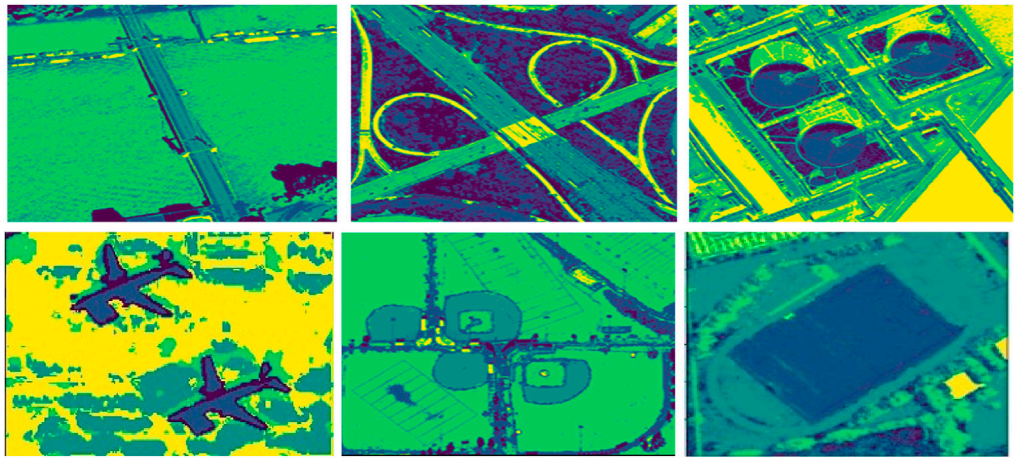

FIGURE 6
MRF-based segmentation over remote sensing images from both dataset (Row 1) AID dataset (Row 2) UMC dataset.

components. The pixel is subsequently assigned to numerous clusters instead of just one using fuzzy logic (Zhou, et al., 2023), producing a fuzzy assignment. To get the desired result, the objective function in FCM is iteratively optimized (Naseer and Jalal, 2023). To deconstruct the image, this iterative procedure entails changing membership degrees and clustering centers (Chen, et al., 2018). Performance index H_{FCM} is formulated using Equation 4.

$$H_{FCM} = (Q,S) = \sum_{i=1}^{r}\sum_{b=1}^{N} z_{ib}^{t} \| q_b - s_i \|^2, 1 < t < \infty \quad (4)$$

where "r" is the set size of clusters, N is the size of pixels, q_b is the bth pixel, s_i is the ce nter of the ith cluster, and t is the blur exponent. Each cluster center and membership function are updated using Equations 5, 6.

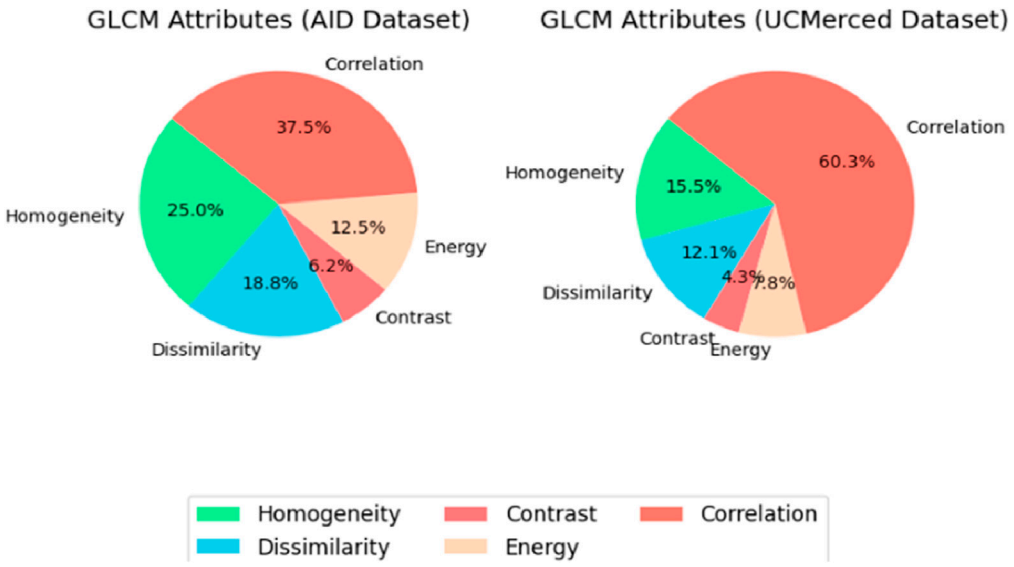

FIGURE 7
Distribution of GLCM attributes in AID and UC merced datasets.

TABLE 1 Comparison of computational time and object segmentation accuracy.

Datasets	Computational time			Segmentation accuracy (%)		
	DBSCAN	FCM	MRF	DBSCAN	FCM	MRF
UCM	162.13s	165.10s	**148.13s**	86.65	89.32	**91.18**
AID	175.30s	140.15s	**142.25s**	89.50	90.43	**91.67**

DBSCAN, Density Based Cpatial Clustering; FCM, Fuzzy Cmean; MRF, Markov Random Field; UCM, UC Merced; AID, Aerial Image Dataset. Bold values indicates proposed results (highlighed).

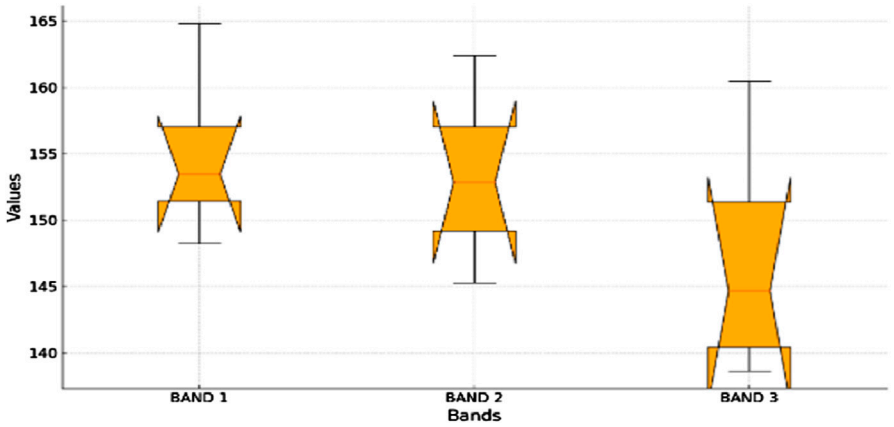

FIGURE 8
Box plot representing distribution of mean values across the different bands for all objects.

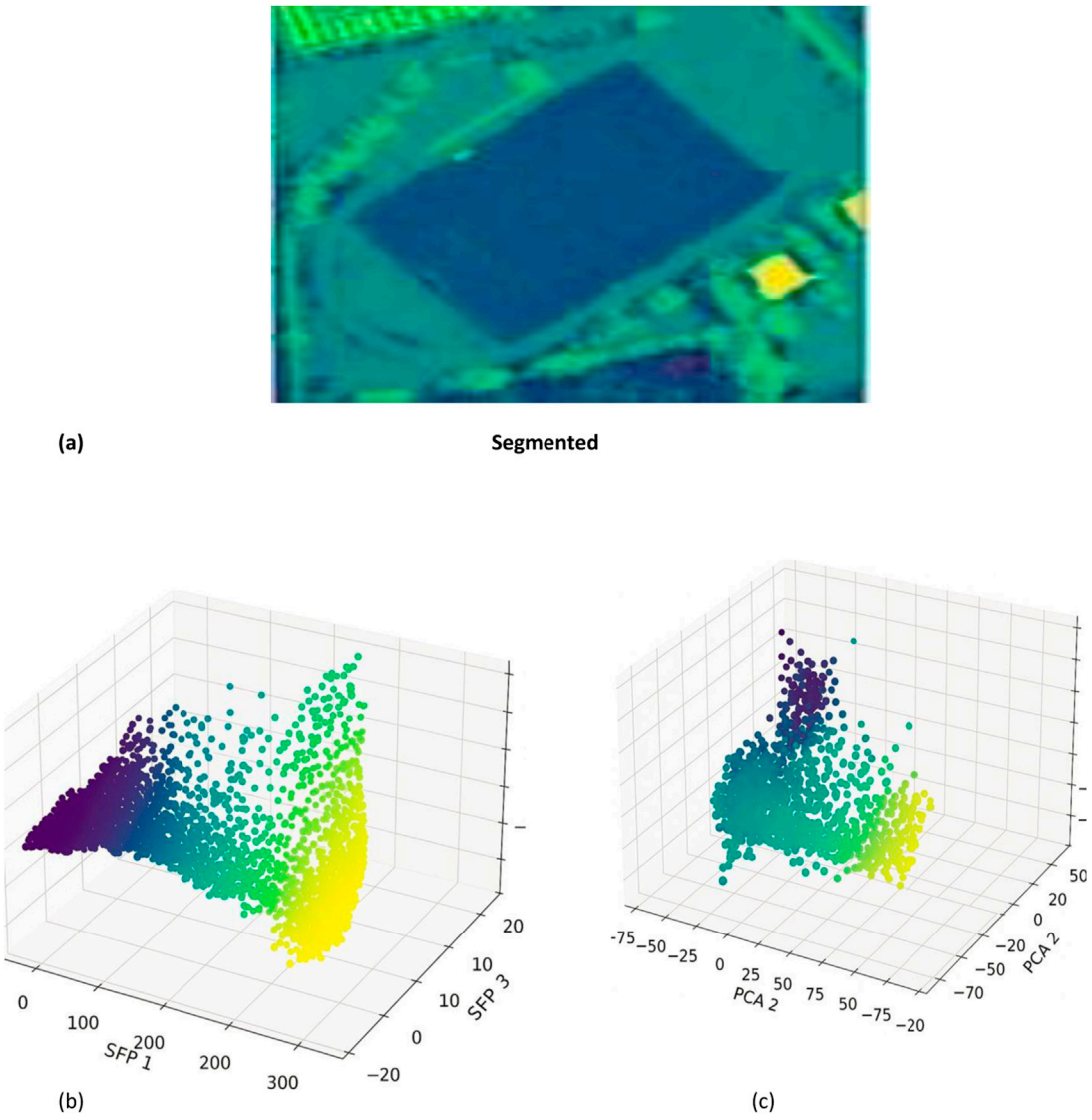

FIGURE 9
Dimensionality reduction **(a)** segmented **(b)** corresponding SSF extraction **(c)** PCA.

