JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 184-215
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Machine and Deep Learning Approaches for Multi-Class Land Cover Classification Using High-Resolution Satellite Imagery

Ahmed Qusay Subhe[1], Muntadher Aidi Shareef[1], Sadra Karimzadeh[1]*

[1] *Northern Technical University, Technical Engineering College / Kirkuk, Department of Surveying Technical Engineering, Kirkuk 36001, IRAQ*

*Corresponding Author: sadra.karimzadeh@gmail.com
DOI: https://doi.org/10.30880/jscdm.2026.07.02.012

Article Info

Received: 7 February 2026
Accepted: 6 June 2026
Available online: 30 June 2026

Keywords

Land use land cover classification, very high-resolution satellite imagery, deep learning, object-based image analysis, graph convolutional network, SPOT-5, remote sensing

Abstract

Land use and land cover (LULC) classification using Very High Resolution (VHR) satellite data is significant for a wide scope of applications in urban planning, environmental monitoring, engineering, and sustainable development. However, there is a lack of research on the performance comparison of different classifiers in optimal computing environments. In this research, twelve classification algorithms from Pixel-based, object-based image analysis (OBIA), patch-based deep learning, and graph and theory-based approaches are the four methodological paradigms that are looked at. In this study, we make a complete comparison between 12 classification algorithms belonging to four methodological paradigms for LULC classification using SPOT-5 multispectral imagery of the Iraqi city of Mosul. The methods that are being taken into consideration are Support Vector Machine (SVM), Random Forest (RF), Support Vector Machine (SVM) and Random Forest (RF) classifiers in SNIC-based OBIA, Seven models for deep learning (including CNN, ResNet, ViT, DaViT, U-Net with CNN, U-Net with ConvNeXt, and EfficientNet), and a Graph Convolutional Network (GCN). Our results demonstrate that while patch-based deep-learning techniques achieve higher accuracy, there is a significant difference between the accuracy of the various techniques. When using DaViT together with CNN, the best overall accuracy was reached, which is 99.83%, and has a kappa value of 0.9977. EfficientNet was in second place with 99.66%. The best performance was given by U-Net at 99.49%, followed by ViT at 99.41%. The classical pixel-wise RF showed an accuracy of 98.63% and can compete with much lower computational costs. The OBIA method with RF resulted in an accuracy of 98.67%, whereas the accuracy of the GCN-based method was only 78.16%. The McNemar's test result was very significant when comparing the best and worst performing models ($p = 0.02$). The analysis by class showed that the Road class was the hardest to classify in various approaches, while the Water class had a high level of accuracy. On average, Agriculture yielded the highest percentage of pixel misclassifications across all paradigms with only 12% pixels being correctly classified. Our findings provide empirical advice for selecting a method for VHR satellite image classification, taking into account the

trade-offs among accuracy, computational efficiency, and practical implementation.

1. Introduction

Getting accurate information about how land is used and covered from satellite images is very important in remote sensing. This helps a lot in keeping track of the environment, planning cities, managing natural resources, and working towards sustainable development . Because cities are growing quickly and humans are affecting nature more than ever, there is a big need for maps that show LULC in a way that is accurate, up-to-date, and able to cover large areas . High-resolution imagery from sensors like SPOT-5, WorldView, and GaoFen has changed the way we think about remote sensing. These systems can see much smaller details in the landscape that couldn't be seen with less detailed data before. Even though these technological advances have brought many benefits, they have also introduced new challenges in how we analyze the data. This is because high-resolution images often have complex patterns and variations across different areas, making it harder to classify the information accurately. As a result, more sophisticated techniques are needed to handle this type of data effectively [1-5].

VHR image classification has gone through various different approaches over time. Using models like SVMs and RF, pixel-based classification is still a common approach . Even though these methods are effective, they often create salt-and-pepper noise because they don't take into account the spatial context . Object-Based Image Analysis (OBIA) was developed to address this by using near-segmentation to divide imagery into areas that are similar in their surroundings, so that shape and texture features can be better understood . With the arrival of deep learning, methods that use CNNs based on patches became popular proposed to learn hierarchical features from a larger image window. Patch-level road extraction is done with a simple CNN model . A binary detection head with dense layers helps the model recognize and use local spatial patterns via the U-Net path and global context information via the transformer path to accurately segment boundaries . Graph Neural Networks (GNNs) have given a way to show non-Euclidean spatial connections in different types of fields [6-8].

Even though these models have been studied a lot, there aren't many controlled studies comparing them directly, and when there are, they usually only compare two at a time . In this study, the authors take a look at twelve different approaches to information classification: pixel-wise, object-based, patch-based deep learning and graph-based. They investigate them in a specific case study on the city of Mosul in Iraq, using satellite images from SPOT-5. Pixel-wise classification, where each pixels in the hyperspectral images is assigned to a category based on its spectral information, remains a key approach [9-13]. Gave a detailed explanation about how SVM is used and its capability to handle images from the air and from satellites. discussed SVM methodologies, and also described the latest techniques for this procedure, like active and semi-supervised SVMs in VHR data. However, a meta-analysis indicated the contrary. Noted that the absence of spatial dependencies made SVM and RF disadvantaged by nature. Further compared the merits of RF to SVM to demonstrate that the performance of an optimized RF classifier could be superior to classic SVM and default RF feature on Sentinel-2 images [14-17].

Compared to the Maximum Likelihood classification, [18] indicated SVMs were also as effective as SPOT-5 imagery in classifying crops. In comparison to pixel-based methods, the object-based methods, and found that RF and SVM yielded more visually pleasing land cover maps than the Decision Trees. This is the work of [17], which used RF, SVM, and CNNs on SPOT-5 and Sentinel-2 images for change detection. While introduced the still existing popularity of SVM, [19] proved that RF was more accurate compared to SVM in the classification of tree species in pansharpened VHR images.

Deep Learning and Patch Inference-Based Approaches. Deep learning models have yielded the pioneering accuracies of classification in patch-based scenarios. presented a deep patch-based CNN network that was better than pixel-based neural networks. used this paradigm in an adopted fashion to choose the type of the roof, depending upon a pre-trained CNN on WorldView-2 images. Benchmark datasets have pushed the field in the right direction; proposed the EuroSAT dataset, and 98.57% accuracy is attained using the deep CNNs, on the move, [19] prepared a data set of Indian landscape, and identified that under various deep learning approaches, the recognition rate was more than 90%.

In the van is coming the new contact-architectural discoveries. introduced the ResASPP-Unet trained on the urban land cover, which was highly accurate on WorldView imagery. studied ten segmentation construction models, including deep convolutional NN and transformers. Vision Transformers (ViT): New transformer-based networks. Transformer-based networks were also recently introduced in Wav2Vec 2 as a possible alternative to ConvNets, whose performance improves with network size [20, 21]. A ViT architecture based on small datasets was created, and the model was suggested with multi-head local self-attention. Furthermore, merges the Data-efficient Image Transformer with U-Net, which compares state-of-the-art models, e.g., ConvNeXt and Swin Transformer, on EuroSAT. Recent studies from 2023 to 2026 have further pushed the boundaries of remote sensing classification.

For instance, Rad [22] proposed advanced vision transformer configurations specifically optimized for multispectral satellite imagery, demonstrating significant improvements in classification robustness. Similarly,

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

Lee et al. [23] and later Vinod et al. [24] introduced a novel multitask transformer architecture that integrates local self-attention with a UNet decoder for joint classification and segmentation. In parallel, hybrid deep learning models that combine CNNs and Transformers have been explored to capture both local fine-grained boundaries and global contextual relationships, such as the CNN-Transformer integration for bi-temporal SAR flood detection by Noori et al. [14]. These models highlight the shift towards hybrid spatial-spectral learning architectures in VHR satellite image analysis.