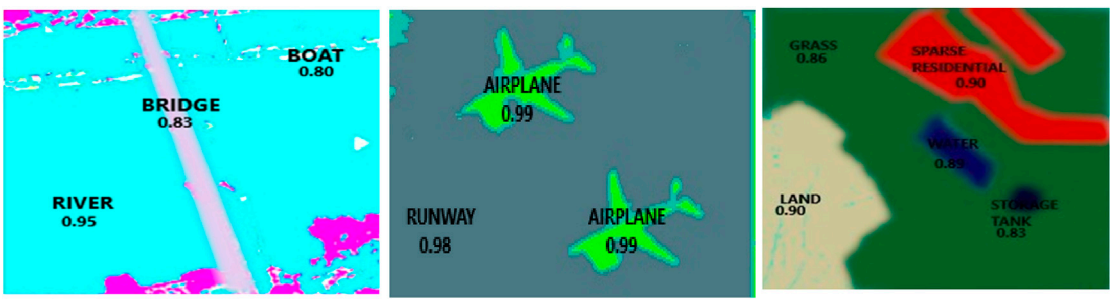

FIGURE 10
Jaccard jaccard similarity on multiple objects over some images.

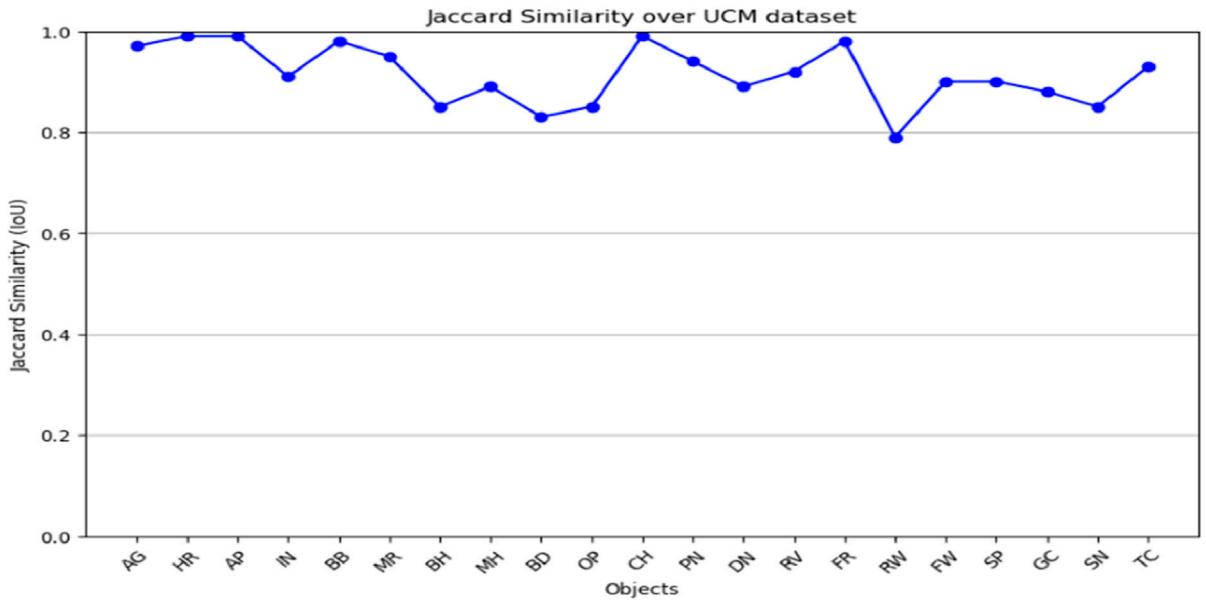

FIGURE 11
Jaccard similarity graph shown over UCM dataset.

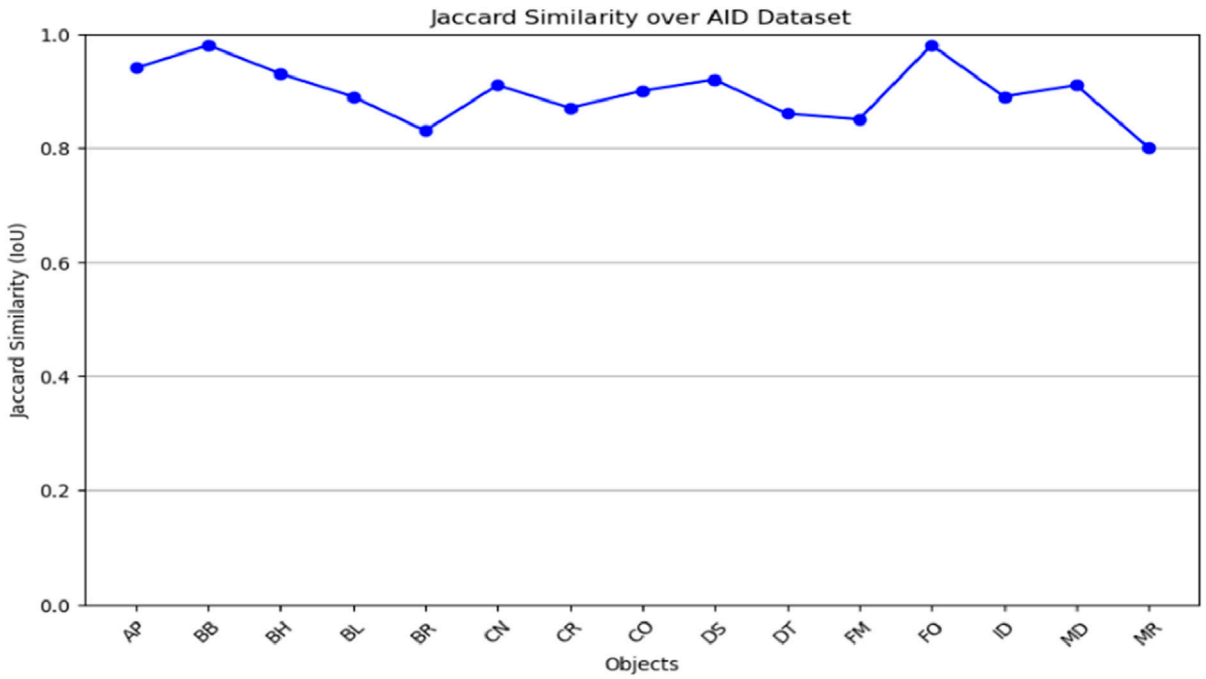

FIGURE 12
Jaccard similarity graph shown over AID dataset.

$$z_{ib}^{t} = \frac{1}{\sum_{j=1}^{c}\left(\frac{1}{(t-1)}\right)} \tag{5}$$

$$s_j = \frac{\sum_{k=1}^{N} z_{ib}^{t} q_b}{\sum_{k=1}^{N} z_{ib}^{t}} \tag{6}$$

where $z_{ib} \in [0,1], for\, b = [1,\ldots,c]$; l_{ib}^{2} represents the distance between pixel q_b and cluster centroid s_i and z_{ib}^{t} stands for the membership matrix that belongs to [0, 1]. According to (Halder et al., 2011), FCM gives pixels close to the center of their class high membership values, whereas pixels far from the center receive low membership values. As demonstrated in Figure 3, which shows the segmented results of images from the UC Merced dataset, this processing complexity is applied to every nearby pixel in the images.

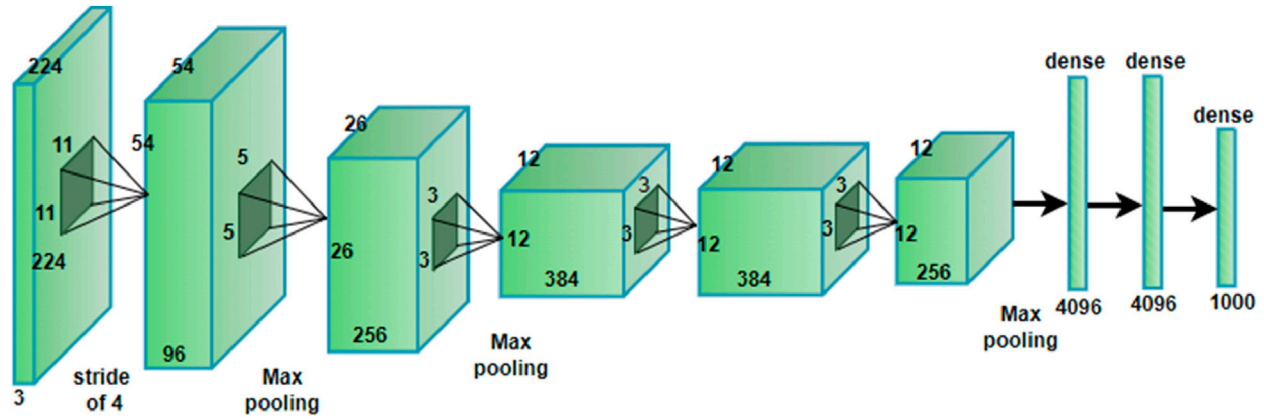

FIGURE 13
The layered architecture of the Alex Net.

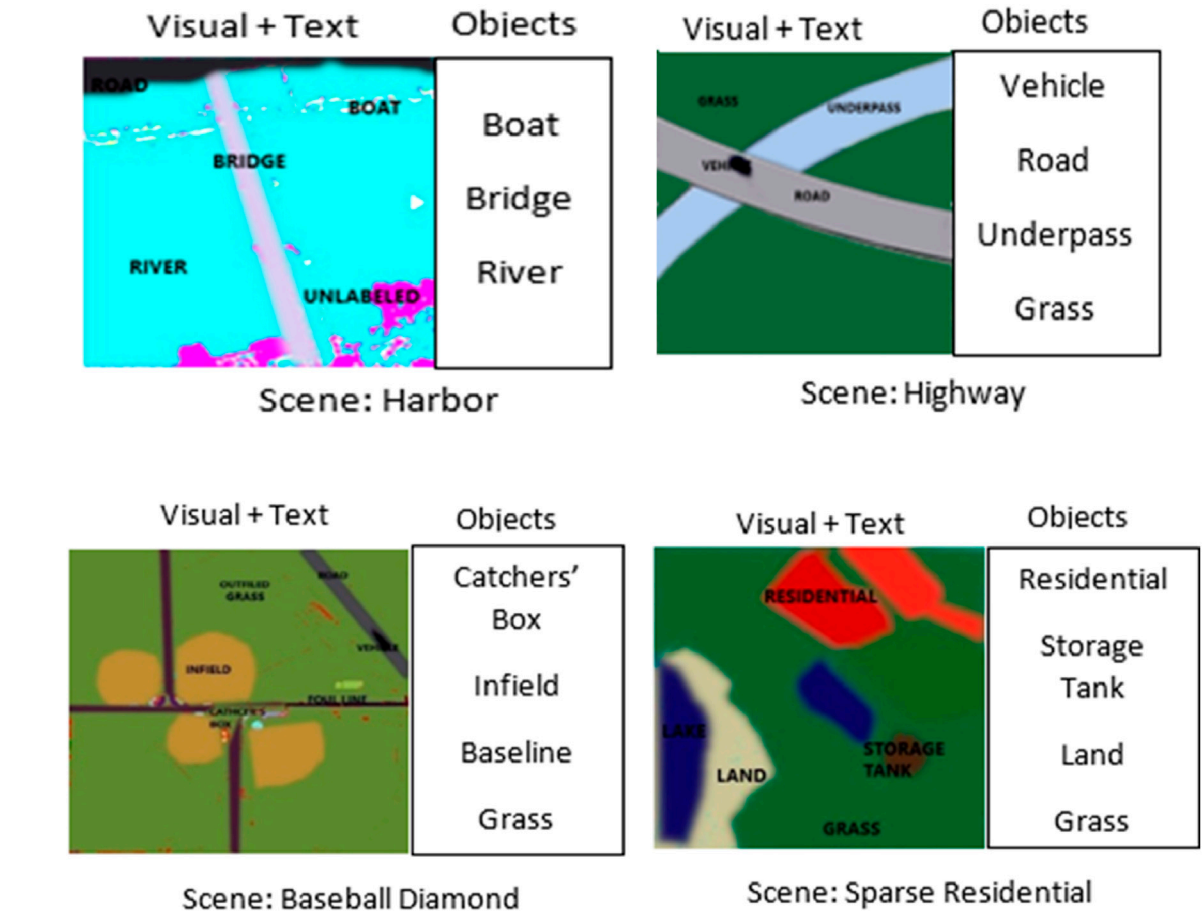

FIGURE 14
Scene recogntion using contextual relations between objects.

3.3.2 DBSCAN clustering

Density-Based Spatial Clustering or DBSCAN, is a popular clustering technique in the data analysis and machine learning domains. As opposed to conventional clustering techniques, which demand that the number of clusters be pre-specified, DBSCAN adopts a more data-driven methodology as in Equation 7. It is especially useful for discovering irregularly shaped and varying-sized clusters in complicated dataset since it clusters data

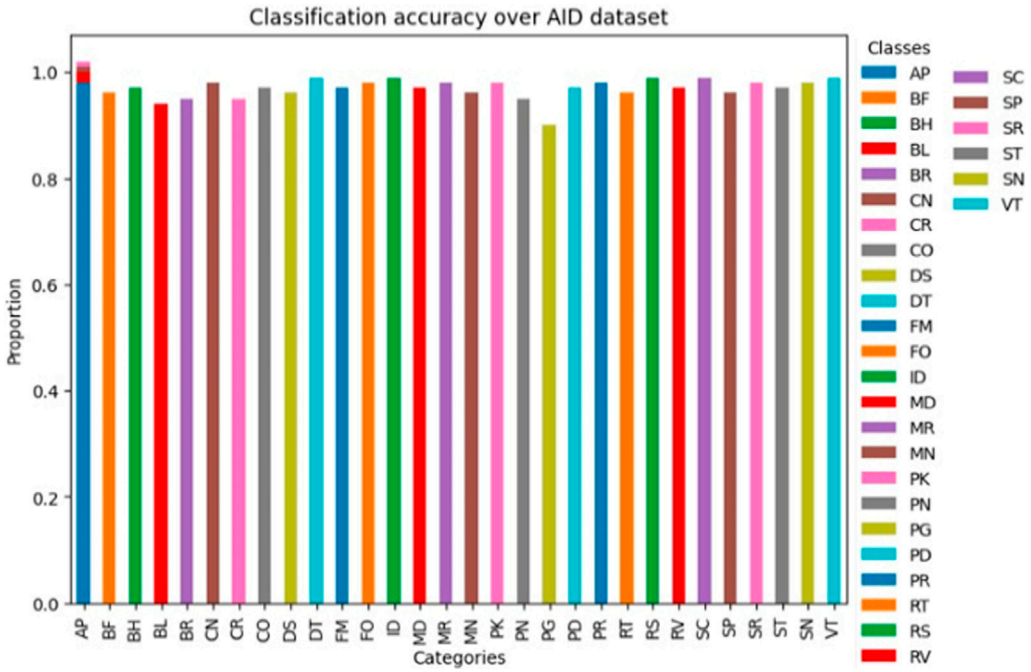

FIGURE 15
Scene classification accuracy over AID dataset.

points according to their density and closeness (Nawaz and Yan, 2020). The algorithm identifies core points as those with the fewest neighboring data points within a given distance Equation 8. After that, it adds neighboring data points that satisfy the density requirements, enlarging these core points into clusters (Jalal et al., 2021). Noise is defined as data points not fitting into any cluster or core point classification.

$$N_\varepsilon(x_i) = \left\{ x_j \in X \middle| dist\left(x_i, x_j\right) \le \varepsilon \right\}, X = \{ x_1, x_2, \ldots\ldots x_n \}, \quad (7)$$

$$C = \{ x_i \in X \| N_\varepsilon(x_i) \ge MinPts \} \quad (8)$$

where $x_i \in X \| N_\varepsilon(x_j)$ *and* $x_j \in C$; x_i is the epsilon neighborhood of x_j and x_j is the core point. Figure 4 representing the outcomes in which density based clusters are form.

3.3.3 Markov Random Field (MRF)

Consider G = v, e be the Markov Random Field (MRF) model's probabilistic network (Zheng et al., 2019) The vertex collection is represented by v = $\{v_s \mid s \in S\}$, while the edge set is written as E = $\{e_{s,t} \mid s, t \in S\}$. In the probabilistic graph, a single site is denoted by "s", while the whole collection of these sites is repesented by "S". $e_{s,t} = 1$ if vs and vt are next to one other in space; $e_{s,t} = 0$ otherwise. In the traditional pixel-based MRF model, G is a probabilistic graph where each "s" denotes a pixel. In the MRF model, G is employed if "s" denotes an over-segmented region. Figure 5 shows the MRF model (Li et al., 2018), where I_s stands for the observed data of vs in the image I = $\{I_s \mid s \in S\}$ S stands for the label field $X = [Xs|s \in S]$ in which the label class of each vs is represented by Xs, a random variable with values from $\Lambda = \{1, 2, \ldots, k\}$. For instantiation of $A, a = \{as|s \in S\}$, the posterior probability $P\{A = a|I\}$ may be found using Equation 9 on observed image I.

$$P(A|I) = \frac{P\left(I \middle| A = a\right).P(A)}{P(I)} \quad (9)$$

where $P(A|I)$ is the posterior probability, $(A|I)$ is the likelihood of observing, P(A) is prior Probability and P(I) is the probability of observing I.

$$P(I|A) = \prod_{i=1}^{n} f(I_i; A) \quad (10)$$

where I_i represents individual data points in the observed image (Lu et al., 2016). The spatial neighborhood interactions between labels of several places can be captured by the joint distribution as shown in Equations 11, 12 below.

$$P\left[Ac \middle| At, t \in \frac{v}{vs} \right] = P[Ac|At, t \in Ns] \quad (11)$$

In this instance, if est = 1, then vt is in Ns, which means that Ns contains vs's surrounding vertices. The Hammersley-Clifford theorem (Chen et al., 2017) uses potential functions to construct the joint probability distribution in MRFs.

$$\mathrm{P}[\mathrm{Ac}|\mathrm{At}, \mathrm{t} \in \mathrm{Ns}] = \frac{1}{Z} \prod_{c \in C} \varphi_c(A_c) \quad (12)$$

Z serves as the partition function to guarantee that the probabilities add up to 1, A_c stands for the variables in each clique, and C is the graph's collection of maximal cliques. The possible function that relates to cliques are φ_c, which represents the interaction between clique variables represented by c. According to the posterior prob. Of Equation 10. The segmentation of the image provided may be accomplished by finding the best realization by applying the MAP criterion Equation 13, i.e.,

10.3389/fbioe.2025.1430222

TABLE 2 Results for scene recognition among three classifiers over AID dataset.