Object-Based Image Analysis (OBIA) OBIA solves the problem of per-pixel by categorizing segmented regions. had already demonstrated that object representations can be used with SVM. introduced the ensemble model SVM/RF (Support Vector Machine/Random Forest) in an OBIA setting on the basis of multiresolution segmentation. cumulative multi-scale segmentation and RF ensemble, outperforming the SVM and Decision Trees on complex urban settings [25-27].

The quality of segmentation is also a particularly critical parameter in OBIA; observed that the quality of segments is directly proportional to classification accuracy on WorldView-2 data. performed an analysis of the various segmentation parameters and their effects on the performance of the RF classifier. The recent developments include [13], which used Simple Non-iterative Clustering (SNIC) in combination with SVM and RF deployed on Google Earth Engine; and [28], which used the Superpixel Contour algorithm and mechanism.

GNNs may be considered as one of the promising ways of spatial relationship modeling. Formulated deep GCNs to classify VHR scenes by image representation using Region Adjacency Graphs. proposed an Object-Based Image Classification pipeline that used GCNs and could achieve great classification accuracy at a decent computation time. Applied superpixel-based GNNs to waterbody extraction and semantic segmentation problems, respectively, and made the contextual information represented in the form of a graph [29-33].

The hybrid architectures are also changing. proposed a spatio-temporal method that combined CNNs at the feature extraction stage and GraphSAGE to classify nodes. We obtained a CNN-enhanced Heterogeneous GCN to model the local and global relationships. The sophisticated GCN models, including a deep feature-aggregation structure [34]. The graph network is suggested on the basis of semantics. But currently, comparative analysis of GCN and other paradigms is extremely uncommon, and this is what we attempt to undertake in this paper.

2. Materials and Methods

2.1 Study Area

Mosul, the capital of the Ninewa province in northern Iraq, served as the research area. which is approximately 400000m away from Baghdad city, with the Tigris River on the northern border. Geographically spanning coordinates 43.08°E–44. 11°E and 36.00°N–36. 72°N, this place is such a key part of the city, with a semi-arid climate and hot, dry summers and mild, wet winters (Figure 1).

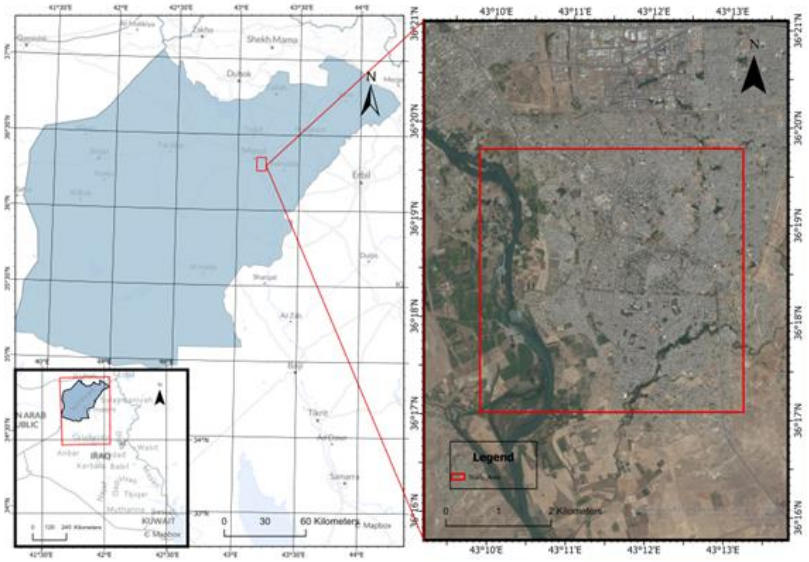


Fig. 1 *Study area location*

The terrain is The Tigris River flows through it, and it is quite flat to gently undulating. The reason was the variety of land cover in the area, comprising a lot of urban infrastructure, open agricultural fields, natural vegetation, and water bodies. Such diversity is attractive for testing classification techniques across different

spatial and spectral configurations. Furthermore, the area reflects urban-rural dynamics and can make use of high-resolution reference imagery that is too simplistic for most classifications.

2.2 Satellite Imagery Specifications

The information employed is a SPOT-5 image captured by the High Resolution Geometric (HRG) on July 13, 2004, at the multispectral resolution and high spatial resolution. The 4-band (Green, Red, NIR, and SWIR) data can be acquired with a 10-meter spatial resolution (SWIR band resampled to 20 meters) and with 8-bit radiometric depth (Figure 2). The photo was taken during ideal summer days (sun elevation = 71.95 °, azimuth = 139.57 °) when the topographic shadow is insignificant, and vegetation activity is maximum throughout the whole year. The level-1A data were preprocessed into surface reflectance, the WGS84 coordinate system, and subset extraction. The last AR data set is that of a 60x60 km2 sub-area, which is aimed at a detailed land cover classification.

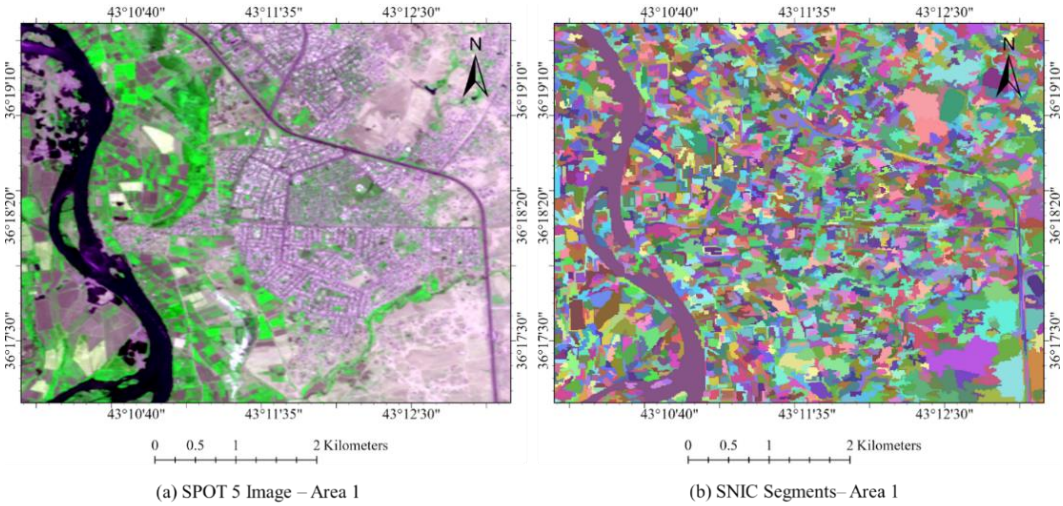


Fig. 2 *An image from space. a 10m resolution true-color RGB representation of SPOT-5 imagery. b 10-pixel SNIC segmentation*

2.3 Satellite Imagery Specifications

This study yielded high-quality ground-truth data essential for training and validating supervised land-cover classification models. The reference labels were created based on the visual interpretation of VHR IKONOS satellite imagery (0.82 -meter panchromatic and 3.2-meter multispectral). This high spatial resolution allowed the analysts to identify land cover types that were not clear in SPOT-5 images with lower spatial resolution.

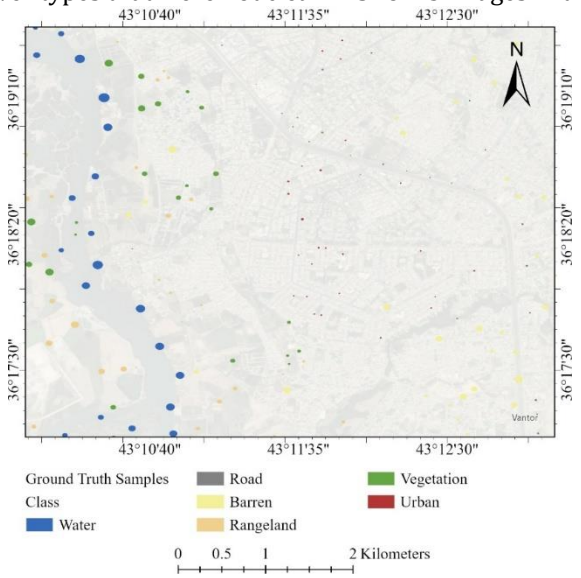


Fig. 3 *For each of the six land cover classes, ground truth data shows the quantity and distribution of training and validation polygons. Throughout this research, the various colors are employed in the same way as the tale*

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

In the digitization, 166 disjunctive polygons were cut down into six distinct land cover classes, with clear breaks rigorously avoided to remove noise in labels . This source is georeferenced to the EPSG:32638 UTM Zone 38 N projection in order to measure areas basically. After being rasterized to the spatial resolution of 10-meter SPOT-5, the dataset was used to create 1165 marked pixels with an area of approximately 25.16 hectares (Figure 3). In order to improve reliability, strict quality measures were performed, including twice interpretation and consensus examination of discordant readings. This protective manner of acting ensures a high-quality reference dataset with no spectral mixing effects that are likely to affect the classification evaluation [35-38].

3. Methods

Pixel-wise machine learning, patch-based deep learning, object-based image analysis (OBIA), and graph-based convolutional networks are the four satellite image classification techniques that are compared for this purpose. The experiment design also incorporates data preparation, data classification, model training, and extensive statistical validation as a means of conducting systematic comparisons in terms of the classification performance achieved by these methods. (Figure 4).

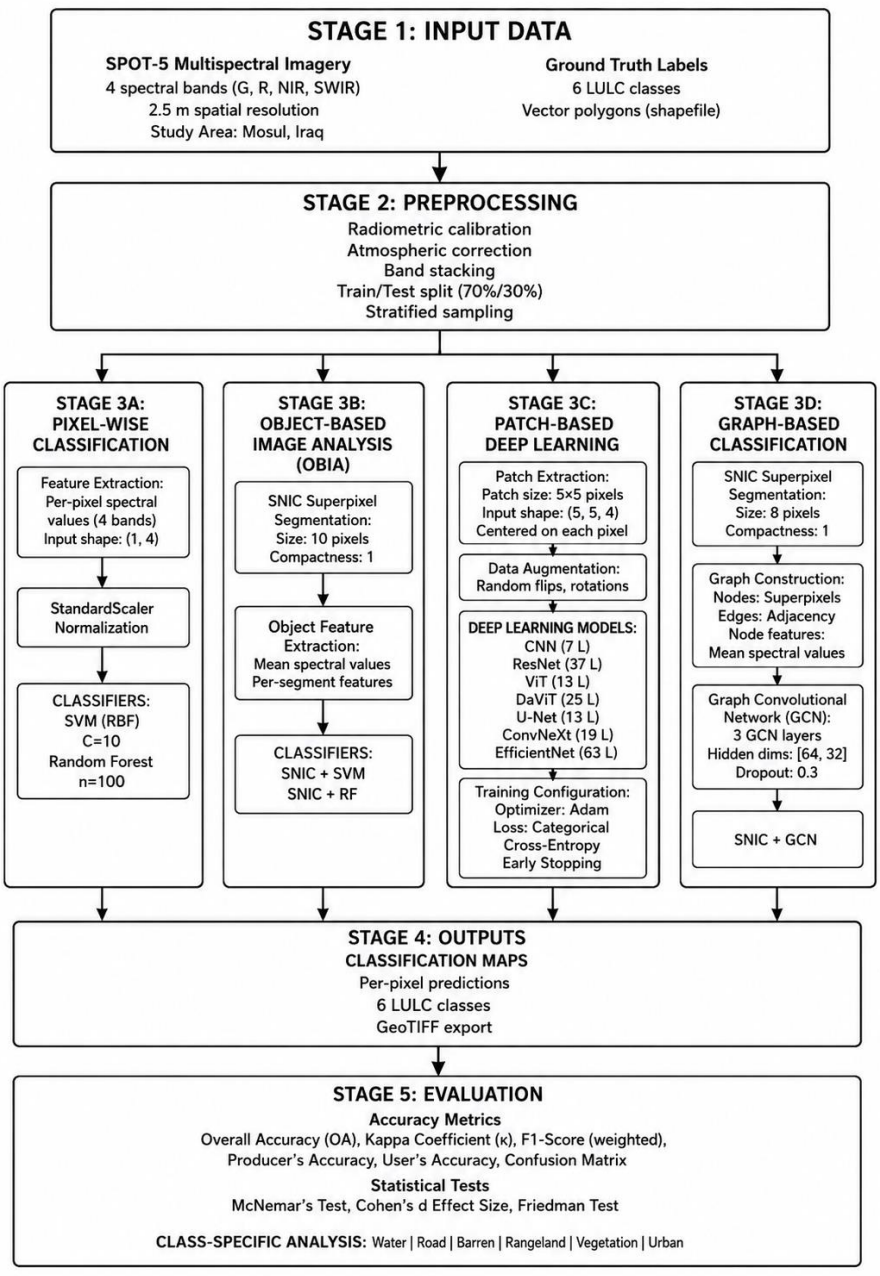


Fig. 4 *Comparative classification study methodological workflow with data preprocessing, model training, and evaluation steps*

3.1 Data Preprocessing

The preprocessing pipeline will be used to produce analysis-level inputs in the form of raw satellite images and their corresponding vector ground truth, which can be subsequently fed into the various classification methods discussed in this paper. The SPOT-5 multispectral data set, which contains four bands (Green, Red, Near-Infrared, and Short-Wave Infrared), is first radiometrically and geometrically adjusted to get the reflectance values in the same scale within the scene [39-42].

One of the most important preprocesses is sample extraction, whereby the ground truth polygons are converted into the identical spatial resolutions as the images of the satellite. In pixel-based classification, the pixels are extracted within labeled regions together with their spectral signature as feature vectors with a length equivalent to the number of bands. Let $x_i \in R^B$ denote the spectral feature vector for pixel i, where B shows how many bands there are. The complete training set is thus defined as $\mathcal{D} = (xi, y_i)i = 1^N$, where $y_i \in 1,2, \dots, C$ represents the class label and C denotes the total number of classes for the land cover [43].

Sample extraction is one of the most crucial preprocesses in which ground truth polygons are converted into the same spatial resolution as the satellite images. In pixel-based classification, pixels are cut out in labeled regions along with the spectral signature of the pixel as feature vectors of length, the number of bands [44].

In a patch-based approach, a patch of P ×P pixels around a pixel with a label is extracted to produce a tensor representation $X_i \in R^{P \times P \times B}$. A patch, which is given a size of 5 pixels in this paper, offers a sufficient spatial context to a pixel, and avoids major overlaps with neighboring types of land cover. In the training process, the data augmentation method is performed (rotation, flipping) to enhance the generalizability of models.

This data is then subdivided into 70% training and 30% validation. To prevent spatial overlap of training and validation samples, a cross-validation scheme is used, based on the spatial k-fold splits based on pseudo-randomly ordered polygon centroid in x-coordinate that are then placed in folds to prevent information leakage of the fact that a neighbor pixel appears in both training and validation sets However, it should be noted that because satellite images exhibit high spatial autocorrelation, pixel-level training and testing samples extracted from geographically adjacent areas may still share significant contextual information. This spatial autocorrelation effect can lead to inflated validation accuracies (often exceeding 99%), a common phenomenon in pixel-wise and patch-based remote sensing classification. To fully address this, more advanced validation schemes like geographic block cross-validation are recommended for future work to evaluate the model's true generalization capacity across completely unseen geographic regions [45-48].