Categories	ANN			XGBoost			AlexNet		
	Pn	Rc	F1 Scr	Pn	Rc	F1 Scr	Pn	Rc	F1 Scr
AP	0.811	0.855	0.832	0.768	0.732	0.751	0.895	0.977	0.937
BB	0.871	0.917	0.893	0.883	0.857	0.869	0.960	0.955	0.957
BH	0.915	0.955	0.934	0.995	0.951	0.972	0.924	0.972	0.947
BL	0.903	0.845	0.873	0.986	0.937	0.960	0.977	0.911	0.945
BR	0.944	0.933	0.938	0.967	0.903	0.933	0.899	0.935	0.916
CN	0.887	0.841	0.865	0.844	0.875	0.850	0.844	0.889	0.865
CR	0.935	0.798	0.862	0.755	0.839	0.795	0.915	0.887	0.903
CO	0.868	0.899	0.895	0.872	0.921	0.895	0.872	0.954	0.911
DS	0.933	0.884	0.907	0.886	0.938	0.909	0.928	0.971	0.950
DT	0.887	0.815	0.850	0.985	0.954	0.969	0.971	0.892	0.969
FM	0.809	0.856	0.837	0.901	0.856	0.877	0.915	0.977	0.937
FO	0.845	0.881	0.862	0.883	0.965	0.922	0.811	0.892	0.922
ID	0.929	0.862	0.895	0.995	0.951	0.972	0.913	0.928	0.972
MD	0.899	0.918	0.902	0.798	0.937	0.936	0.886	0.966	0.925
MR	0.975	0.955	0.965	0.967	0.903	0.933	0.897	0.937	0.916
MN	0.798	0.912	0.859	0.844	0.889	0.879	0.912	0.901	0.906
PK	0.845	0.947	0.893	0.889	0.839	0.863	0.977	0.887	0.929
PN	0.811	0.886	0.846	0.872	0.921	0.895	0.900	0.946	0.954
PG	0.869	0.818	0.842	0.886	0.938	0.909	0.855	0.891	0.872
PD	0.905	0.875	0.842	0.985	0.954	0.969	0.925	0.917	0.921
PR	0.956	0.836	0.891	0.901	0.859	0.877	0.937	0.977	0.956
RT	0.854	0.825	0.840	0.883	0.965	0.922	0.871	0.871	0.871
RS	0.957	0.851	0.900	0.879	0.851	0.864	0.995	0.951	0.972
RV	0.991	0.888	0.936	0.986	0.937	0.960	0.956	0.879	0.915
SC	0.899	0.835	0.865	0.967	0.809	0.880	0.891	0.903	0.896
SP	0.879	0.925	0.899	0.844	0.855	0.850	0.819	0.916	0.864
SR	0.933	0.918	0.925	0.899	0.839	0.867	0.977	0.921	0.948
ST	0.789	0.857	0.821	0.872	0.875	0.873	0.887	0.911	0.898
SN	0.887	0.877	0.881	0.886	0.913	0.810	0.793	0.935	0.858
VT	0.895	0.798	0.843	0.845	0.866	0.800	0.985	0.905	0.943
Mean	**0.859**	**0.875**	**0.880**	**0.897**	**0.885**	**0.875**	**0.903**	**0.921**	**0.936**

[a]AP, airplane; BB, baseball diamond; BH, beach; BL, bare land; BR, bridge; CN, center; CR, church; CO, commercial; DS, dense residential; DT, desert; FM, farmland; FO, forest; ID, industrial; MD, meadow; MR, medium residential; MN, mountain; PK, park; PN, Parking;PG, playground; PD, pond; PR, port; RT, railway station; RS, resort; RV, river; SC, school; SP, sparse residential; SR, square; ST, stadium; SN, storage tank; VD, viaduct; Pn, Precision, Rc, Recall. Bold values indicates proposed results (highlighed).

$$\tilde{a} = argmaxP(A|I) = argmax[P(I|A).P(A)] \tag{13}$$

Pair-site cliques are typically utilized to compute image segments φ_c (A_c) in the P $[Ac|At,t\epsilon Ns]$ where φ_c (A_c) = $\sum_{t\in N_s} V(a_c, a_t)$ see Equations 14-16.

$$V(a_c, a_t) = \begin{cases} -\vartheta, a_c = a_t \\ \vartheta, a_c \neq a_t \end{cases} \tag{14}$$

where ϑ is the potential parameter Thus enabling the representation of P (A= a):

TABLE 3 Results for scene recognition among three classifiers over UCM dataset.

Categories	ANN			XGBoost			AlexNet		
	Pn	Rc	F1 Scr	Pn	Rc	F1 Scr	Pn	Rc	F1 Scr
AG	0.755	0.788	0.755	0.799	0.899	0.817	0.901	0.977	0.937
AP	0.711	0.744	0.875	0.815	0.875	0.844	0.883	0.965	0.922
BB	0.783	0.711	0.746	0.841	0.819	0.747	0.995	0.951	0.972
BH	0.792	0.658	0.737	0.844	0.889	0.844	0.986	0.937	0.960
BD	0.701	0.725	0.758	0.889	0.839	0.889	0.967	0.903	0.933
CH	0.745	0.715	0.875	0.872	0.921	0.872	0.977	0.839	0.903
DN	0.799	0.791	0.795	0.886	0.938	0.886	0.872	0.921	0.895
FR	0.783	0.711	0.746	0.985	0.954	0.985	0.886	0.938	0.911
FW	0.771	0.792	0.781	0.901	0.859	0.901	0.985	0.954	0.969
GC	0.730	0.717	0.961	0.883	0.965	0.883	0.925	0.917	0.969
HR	0.755	0.788	0.755	0.879	0.851	0.879	0.936	0.977	0.937
IN	0.711	0.744	0.875	0.986	0.937	0.986	0.871	0.871	0.922
MR	0.783	0.711	0.746	0.967	0.809	0.967	0.995	0.951	0.972
MH	0.792	0.658	0.737	0.845	0.856	0.850	0.956	0.879	0.96
OP	0.701	0.725	0.758	0.879	0.851	0.864	0.891	0.903	0.933
PN	0.874	0.845	0.859	0.986	0.937	0.960	0.819	0.916	0.859
RV	0.869	0.829	0.903	0.967	0.809	0.880	0.977	0.921	0.903
RW	0.872	0.851	0.895	0.844	0.855	0.850	0.887	0.911	0.895
SP	0.886	0.918	0.911	0.899	0.839	0.867	0.793	0.935	0.911
SN	0.965	0.934	0.969	0.872	0.875	0.873	0.799	0.916	0.895
TC	0.901	0.859	0.937	0.886	0.913	0.899	0.855	0.891	0.911
Mean	**0.893**	**0.881**	**0.922**	**0.922**	**0.911**	**0.915**	**0.923**	**0.915**	**0.946**

[a]AG, agriculture; AP, airplane; BB, baseball diamond; BH, beach; BD, building; CH, chaparral; DN, dense residential; FR, forest; FW, Freeway GC, golf course; HR, harbor; IN, intersection; MR, medium residential; MH, mobile home park; OP, overpass; PN, parking; RV, river; RW, runway; Sp, Sparse Residential; SN, storage tank; TC, tennis court; ANN, Artificial Neural Network; XGBoost, eXtreme Gradient Bossting. Bold values indicates proposed results (highlighed).

$$P(A = a) = \Pi_{c\epsilon C}\, P[A_c = a_c, \mid t, t \in Ns\,] = \Pi_{c\epsilon C}\, \varphi_c\,(A_c) \qquad (15)$$

Therefore, by optimizing each $\tilde{a}_s$, the optimal realization $\tilde{a} = \{\tilde{a}_s\}$ may be attained progressively. The resultant segmentation obtained by using the MRF is shown in Figure 6.

$$\tilde{a}_s = argmax P\left[A_c = a_c \middle| I_c, A_t, \mathrm{t} \in \mathrm{Ns} = argmax\right] \times \left[P(I_c|A_A = a_c).P\left(A_c = a_c \,\middle|\, A, \mathrm{t} \in \mathrm{Ns}\right)\right] \qquad (16)$$

Based on the estimation time, three segmentation methods are contrasted. MRF-based segmentation is selected for labeling as it takes less time than DBSCAN and FCM. Labeling categorizes pixels by taking contextual information and nearby relationships into account. Using potentials or energies attached to pixel labels, MRF models these relationships (Nguyen et al., 2020).

3.4 Feature extraction

The extraction (Naseer and Jalal, 2023) of pertinent information or characteristics from aerial images is an essential phase in applications associated to remote sensing and image analysis. For object categorization in remote sensing images, a broad range of conventional features—including statistical techniques like texture, spatial, and spatial spectral features—are assessed. A thorough examination of the techniques for feature determination, combination, and selection is given in the following sections.