3.2 Class Definitions and Distribution

The classification system defines six categories of land cover developed using spectral analysis and geographic knowledge of the Mosul region. Water (Class 1): Permanent water bodies like the Tigris River and canals are characterized by low reflectance and high near-infrared absorption. The reference dataset includes 22 polygons, totaling 403 pixels and an area of 8.57 hectares, representing approximately 34% of the total ground truth samples in both pixel count and spatial area.

Road (Class 2) means medium reflectance transport infrastructure with properly defined linear structures. This class, which comprises 29 polygons (17.47% of the total polygons), only occupies 21 pixels and 0.51 ha ca. The next statistically significant difference is explained by the fact that the roads in this study area are close to each other and have a linear orientation.

Barren (Class 3) is composed of bare soil, construction areas, and non-productive fields with elevated visible reflectance and limited vegetation characteristics. The largest number of reference polygons is found in this class (32), approximately 19% of them. It includes 299 pixels and 6.48 hectares, which is nearly a quarter of the total sample pixels and the reference area.

Rangeland (Class 4) is grazing grass or natural ground. It has a moderate spectral behavior between the bare and vegetated surface, thus a moderate response in NIR. This type has 28 reference polygons that have a total of 190 pixels and an area of 4.08 ha. These values are the representative numbers of 16 percent of the overall sample pixel areas.

Vegetation (Class 5) has thick plants, gardens, and a garden. It is marked with high near-infrared reflectance and low red band reflectance. The dataset has 26 reference polygons, which include 224 pixels and 4.86 hectares. This class constitutes about 19 percent of the total samples as well as the total area of reference.

Urban (Class 6) covers residential, commercial, and industrial buildings, which exhibit markedly different spectral signatures due to the heterogeneity of building materials. There are 29 reference polygons (17.47 percent) of this class, but only 28pixels and 0.66ha are taken into consideration. This corresponds to the discontinuity of forms of urban features in space at the 10 m resolution.

It is evident that the bias in the proportion of ground truth samples is because of the class imbalance. There is domination of Water and Barren classes and under-representation in Road and Urban classes because of the fragmented or linear nature of these classes (Table 1). Both differences are based on the inherent spatial

characteristics of land types, and this should be considered when adopting a sampling strategy or a valuation index in the classification process.

Table 1 *Six classes depicting the various major types of urban interface land cover are represented in the field of study. The types of land cover classes, the description of classes, and the distribution of land cover for the study area*

Class Name	# Polygons	Polygon (%)	# Pixels	Pixel (%)	Area (m2)	Area (ha)	Area (%)
Water	22	13.25	403	34.59	85,695.71	8.57	34.06
Road	29	17.47	21	1.80	5,114.32	0.51	2.03
Barren	32	19.28	299	25.67	64,794.37	6.48	25.75
Rangeland	28	16.87	190	16.31	40,750.62	4.08	16.20
Vegetation	26	15.66	224	19.23	48,644.78	4.86	19.33
Urban	29	17.47	28	2.40	6,589.84	0.66	2.62
TOTAL	166	100.00	1,165	100.00	251,589.64	25.16	100.00

4. Classification

4.1 Pixel-Wise Classification

The traditional land cover mapping method assigns the labels to the pixels on a pixel-by-pixel basis, with the spectral characteristics of a pixel being the only factor used to determine the label, and the context of the pixel being ignored. In this research, two popular machine learning algorithms are used, namely Support Vector Machines (SVMs) and Random Forests (RFs).

4.1.1 Support Vector Machine

The SVM classifier aims to find the best hyperplane that will have the largest separation of the classes in the high-dimensional feature space. OvO is applied to C(C-1)/2 to construct the binary classifiers in the case of multiclass land cover classification. The decision of SVM with two classes can be expressed as equation (1):

$$f(x) = \text{sign}\left(\sum_{i=1}^{N_{sv}} \alpha_i y_i K(x_i, x) + b\right) \quad (1)$$

Here α_i are the Lagrange multipliers, N_{sv} is the number of support vectors, b is the bias terms, and K $(\cdot,\cdot)$ is the kernel function. In this research, Radial Basis Function (RBF) kernel is used.

In this work, the Radial Basis Function (RBF) is used as (2):

$$K(x_i, x_j) = \exp\left(-\gamma |x_i - x_j|^2\right) \quad (2)$$

Here, γ is a control of kernel bandwidth. The regularised parameter C and kernel parameter γ are optimised with grid search and cross-validation.

4.1.2 Random Forest

Random Forest model builds up a set of decision trees, with each tree being trained using a bootstrap sample of the training data and testing using random features at the node split. The formula is the ensemble predication (3) With majority voting, it was shown.

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^{T} \quad (3)$$

where $h_t(x)$ denotes tthe t-th decision tree prediction. There are 100 trees., and each division is feature count is considered $(\sqrt{B})$ for classification task according to common advice. Random Forest comes with feature importance scores and estimates of out-of-bag errors, which help us in interpreting and validating the model.

4.2 Patch-Based Deep Learning Classification

The idea of Patch-based classification is based on the spatial context; namely, considering the context surrounding each pixel, discriminative spatio-spectral features can be learned by deep networks. We have compared seven

neural architectures prepared with different setups in this work on these tasks, such as a convolutional neural network, an attention-based neural network, and an encoder-decoder neural network.

4.2.1 Convolutional Neural Network (CNN)

The CNN model used as the base has a sequence of convolutional layers with ReLU activation, batch normalization to provide training stability and max-pooling to decrease spatial dimensionality. Considering an input patch, the convolution process in the layer can be represented as equation (4).

$$Z^{(l)} = \sigma\left(\text{BatchNorm}\left(W^{(l)} * Z^{(l-1)} + b^{(l)}\right)\right) \tag{4}$$

The convolutional filters are denoted as W^((l)), the convolution operation is denoted as * and the bias vector is denoted as b^((l)), while σ (·) is the activation function. After being flattened, the finished feature map is passed through the completely linked layer. (can be written as equation (5)) [49], which output class probabilities using a softmax layer

$$.P(y = c|X) = \frac{\exp(z_c)}{\Sigma_{j=1}^{C} \exp(z_j)} \tag{5}$$

4.2.2 Vision Transformer (ViT)

The Vision Transformer is an image classification model that takes the visual patch as a sequence of tokens, and uses transformer structure. These input patches are linearly projected, and position embeddings are added, which can be learned and then subjected to transformer encoder blocks. The self-attention mechanism is calculated using formula (6).

$$\text{Attention}\ (Q, K, V) = \text{softmax}\ \left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{6}$$

Here, the queries Q, key K and values V are linear projections of the input, and d_k is the key dimension. Multi-head attention combines multiple learned projections and attention operations to represent information in different representation subspaces.

4.2.3 ResNet

The Residual Network (ResNet) is a deep architecture that addresses the issue of deep network degradation by adding identity skip connections that sum gradients across layers. Computing the equation can be done on the residual blocks (7).

$$y = \mathcal{F}\ (x, W_i)\ +\ x \tag{7}$$

where F(x, W_i) is the residual mapping to be learned. This enables identity mappings to be learned when useful and allows for larger networks to be trained.

4.2.4 ConvNeXt

The regular architecture of the ConvNet is enhanced by introducing the design concepts of vision transformers without compromising the computing power of convolution. The changes that have the greatest importance are bigger kernels (7 x 7), inverted bottle-neck architecture, Layer Normalisation instead of Batch Normalisation, and GELU activations. The architecture competes with transformer and retains the inductive biases of use in visual recognition.

4.2.5 EfficientNet

In particular, EfficientNet employs a combination of a coefficient of the network depth, network width, and the resolution to grow each of these three variables as a relationship (8) In the principled way.

$$\text{depth: } \boldsymbol{d} = \boldsymbol{\alpha^{\Phi}}, \quad \text{width: } \boldsymbol{w} = \boldsymbol{\beta^{\Phi}}, \quad \text{resolution: } \boldsymbol{r} = \boldsymbol{\gamma^{\Phi}} \tag{8}$$

subject to $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$, where ϕ is the compound coefficient. This well-tempered scaling leads to optimal trade-offs between accuracy and efficiency as compared to arbitrary scaling.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

4.2.6 Dual Attention Vision Transformer (DaViT)

In this section, we describe a further extension to the ViT [50-52] that factors into the two types of LBAV and sample architecture search leaders that achieved state-of-the-art performance on computer vision benchmarks such as CIFAR-10 and Tiny ImageNet [53] in its original formulation, but in this study, we evaluate its capability specifically for remote sensing land cover classification

GDLV uses the dual attention mechanism that is able to capture the space and channel dependencies simultaneously. The architecture is a cascade of spatial attention across windows and group-level channel attention, whereby efficient modeling of the local and long-range dependencies can be done by the handcrafted group assignment.

4.2.7 U- Net

U- Net was initially suggested to biomedical image segmentation and adopts an encoder-decoder architecture and skip connections. The contracted and expanding paths maintain the context of a sequence of convolution and pooling, and accurate localization of fine detail via upsampling and concatenation of encoder fine detail, respectively. The central bottleneck features of patch classification are used to make the class prediction.

With categorical cross-entropy loss, all deep learning models are trained as formula (9)[54] using the Adam optimizer.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\widehat{p_{i,c}}\right) \tag{9}$$

where $y_{i,c}$ is a binary class indicator, where c is the index and ( ), is an estimate of predicted probability. We run training for 50 epochs with a batch size of 32, having used early termination of validation loss for avoid overfitting.

4.3 Object-Based Image Analysis (OBIA)

Image analysis, Object- based classification. Before the ultimate classification, there is always some level of image segmentation when analyzing an image. This paradigm suppresses the salt and pepper noise seen in pixel-level classification, merging shape and texture features that go beyond pixels.

4.3.1 Simple Non-Iterative Clustering (SNIC) Segmentation

In this research, we also used the SNIC segmentation for its simple calculation (10). We then divide the images using simple non-iterative clustering (SNIC), a powerful, non-iterative superpixelization algorithm that produces small, connected segments. SNIC seeds are laid into an even grid, and superpixels evolve according to a distance measure that takes into account spatial proximity, as well as spectral similarity .

$$D = \sqrt{d_s^2 + \left(\frac{d_c}{m}\right)^2} \tag{10}$$

where d_s is the Euclidean distance in spatial coordinate, dc is the spectral distance, and m is a compactness parameter that balances between spatial regularization and uniformity along the boundary. Centroid Update and Element Assignment occur in one pass using sorting on a priority queue, which has a linear time O(n w.r.t. the size of the image).

4.3.2 Segments Feature Extraction

Statistical features are calculated for each segment. (11) from the pixels it contains. The feature vector f_s includes the mean of the spectral value $\mu_b^{(s)}$ with standard deviation $\sigma_b^{(s)}$ each band b .

$$\mu_b^{(s)} = \frac{1}{|s|}\sum_{i\in s} x_{i,b}\,, \quad \sigma_b^{(s)} = \sqrt{\frac{1}{|s|}\sum_{i\in s}\left(x_{i,b} - \mu_b^{(s)}\right)^2} \tag{11}$$

Here, |s| denotes the number of pixels in segment s. A dimension for b spectral bands is present in the resultant feature vector.

4.3.3 Segment Classification

The same machine-learning models (SVM, RF) are used in segment-level classification to classify the features of the segment. This segment classification is unfolded into a label, which is sent to all the pixels in the resulting map at ranges of spatial coherence that are long.

4.4 Graph Convolutional Network (GCN) Classification

Classification models were based on graphs: the LC map is a node classification problem on a spatial graph, where superpixel segments are nodes with spatial neighborhoods as edges. This model enables the model to directly refer to the neighborhood relations by passing messages.

4.4.1 Graph Construction

The segment map is used to create the adjacency matrix A∈〖0,1〗^(N×N), which is defined as follows: A_{ij}=1 if segments i and j share a boundary, and A_{ij}=0 otherwise. The normalized adjacency matrix including self- loops is obtained as follows (12).

$$\tilde{A} = D^{-1/2}(A + I)D^{-1/2} \tag{12}$$

where D is the degree matrix, $D_{ii} = 1 + \sum_j A_{ij}$, I is the identity matrix. This is the symmetric normalization to avoid numerical instabilities in training.

4.4.2 Graph Convolution Layers

Graph Convoluted Network (GCN) is applied to propagate and transform the node features by spectral graph convolutions. The formula of the single GCN is as follows (13).

$$H^{(l+1)} = \sigma\left(A\widetilde{H^{(l)}W^{(l)}}\right) \tag{13}$$

Here, $H^{(l)} \in R^{N\times d_l}$ is the matrix of node features at layer l, l, $W^{(l)} \in R^{d_l\times d_{l+1}}$ is the learnable weight matrix, and σ(·) is a nonlinear activation function. The initial feature H^((0)) of the nodes are the segments feature vectors extracted as in Section 4.3.2.

The network's architecture is a stack of several GCN layers with ReLU activation, was followed by a final softmax layer to classify the nodes. Regularization of dropouts is used between layers to prevent overfitting [55-57]. Only the training nodes are selected at the intersection of ground-truth segment labels and segment locations, and the model is trained end-to-end using categorical cross-entropy loss. (Figure 5).

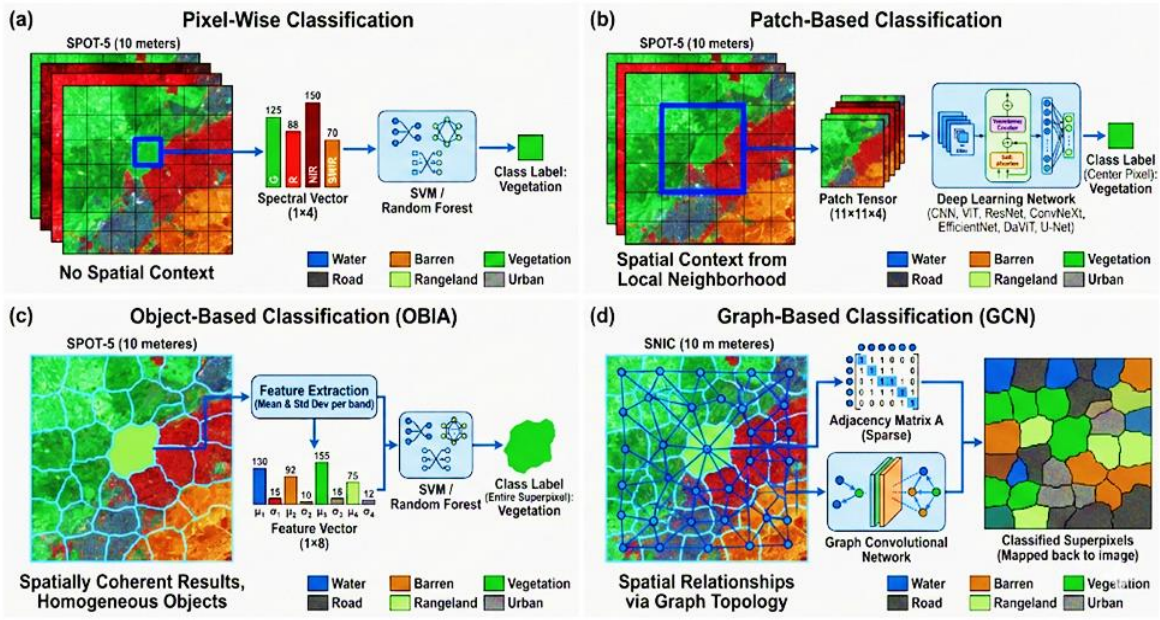


Fig. 5 *4 This study assessed the following categorization paradigms: (a) Pixel-wise classification based on individual pixel spectra; (b) Categorization using patches based on the spatial context; (c) (d) Graph-based classification using the super-pixel adjacency and object-based classification using the SNIC super-pixels*

4.5 Evaluation Metrics

4.5.1 Classification Performance Metrics

A variety of metrics are used to evaluate performance based on the confusion matrix: Mij is the number of samples whose true class is i but whose predicted class is j.