3.4.1 Spatial features

The statistical measures known as Haralick texture characteristics are used to characterize the spatial arrangement or texture of pixel values in an image. The contrast, energy, entropy and correlation are calculated with the help of the gray-level co-

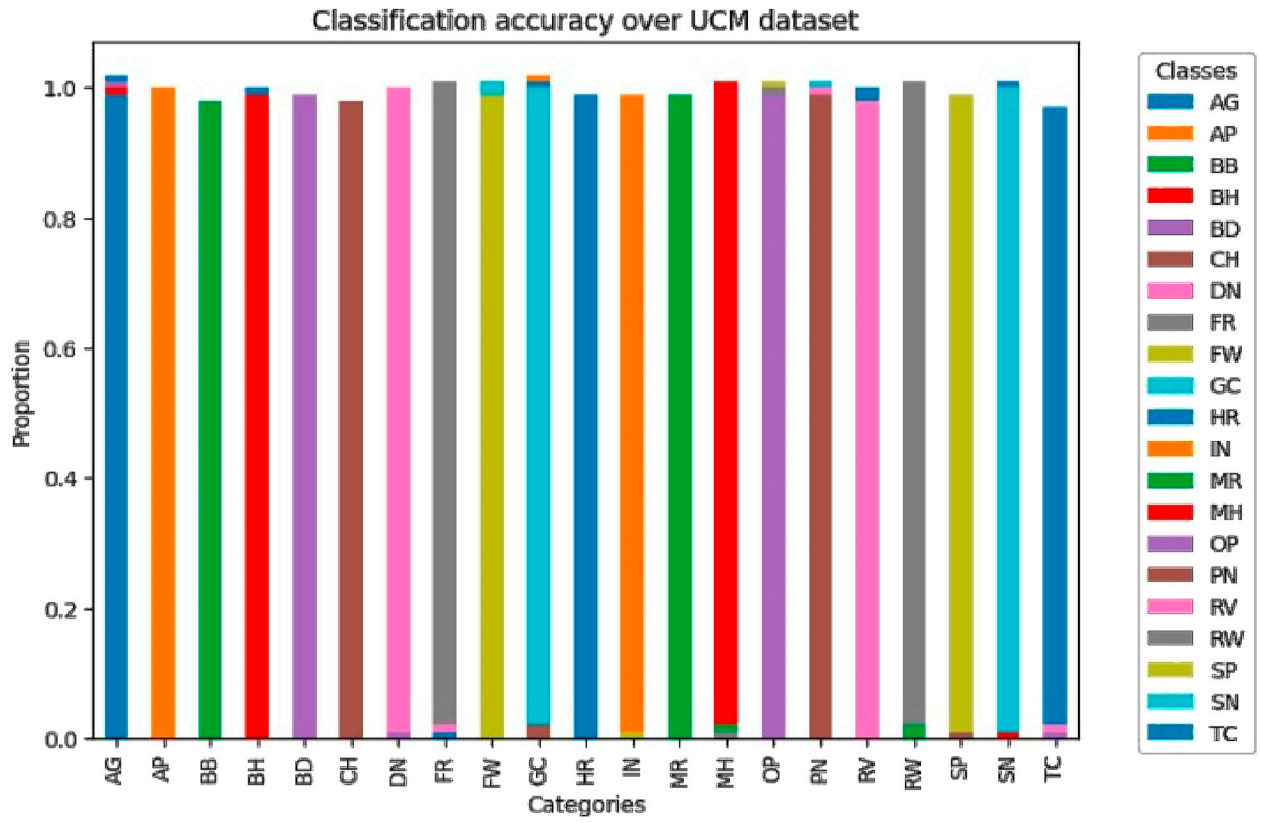

FIGURE 16
Scene classification accuracy over UC merced dataset.

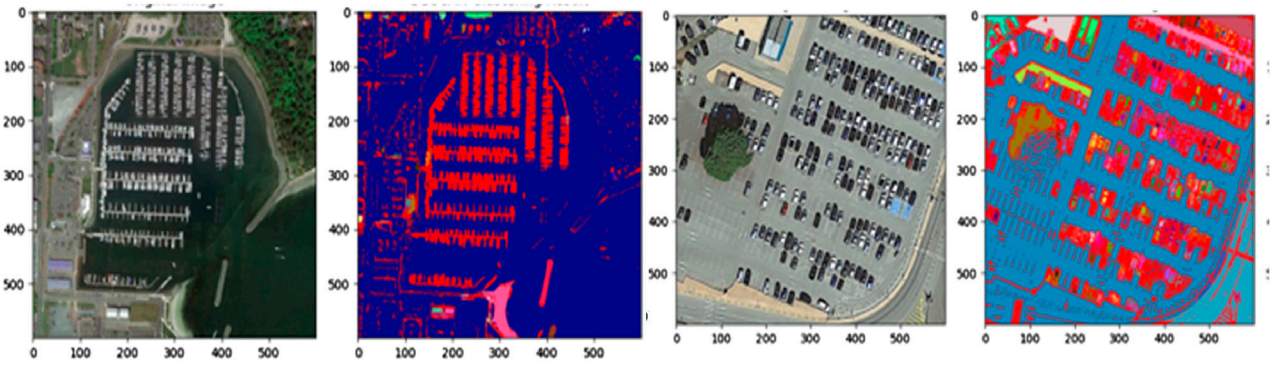

FIGURE 17
Failure cases where model does not perform well.

occurrence matrix (GLCM). Using the GLCM (Sharma et al., 2016), we were able to extract textural properties and deduce four Haralick features shown in Figure 7. Based on the texture of the landscape, these characteristics—energy, correlation, contrast, and homogeneity—can be used to categorize different types of land cover given in Equations 17-20 respectively, such as urban and wooded areas. We utilized the texture data from the GLCM, which Haralick said included texture properties (Hema and Kannan, 2020) Specific formulae can be used to quantitatively compute the important texture properties.

$$Enr(I) = -\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} S(u,v)^2 \tag{17}$$

$$Corr(I) = \frac{\sum_{u=0}^{N-1}\sum_{v=0}^{N-1}(u,v)S(u,v) - \mu_i\mu_j}{\delta_i\delta_j} \tag{18}$$

TABLE 4 Comparison of the SOTA methods with the proposed OCS model.

Authors	Mean accuracies %	
	AID	UCM
Kim and Chi (2021)	86.91	86.79
Cheng et al. (2017)	—	94.17
Yu and Liu (2021)	—	84.00
Xie et al. (2021)	96.01	—
Wang et al. (2018)	88.75	96.81
Kollapudi et al. (2022)	—	90.29
Thirumaladevi et al. (2023)	—	95.00
Proposed	**97.90**	**98.90**

AID, Aerial Image Dataset; UCM, UC Merced Dataset. Bold values indicates proposed results (highlighed).

$$Cont\,(I) = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1}(u-v)^2 P(u,v) \tag{19}$$

$$HMG(I) = -\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} P(u,v)P(u,v) \tag{20}$$

3.4.2 spectral features

For Aerial imaging needs spectral features because they give important information about the makeup and characteristics of the Earth's surface. Many spectral bands are frequently present in aerial shots, and provides distinct details about the electromagnetic energy reflected or emitted (Liang et al., 2018). These spectral properties are critical to many remote sensing and image processing applications. Spectral characteristics are mainly concerned with analyzing color information included in the image, which is usually obtained from many spectral bands. Metrics like the mean (Equation 21), standard deviation (Equation 22), and color histograms are examples of common spectral properties.

$$\bar{X} = \frac{1}{Z}\sum_{i=1}^{Z} x_i \tag{21}$$

$$\delta = \sqrt{\frac{1}{Z}\sum_{i=1}^{Z}(x_i - \bar{X})^2} \tag{22}$$

where $\bar{X}$ is the mean, δ is the entire set of values in the spectral feature, Z is the entire set of values, and x_i is the individual value. Mean values are presented in bands in Figure 8, which indicates that medians for Bands 1 and 2 are relatively close, but somewhat lower in Band 3. It also shows that there is likely little overlap between the notches, or the confidence intervals and hence there is good evidence that the medians of the bands are significantly different. Moreover, it can be observed that Band 3 has a higher degree of variability than Bands 1 and 2, which can also be seen from the whiskers. These observations bring out different features of each band which is important when trying to solve the problem of identifying the features or objects in the dataset.

3.4.3 Spatial spectral features

By separately computing different characteristics, such as textural features, spatial attributes, and spectral features, the approach yields Spectral Spatial Characteristics (SSFs), which are unique feature vectors. SSFs (Zhang et al., 2018) combine spatial and spectral (colour) data from representations of remote sensing. These elements include edge information, texture, spatial autocorrelation, and statistics derived from spectral bands (e.g., mean and standard deviation) that represent the spatial relationships among objects in the image together with their spectral qualities. SSFs are essential for distinguishing across groupings of land cover that have varied spatial layouts but comparable spectral features.

3.5 Feature fusion

In this experiment, we concatenate all the features to the feature vector. To make our dataset as useful as feasible, the objective was to find and keep only those characteristics that demonstrated substantial variance across all features. The Texture (GLCM), Spectral and SSF and are computed separately $Feature_{GLCM}$, $Feature_{SF}$ and $Feature_{SSF}$ respectively. With reference to (Zhao and Du, 2016; Raja et al., 2020) a complete fused feature vector is created by the fusing of many feature vectors. Normalization prior to fusion is essential for balanced representation of features. It keeps a single element from taking center stage in the fused feature vector. Subsequent data processing is enhanced by this integrity preservation. The fused feature vector is composed of elements from the Haralick, Spatial, and SSF features combined together as shown below Equation 23.

$$Fused_F = [\,Feature_{GLCM}\,Featur_{SF}\,Feature_{SSF}] \tag{23}$$

3.6 Dimensionality reduction using PCA

Dividing complex images into two layers produces a high-dimensional feature vector that can be used to manipulate the

TABLE 5 Ablation results for key features in the proposed model.

Experiment	Component removed	Accuracy (AID)	Accuracy (UCM)	F1 score (AID)	F1 score (UCM)
Full model	No	97.90%	98.90%	93.60%	94.16%
Without MRF	Replaced with FCM	92.50%	90.30%	92.10%	89.80%
Without spectral-spatial features	Removed	89.80%	87.60%	89.20%	86.90%
Without Haralick features	Removed	90.30%	88.20%	89.70%	87.50%
Without Bilateral Filter	Removed	85.60%	83.20%	85.10%	82.70%

image. Working with high dimensions, meanwhile, can provide less-than-ideal outcomes. This problem is addressed by PCA feature selection for dimensionality reduction, which projects the data into a lower-dimensional space. This method, called Feature_RF, solves problems with high-dimensional data while optimizing processing time and computational resources.

In order to do this, the reduced dimensional feature vector $Feature_{RF}$ is produced by using PCA feature selection as in Equation 24.

$$Feature_{RF} = PCA[Feature_F] \tag{24}$$

The SSF feature extraction of the combined features is shown in Figure 9, after which PCA is used to reduce dimensionality. The outcomes that followed are shown be.

3.7 Jaccard similarity

The Jaccard Index, is a widely used metric to assess the degree of agreement between predicted masks and ground truth masks in pixel-level precision tasks like image segmentation. Figure 10 provides a visual representation of the segmentation algorithm achieves exceptional accuracy for large, distinct objects like "Airplane" (IoU: 0. While in larger and easier to identify objects such as "Bike" (IoU: 0.99) and "Runway" (IoU: 0.98), there is high accuracy, small and ambiguous object like "Boat" the same partially trained model yields an IoU of 0.80.

The graphs (Figure 11; Figure 12) evaluate segmentation quality across various object classes, highlighting both high-performing and challenging cases. The two graphs demonstrate the jaccard similarity (IoU) scores of objects classes in both AID and UCM datasets for an analysis of the segmentation performance of the different categories of objects. The segmentation algorithm works with reasonable accuracy based on the IoU scores following the delineated ranges: ≤0.5; 0.5–0.7; 0.7–0.8; and >0.8. In the case of both datasets, most of the IoU values are above 0.8. Out of all the classes in the AID dataset, we see near perfect IoU scores for certain classes like AP (Airport) and BB (Baseball Field) which are easily distinguishable under spectral-spatial characteristics. Still, the low values identifying such classes as MR (Meadow) indicate difficulties distinguishing between these objects because of their similarities with the surrounding environment. Likewise, in case of UCM dataset, classes like CH and AG are well detected where IoU value is close to one and FW and SP have relatively low IoU value. Altogether, these results demonstrate the effectiveness of the presented algorithm for most considered object classes with indication on the potential improvement of the algorithm's performance in the most challenging circumstances.

3.8 Scene classification using AlexNet

The famous convolutional neural network (CNN) architecture known as AlexNet (shown in Figure 13) was created in 2012 by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. It demonstrated exceptional performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and was a major contributor to the deep learning for image classification field's broad adoption.

This type of model takes an image input with dimensions of 224 × 224 × three and passing through several convolutions, pooling and fully connected layers to identify the input image and place it into one of the 1,000 categories. The First convolutional layer uses 96 filters of size 11 × 11 with four stride which down samples and aims to capture large scale features. Following layers' use progressively smaller filters (5 × 5 and 3 × 3) with even more feature maps (up to 384) as in Equation 25 to identify the medium and even the fine scale features including the texture and edges. All these layers are said to follow some of the convolutional layers while they minimize the spatial dimensions while preserving the most important features in the images. Finally, following the final convolutional and pooling layers, there is a vectorization of the features maps, two dense layers containing 4,096 neurons each, to amalgamate the spatial features into a global space-quality representation. Last layer is the output layer that consists of 1,000 neurons, and performs softmax activation to drive out probabilities of the classes. Such a structure helps the model to learn higher levels of abstraction and for such tasks as object recognition the model is almost unbeatable.

$$Y = ReLU\left(convoluiton\left(X, W\right) + b\right) \tag{25}$$

The feature map (Y) is produced by convolution between (X, W), which is followed by the ReLU activation function. Spatial dimensions using Equations 26, 27 are then reduced by a max pooling layer using a 3 × 3 filter size and stride 2.

$$Y = \max\_pool\left(X, p\right) \tag{26}$$

$$Y\left(i, j, k\right) = X\left(i, j, k\right)/\left(k + \alpha^{*} \sum\left(X\left(i, j, k^{`}\right)\hat{}2\right)\hat{}\beta \tag{27}$$

where i, j, k representing spatial and depth dimensions, $k^{`}$ represents the neighboring depth indices, α and β are hyper parameters. There are three fully connected levels in AlexNet. The first two layers each include 4,096 neurons, while the output layer contains 1,000 neurons, matching the classes in the UC Merced dataset. Softmax activation is used in the output layer to handle class probabilities. To keep the first two layers from overfitting, 0.5 probability dropout is applied to them during training. The usual training approach is stochastic gradient descent (SGD) with momentum (Thirumaladevi et al., 2023), and to increase the diversity of the training set, data augmentation techniques like random cropping and horizontal flipping are applied. Scene recognition has been done by the contextual relationship between multiple objects from remote sensing images using AlexNet shown in Figure 14.

In the presented framework, the reason for selecting AlexNet is based on its architectural effectiveness and its applicability to the requirements of Remote Sensing Image Analysis (RSIA) tasks. AlexNet's hierarchical convolutional structure outperforms other deep architectures for detecting the small features at the initial layers such as edges and textures, and small semantic features at the later layers that are more relevant to scene recognition in complex data sets. This feature is considered a strength since it requires less computation and works efficiently on large datasets including the datasets AID and UCM it will not compromise performance with computational complexity.

Even though current progresses in architectures like ResNet or DenseNet provide deeper feature extraction in model architectures, they also bring more troubles like higher model complexity and stronger demand on computation resources which can an adversary against efficient implementations in certain situations. Moreover, the integration of AlexNet to the proposed Markov Random Field (MRF)-based segmentation enhances its performance to the maximum level, since it employs accurate and consistent segmentation outcomes for enhancing the feature learning process. The synergy that is tailored accordingly helped the system to obtain even more high classification rates 98.90% for UCM and 97.90% for AID databases, which is higher compared to many other state of the art methods. For that reason, AlexNet is poised to offer the best combination of efficiency, versatility, and reliability for this framework, and optimized for remote sensing use.

4 Experimental setup and dataset

Evaluation of the proposed system is performed on two benchmark datasets: Aerial Image dataset AID and UC Merced (UCM) dataset. The experiment is performed on an intel core i7 with 16 GB of RAM, a 3.2 GHz processor, and 512 GB of SSD.

4.1 Datasets description

4.1.1 The aerial image dataset (AID)

The most current large collection of aerial images is called the Aerial Images Dataset (Xia et al., 2017). This dataset, which consists of 10,000 images overall across thirty classes of different scenarios. The collection includes a variety of aerial scene types, such as beaches, bridges, business districts, barren terrain, baseball fields, and airports.

4.1.2 The UC merced dataset

A publicly accessible benchmark for study, the UCM dataset (Kim and Chi, 2021) consists of 100 images per class, all 256 × 256 pixels in size. These diverse images, which come from the USGS National Map Urban Area collection, include views of residences, beaches, farms, airplanes, and more.

4.2 Experimental evaluation

4.2.1 Precision, recall, and F1-score

We provide recognition accuracies utilizing AlexNet and the OSCM architecture for the UCM and AID datasets. Our method uses ANN trained with SSF, Haralick, and spectral features, then XGBoost. Tables 2, 3 present a comparative evaluation of our OSCM framework and AlexNet for accurate scene recognition on difficult datasets utilizing Precision, Recall, and F1 Score measurements on the AID and UCM datasets.

4.2.2 Second experiment: confusion matrix

These two figures show how the classification rate varies on categories in AID and UCM sets. Every dataset shows a good level of performance of most classes with accuracy proportions being very close to 1.0% suggesting that all classes are well classified. Small deviations in some classes indicate where optimization is necessary more than ever, especially in classes that present a weak spectral and spatial contrast, as are observed within the Figures 15, 16 and failure cases shown in Figure 17.

We investigated comprehensively by contrasting our proposed approach with accepted state-of-the-art techniques. This evaluation was especially concerned with determining the average accuracy in object classification and segmentation. The results, presented in Table 4, provide a thorough comparison with the state-of-the-art methods now in use. These results show a significant improvement in performance, which we attribute to our novel OSC system.

5 Ablation study

The proposed model, while achieving high accuracy and robustness, has certain limitations. First, its evaluation was limited to the AID and UC Merced datasets, which, although diverse, may not fully represent the complexity of real-world remote sensing scenarios. The model has not been tested on datasets containing multi-temporal or multi-sensor imagery, which could reveal its adaptability to varying data sources. Additionally, the lack of data augmentation techniques might have restricted the model's ability to generalize further to unseen variations in the datasets. This will be done by a step-wise elimination or alteration of components, namely, the MRF segmentation, texture features based on Haralick, spectral-spatial features or particular layers of Alex Net and consequently measure the effect on performance. Such experiments will bring more attention to each module and will delete the unnecessary parameters improving the architecture. Ablation analysis shown in Table 5 will also allow for deeper understanding of how such hyper parameters affect model behavior to ensure that the level of accuracy and scalability is well understood.