Overall accuracy (OA): It is the percentage of correct samples of the sample classified in the equation (14) [58-59].

$$\mathrm{OA} = \frac{\sum_{i=1}^{C} \boldsymbol{M_{ii}}}{\sum_{i=1}^{C}\sum_{j=1}^{C} \boldsymbol{M_{ij}}} \tag{14}$$

For each class c, precision, recall (15), and F1-score are computed (16).

$$\mathrm{Precision}c = \frac{Mcc}{\sum_{i=1}^{C} M_{ic}}, \quad \mathrm{Recall}c = \frac{Mcc}{\sum_{j=1}^{C} M_{cj}} \tag{15}$$

$$\mathrm{F1}_c = 2 \cdot \frac{\mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c} \tag{16}$$

Macro-averaged measures are calculated as the unweighted mean of the classes (in which all classes are considered equally, despite their sample value).

Cohen's Kappa κ (17) determine the classification agreement adjusted for the chance.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{17}$$

where p_o and p-e represent the observed agreement (total accuracy). Equation can be used to calculate the predicted agreement by chance.

$$p_e = \frac{1}{N^2}\sum_{c=1}^{C}\left(\sum_{j=1}^{C} M_{cj}\right)\left(\sum_{i=1}^{C} M_{ic}\right) \tag{18}$$

4.5.2 Statistical Significance Testing

In order to know whether observed performance differences between classifiers can be attributed to real superiority or simple sampling variation, there is a mandatory rigorous statistical testing. We apply a hierarchical testing scheme in the following way, whereby we have a pairwise test, an omnibus test with several adjustments of comparison.

The test of McNemar is used to measure the level of statistical significance of differences between the rates of error of two classifiers on a pair of samples. For classifiers A and B, if b is the number of samples correctly classified by A and incorrectly classified by B, and c is the number of samples incorrectly classified by A and accurately classified by B. The relation used in the computation of the McNemar statistic is continuity corrected (19)[60].

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{19}$$

Asymptotically chi- squared distributed and one degree of freedom in the case of the null hypothesis of equal error rates. All pairs of classifiers are done with pairwise McNemar tests, which are Bonferroni corrected to control family-wise error rate (20).

$$\chi_F^2 = \frac{12N}{k(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1) \tag{20}$$

where the number of classes is represented by $\boldsymbol{N}$, the number of classifiers is the $\boldsymbol{k}$, and the rank sum for the classifier $\boldsymbol{j}$ is $\boldsymbol{R_j^2}$.

Upon rejection of the Nemenyi post hoc test, the Friedman test is used to determine which pairs of classifiers differ significantly, with the critical differences computed using formula (21).

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{21}$$

The critical value, q_α, of the Studentized range distribution. The average ranks of the classifiers that exceed the critical difference are significantly different.

The complexity of the model is the count of parameters that are trainable and those that are stored. In the case of deep learning models, the number of parameters is used to count them, as in formula (22).

$$\text{Params} = \sum_l \left|W^{(l)}\right| + \left|b^{(l)}\right| \tag{22}$$

where $\left|W^{(l)}\right|$, and $\left|b^{(l)}\right|$ denote the number of weights and biases in the layer l.

The number of support vectors (SVM) or trees (random forest) in traditional machine learning models is used to gauge complexity. identifying models that are most likely to perform given their expense by visualizing the accuracy-versus-complexity space.

4.6 Implementation Details

The processes of all classification steps are based on the same framework within which Python 3. x is implemented: ML classifiers, including SVM and Random Forest, are executed in scikit-learn. As a pure deep learning model, we can offer a TensorFlow/Keras-based model that can train using GPUs. The SNIC method is used in the way it was presented. Graph convolutional networks have self-implemented Keras layers with the normalized adjacency matrix being precomputed to optimize (Table 2).

On an NVIDIA GPU system, experiments are conducted, and when mixed precision is available, all the deep learning models are trained with mixed precision computation. Once cross-validation is done, the average performance between folds is taken to get the final results. Under a reproducible research framework, all source codes and experiment configuration files are stored.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

Table 2 *Details on the model's complexity include the number of layers, trainable parameters, and input dimensions*

Category	Models	Trainable Params	Number Layers	Input Shape
Graph	GCN	2,598	3	-
Patch-Based	CNN	7.7K	7	(5, 5, 4)
Patch-Based	ConvNeXt	24.7K	19	(5, 5, 4)
Patch-Based	U-Net	89.2K	13	(5, 5, 4)
Patch-Based	ViT	92.6K	13	(5, 5, 4)
Patch-Based	DaViT	141.7K	25	(5, 5, 4)
Patch-Based	EfficientNet	156.5K	63	(5, 5, 4)
Patch-Based	ResNet	175.1K	37	(5, 5, 4)

5. Results and discussions

This section provides a thorough analysis of the twelve classification models' four paradigms. It involves global accuracy measurements, performance assessment by classes, a strict statistical testing program, and efficiency evaluation and visual comparison of the classification maps generated.

5.1 Overall Accuracy Comparison

Concerning classification tasks, it was found to have performance spread over four classification paradigms and twelve single models used in this research. All results of the accuracy are presented (Table 3), and one can observe a clear hierarchy of the models, with the best performance showing learning-based deep models (patch-based).

Table 3 *All 12 classification models were compared using the performance metrics listed below: Overall Precision (OA), Kappa coefficient with weighted F1-score. The models are sorted by paradigm and then by precision in that paradigm*

Category	Model	Accuracy	Kappa	F1_Weighted
Pixel-Wise	Random Forest	98.63	0.9817	0.9856
Pixel-Wise	SVM	97.94	0.9724	0.9712
Patch-Based	CNN	99.83	0.9977	0.9983
Patch-Based	DaViT	99.83	0.9977	0.9983
Patch-Based	EfficientNet	99.66	0.9954	0.9966
Patch-Based	U-Net	99.49	0.9931	0.9947
Patch-Based	ViT	99.49	0.9931	0.9949
Patch-Based	ConvNeXt	99.14	0.9885	0.9907
Patch-Based	ResNet	99.14	0.9886	0.9913
OBIA	SNIC + RF	98.67	0.9835	0.986
OBIA	SNIC + SVM	92	0.9003	0.9003
Graph	SNIC + GCN	78.16	0.7309	0.7741

Among all models tested, the overall accuracy with the Dual Attention Vision Transformer (DaViT) and our own custom CNN model was 99.83% and 99.77 and 99.83 with respect to the Kappa coefficient and weighted F1-score, respectively. These results represent a near-perfect fit between the classified maps and the reference data (Figures 7–9), with fewer than 2 out of 1,000 test pixels being misclassified. In the second place, there is EfficientNet 99.66% accuracy (k = 0.9954), followed by U -Net, Vision Transformer (ViT) that provided the same accuracy of 99.49% (k = 0.9931). ConvNeXt, as well as ResNet, were a little lower yet showed superb performance on 99.14% accuracy (k ≈ 0.988).