6 Discussion and future work

The methodology proposed in this paper outperforms others through the use of MRF based segmentation, contextual feature extraction, and the recognition of scenes using Alex Net. Minimum segmentation accuracies of 91.18%, 91.67% of UCM, and AID datasets, respectively, further enhances the credibility of the proposed MRF model compared to DBSCAN and FCM. The addition of spatial, spectral and Haralick texture features aids in object discrimination whilst the incorporation of AlexNet aids in scene discrimination with accuracies of 97.90% on AID dataset and 98.90% on the UCM dataset. The novel method outperforms Xie et al. (96.01%) and Thirumaladevi et al. (95%) and establishes that the system can be successfully utilised for remote sensing. The generalizability of our proposed system to new circumstances is limited since it does not show how the OSC system works in real-world applications outside of the benchmark datasets.

Future research endeavors aimed at optimizing the OSC system's efficacy will encompass the integration of advanced deep learning techniques, exploration of temporal and spatial transferability analysis, and evaluation of its robustness under

diverse conditions. In addition, the model will be thoroughly examined by testing on several benchmark datasets and comparing its effectiveness with state-of-the-art technology.

7 Conclusion

The suggested OSC method performs exceptionally well at segmenting and categorizing objects in complex aerial data. We use a two-step strategy, where we use MRF for accurate segmentation and deep learning model Alex Net for scene recognition, using benchmark datasets. When processing complicated aerial images, our model performs better than state-of-the-art techniques, exhibiting outstanding accuracy and reliability. Our OSC system offers a dependable and complex solution with state-of-the-art methodologies and benchmark datasets, expanding the field of remote sensing research and creating new opportunities for aerial image analysis. In some images of classes (Parking, Port), our model does not perform well in distinguishing small and overlapping objects, accurately segmenting boundaries, and handling areas with similar textures, such as water and shadows. These limitations highlight the challenges in achieving precise segmentation in complex and densely packed regions.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/datasets/abdulhasibuddin/uc-merced-land-use-dataset.

Author contributions

AN: Investigation, Writing – original draft. NA: Investigation, Writing – review and editing. HA: Data curation, Writing – review and editing. AbA: Project administration, Writing – review and editing. YA: Conceptualization, Writing – review and editing. AsA: Project administration, Writing – review and editing. AJ: Supervision, Writing – original draft. HL: Validation, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The APC was funded by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. The research team thanks the Deanship of Graduate Studies and Scientific Research at Najran University for supporting the research project through the Nama'a program, with the project code NU/GP/SERC/13/18-4. The authors acknowledge Princess Nourah bint Abdulrahman University Researchers supporting Project number (PNURSP2025R54), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Group Project under grant number (RGP.2/568/45).

Conflict of interest

Author HL was employed by Guodian Nanjing Automation Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ahmed, A., Jalal, A., and Kim, K. (2020). A novel statistical method for scene classification based on multi-object categorization and logistic regression. *Sensors* 20 (14), 3871. doi:10.3390/s20143871

Ahmed, M. W., and Jalal, A. (2024). *Dynamic adoptive Gaussian mixture model for multi-object detection over natural scenes*. Lahore, Pakistan: ICACS.

Alazeb, A., Azmat, U., Al Mudawi, N., Alshahrani, A., Alotaibi, S. S., Almujally, N. A., et al. (2023). Intelligent localization and deep human activity recognition through IoT devices. *Sensors* 23 (17), 7363. doi:10.3390/s23177363

Ansith, S., and Bini, A. A. (2022). Land use classification of high-resolution remote sensing images using an encoder-based modified GAN architecture. *Displays* 74, 102229. doi:10.1016/j.displa.2022.102229

Ball, J. E., and Bruce, L. M. (2007). Level set hyperspectral image classification using best band analysis. *IEEE Trans. Geosci. Remote Sens.* 45, 3022–3027. doi:10.1109/tgrs.2007.905629

Chen, J., Yang, C., Xu, G., and Ning, L. (2018). Image segmentation method using Fuzzy C mean clustering based on multi-objective optimization. *J. Phys.* 1004, 012035. doi:10.1088/1742-6596/1004/1/012035

Chen, X. H., Zheng, C., Yao, H. T., and Wang, B. X. (2017). Image segmentation using a unified Markov random field model. *IET Image Process* 11, 860–869. doi:10.1049/iet-ipr.2016.1070

Cheng, G., Han, J., and Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. *Proc. IEEE* 105 (10), 1865–1883. doi:10.1109/jproc.2017.2675998

Ghanbari, M., Xu, L., and Clausi, D. A. (2023). Local and global spatial information for land cover semi-supervised classification of complex polarimetric SAR data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 16. 3892–3904. doi:10.1109/JSTARS.2023.3264452

Grzeszick, R., Plinge, A., and Fink, G. A. (2017). "Bag-of-Features Methods for Acoustic Event Detection and Classification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25 (6), 1242–1252. doi:10.1109/taslp.2017.2690574

Guo, H., Liu, J., Xiao, Z., and Xiao, L. (2020). Deep CNN-based hyperspectral image classification using discriminative multiple spatial-spectral feature fusion. *Remote Sens. Lett.* 11 (9), 827–836. doi:10.1080/2150704x.2020.1779374

Halder, A., Pramanik, S., and Kar, A. (2011). Dynamic image segmentation using fuzzy c-means based genetic algorithm. *Int. J. Comput. Appl.* 28 (6), 15–20. doi:10.5120/3392-4714

Han, X., Zhong, Y., Cao, L., and Zhang, L. (2017). Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sens.* 9 (8), 848. doi:10.3390/rs9080848

He, Q. (2024). Prompting multi-modal image segmentation with semantic grouping. *Proc. AAAI Conf. Artif. Intell.* 38 (3), 2094–2102. doi:10.1609/aaai.v38i3.27981

He, Q., Xiao, Z., Huang, Z., Yuan, H., and Sun, L. (2024). "Orientation-Aware Multi-Modal Learning for Road Intersection Identification and Mapping," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 16185–16191. doi:10.1109/ICRA57147.2024.10610015

He, Q., Yan, Z., Diao, W., and Sun, X. (2024). DLC: dynamic loss correction for cross-domain remotely sensed segmentation. *IEEE Trans. Geoscience Remote Sens.* 62, 1–14. doi:10.1109/tgrs.2024.3402127

Hema, D., and Kannan, S. (2020). Hybridizing local and global features by sequential fusion technique for object detection and learning. *Int. J. Adv. Sci. Technol.* 29 (05), 2401–2407. Available online at: https://www.researchgate.net/publication/340978840

Hu, Q., Xu, W., Liu, X., Cai, Z., and Cai, J. (2023). Hyperspectral image classification based on bilateral filter with multispatial domain. *IEEE Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2021.3058182

Islam, T., Mukhopadhyay, S. C., and Suryadevara, N. K. (2016). Smart sensors and internet of things: a postgraduate paper. *IEEE Sensors J.* 17 (3), 577–584. doi:10.1109/jsen.2016.2630124

Jain, G., and Anto, S. (2022). Satellite image processing using fuzzy logic and modified K-means clustering algorithm for image segmentation. *Comput. Intell. Mach. Learn.* 3 (2), 57–61. doi:10.36647/ciml/03.02.a008

Jalal, A., Ahmed, A., Rafique, A. A., and Kim, K. (2021). Scene Semantic recognition based on modified Fuzzy c-mean and maximum entropy using object-to-object relations. *IEEE Access* 9, 27758–27772. doi:10.1109/access.2021.3058986

Jayanetti, J. M., Meedeniya, D. A., Dilini, M. D. N., Wickramapala, M. H., and Madushanka, J. H. (2017). "Enhanced land cover and land use information generation from satellite imagery and foursquare data," in Proceedings of the 6th International Conference on Software and Computer Applications, Bangkok, Thailand, February 26–28, 2017, 149–153. doi:10.1145/3056662.3056681

Khan, M. A., Sharif, M., Akram, T., Raza, M., Saba, T., and Rehman, A. (2020). Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition. *Appl. Soft Comput.* 87, 105986. doi:10.1016/j.asoc.2019.105986

Kim, J., and Chi, M. (2021). SAFFNet: self-attention-based feature fusion network for remote sensing few-shot scene classification. *Remote Sens.* 13 (13), 2532. doi:10.3390/rs13132532

Kollapudi, P., Alghamdi, S., Veeraiah, N., Alotaibi, Y., Thotakura, S., and Alsufyani, A. (2022). A new method for scene classification from the remote sensing images. *Comput. Mater. and Continua* 72 (1), 1339–1355. doi:10.32604/cmc.2022.025118

Längkvist, M. J., Kiselev, A., Alirezaie, M., and Loutfi, A. (2016). Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sens.* 8, 329. doi:10.3390/rs8040329

Li, S. Z. (1995). *Markov Random Field Modeling in Computer Vision. Computer Science Workbench.Book*, New York, NY: Springer, 88–90.

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., and Sun, J. (2018). Detnet: A backbone network for object detection, arXiv preprint arXiv:1804.06215. doi:10.48550/arXiv.1804.06215

Liang, J., Zhou, J., Tong, L., Bai, X., and Wang, B. (2018). Material-based salient object detection from hyperspectral images. *Pattern Recognit.* 76, 476–490. doi:10.1016/j.patcog.2017.11.024

Liu, G. Y., Zhang, Y., and Wang, A. M. (2015). Image fuzzy clustering based on the region-level Markov random field model. *IEEE Geoscience Remote Sens. Lett.* 12, 1770–1774. doi:10.1109/lgrs.2015.2425225

Liu, Y., Liu, Y., and Ding, L. (2017). Scene classification based on two-stage deep feature fusion. *IEEE Geoscience Remote Sens. Lett.* 15 (2), 183–186. doi:10.1109/lgrs.2017.2779469

Lu, Q. L., Huang, X., Li, J., and Zhang, L. P. (2016). A novel MRF-based multi-feature fusion for classification of remote sensing images. *Remote Sens.* 13, 515–519. doi:10.1109/lgrs.2016.2521418

Ma, H. C., and Yang, Y. (2009). Two Specific multiple-level-set models for high-resolution remote-sensing image classification. *IEEE Geosci. Remote Sens. Lett.* 6, 558–561. doi:10.1109/LGRS.2009.2021166

Mahakalanda, I., Demotte, P., Perera, I., Meedeniya, D., Wijesuriya, W., and Rodrigo, L. (2022). Deep learning-based prediction for stand age and land utilization of rubber plantation. *Appl. Mach. Learn. Agric.*, 131–156. doi:10.1016/B978-0-323-90550-3.00008-4

Manfreda, S., McCabe, M. F., Miller, P. E., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., et al. (2018). On the use of unmanned aerial systems for environmental monitoring. *Remote Sens.* 10 (4), 641. doi:10.3390/rs10040641

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016). Semantic segmentation of aerial images with an ensemble of CNSS. ISPRS Annals of the Photogrammetry. *Remote Sens. Spatial Inf. Sci.* (3), 473–480. doi:10.5194/isprsannals-iii-3-473-2016

Martin, S. (2011). "Sequential bayesian inference models for multiple object classification," in 14th International Conference on Information Fusion 2011, Chicago, IL, July 05–08, 2011 (IEEE), 1–6.

Meedeniya, D. A., Jayanetti, J. M., Dilini, M. D. N., Wickramapala, M. H., and Madushanka, J. H. (2020). *Land-use classification with integrated data. Machine vision inspection systems*, Hoboken, NJ: Wiley. 1–36.

Muhammad, U., Wang, W., Chattha, S. P., and Ali, S. (2018). "Pre-trained VGGNet architecture for remote-sensing image scene classification," in 2018 24th International Conference on Pattern Recognition (ICPR) 2018, Beijing, China, August 20–24, 2018 (IEEE), 1622–1627.

Naseer, A., Almujally, N. A., Alotaibi, S. S., Alazeb, A., and Park, J. (2024a). Efficient object segmentation and recognition using multi-layer perceptron networks. *Comput. Mater. and Continua* 78 (1), 1381–1398. doi:10.32604/cmc.2023.042963

Naseer, A., Alzahrani, H. A., Almujally, N. A., Al Nowaiser, K., Al Mudawi, N., Algarni, A., et al. (2024b). Efficient multi-object recognition using GMM segmentation feature fusion approach. *IEEE Access* 12, 37165–37178. doi:10.1109/access.2024.3372190

Naseer, A., and Jalal, A. (2023). "Pixels to precision: features fusion and random forests over labelled-based segmentation," in IEEE International Bhurban Conference on Applied Sciences and Technology, Bhurban, Pakistan, August 22–25, 2023 (IEEE), 1–6.

Naseer, A., and Jalal, A. (2024). *Integrating semantic segmentation and object detection for multi-object labeling in aerial images*. Lahore, Pakistan: ICACS.

Nawaz, M., and Yan, H. (2020). Saliency detection via multiple-morphological and superpixel based fast fuzzy C-mean clustering network. *Expert Syst. Appl.* 161, 113654. doi:10.1016/j.eswa.2020.113654

Nguyen, H. T., Lee, E. H., Bae, C. H., and Lee, S. (2020). Multiple object detection based on clustering and deep learning methods. *Sensors* 20 (16), 4424. doi:10.3390/s20164424

Raja, R., Kumar, S., and Mahmood, M. R. (2020). Color object detection based image retrieval using ROI segmentation with multi-feature method. *Wirel. Personal. Commun.* 112 (1), 169–192. doi:10.1007/s11277-019-07021-6

Saha, S., Bovolo, F., and Bruzzone, L. (2019). Unsupervised deep change vector analysis for multiple-change detection in VHR images. *IEEE TGRS* 57 (6), 3677–3693. doi:10.1109/TGRS.2018.2886643

Salleh, S. S., Aziz, N. A. A., Mohamad, D., and Omar, M. (2012). Combining mahalanobis and jaccard distance to overcome similarity measurement constriction on geometrical shapes. *Int. J. Comput. Sci. Issues (IJCSI)* 9 (4), 124.

Sangeetham, R., and Sanam, N. R. (2023). High-resolution image classification using a support vector machine. *AIP Conf. Proc.* 2754 (1). doi:10.2991/assehr.k.220301.164

Sharma, A., Malik, A., and Rohilla, R. (2016). A robust mean shift integrating color, GLCM-based texture features and frame differencing. *Int. J. Sci. Eng. Res.* 7 (2), 1386–1398. Available online at: https://scholar.google.com/scholar?oi=bibs&cluster=10207701677313736187&btnI=1&hl=en

Srivastava, P. K., Han, D., Rico-Ramirez, M. A., Bray, M., and Islam, T. (2012). Selection of classification techniques for land use/land cover change investigation. *ASR* 50 (9), 1250–1265. doi:10.1016/j.asr.2012.06.032

Sumbul, G., Cinbis, R. G., and Aksoy, S. (2019). Multisource region attention network for fine-grained object recognition in remote sensing imagery. *IEEE Trans. Geoscience Remote Sens.* 57 (7), 4929–4937. doi:10.1109/tgrs.2019.2894425

Tang, P., Wang, H., and Kwong, S. (2017). G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing* 225, 188–197. doi:10.1016/j.neucom.2016.11.023

Thirumaladevi, S., Swamy, K. V., and Sailaja, M. (2023). Remote sensing image scene classification by transfer learning to augment the accuracy. *Meas. Sensors* 25, 100645. doi:10.1016/j.measen.2022.100645

Tripathi, A. K., and Mukhopadhyay, S. (2012). "Single image fog removal using a bilateral filter," in 2012 IEEE International Conference on Signal Processing, Computing and Control, Solan, India, March 15–17, 2012 (IEEE), 1–6.

Wang, J., Zhao, T., Jiang, X., and Lan, K. (2022). A hierarchical heterogeneous graph for unsupervised SAR image change detection. *Geoscience Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2022.3224454

Wang, L., Huang, X., Zheng, C., and Zhang, Y. (2017). A Markov random field integrating spectral dissimilarity and class co-occurrence dependency for remote sensing image classification optimization. *ISPRS J. Photogram. Remote Sens.* 128, 223–239. doi:10.1016/j.isprsjprs.2017.03.020

Wang, M., Zhang, X., Niu, X., Wang, F., and Zhang, X. (2019). Scene classification of high-resolution remotely sensed image based on ResNet. *J. Geo Vis. Spatial Analysis* 3 (2), 16–19. doi:10.1007/s41651-019-0039-9

Wang, Q., Liu, S., Chanussot, J., and Li, X. (2018). Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geoscience Remote Sens.* 57 (2), 1155–1167. doi:10.1109/tgrs.2018.2864987

Wang, Y., Yang, L., Liu, X., and Yan, P. (2024). An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. *Sci Rep* 14, 9716. doi:10.1038/s41598-024-60375-1

Xia, G. S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., et al. (2017). AID: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geoscience Remote Sens.* 55 (7), 3965–3981. doi:10.1109/tgrs.2017.2685945

Xie, H., Chen, Y., and Ghamisi, P. (2021). Remote sensing image scene classification via label augmentation and intra-class constraint. *Remote Sens.* 13 (13), 2566. doi:10.3390/rs13132566

Yu, Y., and Liu, F. (2021). A two-stream deep fusion framework for high-resolution aerial scene classification. *Comput. Intell. Neurosci.* 2018, 1–13. doi:10.1155/2018/8639367

Zhang, C., Cheng, J., Li, L., Li, C., and Tian, Q. (2017). "Object Categorization Using Class-Specific Representations," in IEEE Transactions on Neural Networks and Learning Systems, 29 (9), 4528–4534. doi:10.1109/tnnls.2017.2757497

Zhang, L., Zhang, Y., Yan, H., Gao, Y., and Wei, W. (2018). Salient object detection in hyperspectral imagery using multi-scale spectral-spatial gradient. *Neurocomputing* 291, 215–225. doi:10.1016/j.neucom.2018.02.070

Zhao, W., and Du, S. (2016). Spectral–spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach. *IEEE Trans. Geoscience Remote Sens.* 54 (8), 4544–4554. doi:10.1109/tgrs.2016.2543748

Zhao, Y., Yuan, Y., and Wang, Q. (2019). Fast spectral clustering for unsupervised hyperspectral image classification. *Remote Sens.* 11, 399, doi:10.3390/rs11040399

Zheng, C., Pan, X., Chen, X., Yang, X., Xin, X., and Su, L. (2019). An object-based Markov random field model with anisotropic penalty for semantic segmentation of high spatial resolution remote sensing imagery. *IEEE Trans. Geoscience Remote Sens. Remote Sens.* 11 (23), 2878. doi:10.3390/rs11232878

Zhou, P. C., Xue, Y., and Xue, M. G. (2023). Adaptive side window joint bilateral filter. *Vis. Comput.* 39 (4), 1533–1555. doi:10.1007/s00371-022-02427-z