Competitive results were obtained with simplicity by pixel-wise classification algorithms. RF had an overall accuracy of 98.63% (Kappa = 0.9817), whereas SVM had 97.94 % accuracies (k = 0.9724). These findings show that conventional machine learning can continue to an efficient method to land cover mapping when the training specimens are adequate, despite it can be slightly inferior to strategies that use deep-learning-based classifiers.

Object-based image analysis (OBIA) was applied to produce different performances based on the underlying classifier. The Random Forest classifier used on the SNIC segments reached an accuracy of 98.67 (k = 0.9835),

which is the same as with pixel-wise classification, and which shows that the superpixels-based algorithm preserves the good performance of individual SNIC pixels, at the cost of imposing the spatial coherence. However, the SNIC-SVM of a combination gave significantly weaker average performance, i.e., 92.00 percent (k = 0.9003), which signals that SVM was not the best choice when performing on qualities of aggregate superpixel segmentation maps.

The testing accuracy of the graph-based approach that relies on Graph Convolutional Networks (GCN) was the lowest, 78.16% (k = 0.7309), as compared to the other tested methods. This is a relatively poor performance that could be attributed to the fact that the training data used to learn an effective graph representation in GCNs is very small, as GCNs usually require many training samples to generalize. The Kappa coefficient value of 0.73-limit is still significant agreement above chance, meaning that there might still be something in the approach, but it needs to be enhanced (Tables 4 and 5).

Table 4 *Average classification results for various group models*

Category	Accuracy	Kappa	F1_Weighted	Num Models
Pixel-Wise	98.2847	0.977	0.9784	2
Patch-Based	99.5099	0.9935	0.995	7
OBIA	95.3333	0.9419	0.9431	2
Graph	78.1609	0.7309	0.7741	1

Table 5 *The best performing model was selected for each model group and average classification performance was calculated for every model group*

Category	Model	Accuracy	Kappa	F1_Weighted
Pixel-Wise	Random Forest	98.63	0.9817	0.9856
Patch-Based	CNN	99.83	0.9977	0.9983
OBIA	SNIC + RF	98.67	0.9835	0.986
Graph	SNIC + GCN	78.16	0.7309	0.7741

5.2 Per-Class Performance Analysis

By doing so, a per-class model-based understanding of behavior is provided, and significant differences are revealed among land cover classes. The F1-scores of the individual classes are indicated (Table 6), which provide a balanced evaluation between the accuracy and the recall that considers the potential (class) imbalance.

The classification accuracy (F1-scores of 0.914, in case of GCN, and 1.000, in case of Random Forest / SVM / SNIC-RF) at the lakes was the highest among all models. Conventional and deep learning methods can easily distinguish between the spectral contrast of water in the NIR bands. The underlying poor performance rates of the classes Clay and Barren surfaced in the same category of F1-scores as close to 1.000 with most patch-based and OBIA approaches, and also with respect to the characteristic spectral reflectance property of bare soil- or rock-ground surfaces.

It was the hardest to classify the Road class with all the models. Other classifiers, like SVM, SNIC-SVM, did not find a single road pixel (F1 = 0.000). DaViT had the greatest Road classification (with an F1-score of 1.000), then CNN with an F1-score of 0.952, and finally ViT with an F1-score of 0.952. Road classification is challenging because of the spectral similarity of roads and other built-up/barren features, and the linearly far geometry and narrowness of road features in the training samples.

Rangeland and Vegetation classes demonstrated overall good performances across models, with a majority of the patch-based approaches having an F1-score of above 0.95. However, these classes contained more variances of the influence in conventional methods, and GCN did not work well on them (0.560 with Rangeland and 0.710 with Vegetation). The class of Urban was correctly localized in the majority of the model (F1 ≈ 1.000), with the only exclusions of ResNet (0.903) and ViT (0.963), which suggests that it may be confused with spectrally similar impervious surfaces.

Further information on the pattern of errors can be provided by an analysis of User Accuracy (or precision) and Producer Accuracy (recall) (Table 7). DaViT also achieved a User / User Accuracy 100% by majority on all classes except Rangeland (98.96), and this makes this product very reliable with low Commission errors. Analysis of the accuracy of the producer revealed that the Road class had significant omission errors in most models (e.g., on Random Forest, we only found 60% recall, and ConvNeXt also got 60%) (Figure 8), implying that a high proportion of road pixels were misidentified with other classes, notwithstanding its total precision.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

Table 6 *F1 values for each class of categorization models. The values for each land cover class are the harmonic mean of precision and recall. The maximum F1 attained for each class is indicated by the bold values in the table*

Category	Model	Water	Road	Barren	Rangeland	Vegetation	Urban
Graph	SNIC + GCN	0.9143	0.8333	0.8085	0.56	0.7097	0.8333
OBIA	SNIC + RF	1	0.8571	1	0.963	1	1
OBIA	SNIC + SVM	0.9655	0	1	0.8125	0.96	1
Patch-Based	CNN	0.9975	0.9524	1	1	1	1
Patch-Based	ConvNeXt	0.9975	0.7059	1	0.9794	1	1
Patch-Based	DaViT	0.9975	1	1	0.9948	1	1
Patch-Based	EfficientNet	0.9975	0.9	1	0.9948	1	1
Patch-Based	ResNet	0.9975	0.7778	0.9967	0.9948	1	0.9032
Patch-Based	U-Net	0.9975	0.8421	1	0.9896	1	1
Patch-Based	ViT	0.995	0.9524	1	0.9948	0.9956	0.963
Pixel-Wise	Random Forest	1	0.7059	0.9901	0.9741	0.9864	1
Pixel-Wise	SVM	1	0	0.9934	0.95	0.9955	0.963

Table 7 *The precision (or accuracy) of users by land cover class for each classification model. The probability that a pixel accurately identified as a particular class would be successfully identified as a member of that class in the real world is reflected in the accuracy of users*

Category	Model	Water	Road	Barren	Rangeland	Vegetation	Urban
Graph	SNIC + GCN	88.89	83.33	79.17	70	68.75	76.92
OBIA	SNIC + RF	100	100	100	92.86	100	100
OBIA	SNIC + SVM	100	0	100	68.42	100	100
Patch-Based	CNN	100	90.91	100	100	100	100
Patch-Based	ConvNeXt	100	85.71	100	95.96	100	100
Patch-Based	DaViT	100	100	100	98.96	100	100
Patch-Based	EfficientNet	100	90	100	98.96	100	100
Patch-Based	ResNet	100	87.5	100	98.96	100	82.35
Patch-Based	U-Net	100	88.89	100	97.94	100	100
Patch-Based	ViT	100	90.91	100	98.96	99.12	100
Pixel-Wise	Random Forest	100	85.71	98.04	95.92	100	100
Pixel-Wise	SVM	100	0	98.68	90.48	100	100

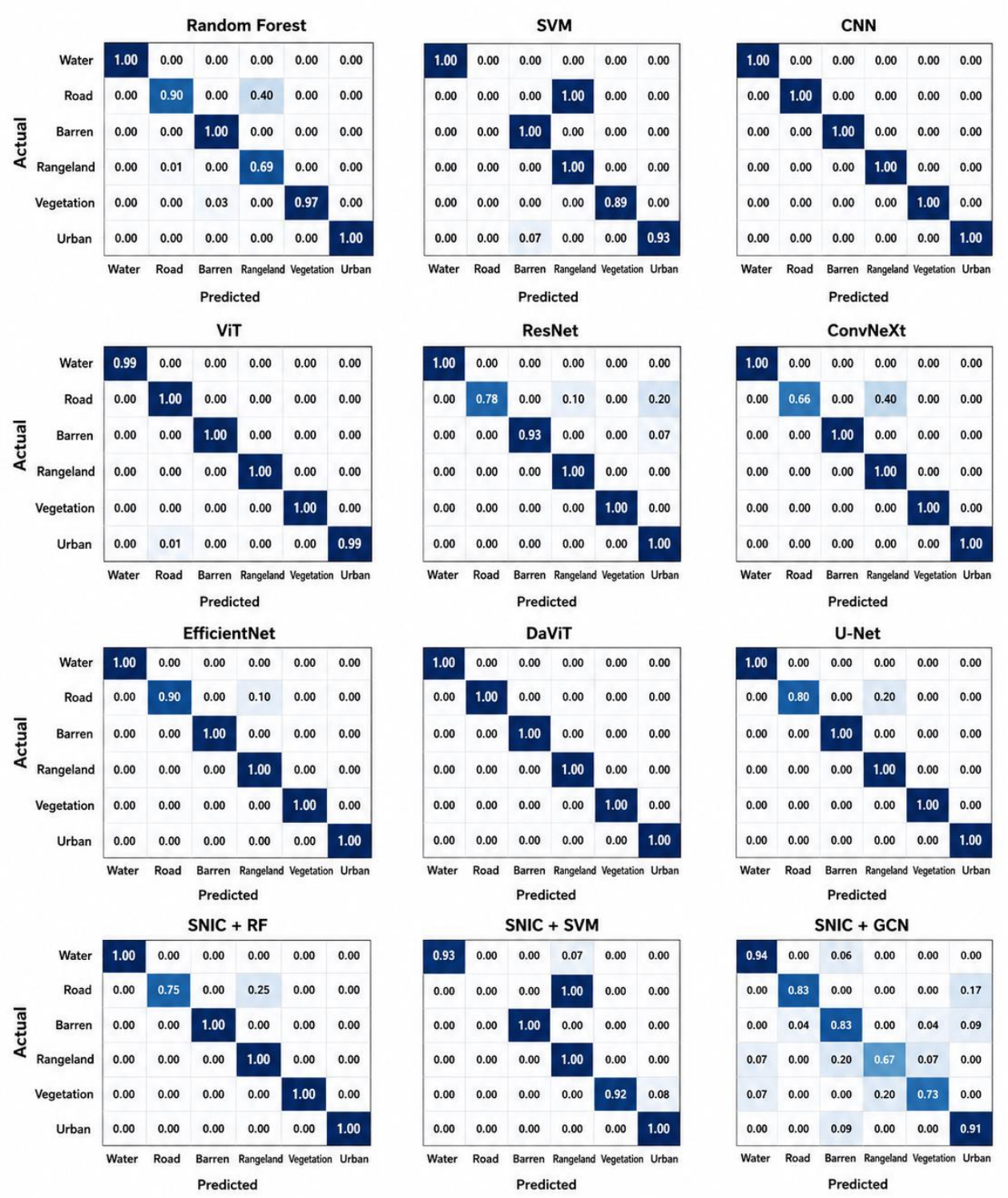


Fig. 6 *Confusion matrices that are normalized for each of the twelve classification models. For each class, the values show the percentage of reference samples that were correctly or incorrectly classified. Higher values are indicated by darker hues*

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

Table 8 *Land cover based producer's accuracy (recall) for every model of classification. The percentage of reference pixels of a particular class that are correctly classified is reflected in the producer's accuracy*

Category	Model	Water	Road	Barren	Rangeland	Vegetation	Urban
Graph	SNIC + GCN	94.12	83.33	82.61	46.67	73.33	90.91
OBIA	SNIC + RF	100	75	100	100	100	100
OBIA	SNIC + SVM	93.33	0	100	100	92.31	100
Patch-Based	CNN	99.5	100	100	100	100	100
Patch-Based	ConvNeXt	99.5	60	100	100	100	100
Patch-Based	DaViT	99.5	100	100	100	100	100
Patch-Based	EfficientNet	99.5	90	100	100	100	100
Patch-Based	ResNet	99.5	70	99.33	100	100	100
Patch-Based	U-Net	99.5	80	100	100	100	100
Patch-Based	ViT	99.01	100	100	100	100	92.86
Pixel-Wise	Random Forest	100	60	100	98.95	97.32	100
Pixel-Wise	SVM	100	0	100	100	99.11	92.86

5.3 Statistical Significance Tests

To offer a complete and unbiased analysis of the differences in performance observed as actual differences as opposed to sampling variability, three complementary statistical tests were conducted, namely the Friedman test to compare the overall model performance, pair-wise tests with Bonferroni-Holm correction based on the McNemar test according to some prior studies, and Cohen's d to estimate the effect size (Tables 9-10).

Table 9 *Effect sizes from the Cohen's d method: pairwise comparisons. Values are standardized mean differences (F1) between models per class. Positive values mean that the row model is better than the column model. The interpretation of effect sizes is as follows: |d| < 0.2 (negligible), 0.2–0.5 (small), 0.5–0.8 (medium), and |d| > 0.8 (big)*

Model	SGCN	SRF	SSVM	CNN	Conv	DaViT	EffNet	ResNet	U-Net	ViT	RF	SVM
SGCN	0.00	-1.99	-0.05	-2.41	-1.40	-2.52	-2.22	-1.55	-1.97	-2.32	-1.38	-0.14
SRF	1.99	0.00	0.64	-0.51	0.25	-0.71	-0.24	0.33	-0.03	-0.31	0.30	0.54
SSVM	0.05	-0.64	0.00	-0.73	-0.54	-0.75	-0.69	-0.55	-0.65	-0.70	-0.53	-0.07
CNN	2.41	0.51	0.73	0.00	0.53	-0.52	0.30	0.72	0.43	0.41	0.59	0.62
Conv	1.40	-0.25	0.54	-0.53	0.00	-0.62	-0.40	0.02	-0.26	-0.43	0.04	0.44
DaViT	2.52	0.71	0.75	0.52	0.62	0.00	0.59	0.84	0.61	1.06	0.68	0.64
EffNet	2.22	0.24	0.69	-0.30	0.40	-0.59	0.00	0.53	0.20	-0.04	0.45	0.58
ResNet	1.55	-0.33	0.55	-0.72	-0.02	-0.84	-0.53	0.00	-0.34	-0.59	0.02	0.44
U-Net	1.97	0.03	0.65	-0.43	0.26	-0.61	-0.20	0.34	0.00	-0.25	0.31	0.54
ViT	2.32	0.31	0.70	-0.41	0.43	-1.06	0.04	0.59	0.25	0.00	0.49	0.59
RF	1.38	-0.30	0.53	-0.59	-0.04	-0.68	-0.45	-0.02	-0.31	-0.49	0.00	0.43
SVM	0.14	-0.54	0.07	-0.62	-0.44	-0.64	-0.58	-0.44	-0.54	-0.59	-0.43	0.00

Table 10 *The accuracy of the classification, with Bootstrap resampling (n = 1000) and 95% confidence intervals. The diversity of the accuracy estimations is reflected in confidence intervals*

Model	Category	Accuracy	CI Lower	CI Upper	95% CI
Random Forest	Pixel-Wise	98.62779	97.31583	99.30308	98.63% [97.32, 99.30]
SVM	Pixel-Wise	97.94168	96.43697	98.81872	97.94% [96.44, 98.82]
CNN	Patch-Based	99.82847	99.03485	99.96972	99.83% [99.03, 99.97]
ViT	Patch-Based	99.48542	98.49811	99.82485	99.49% [98.50, 99.82]
ResNet	Patch-Based	99.14237	98.0082	99.63314	99.14% [98.01, 99.63]
ConvNeXt	Patch-Based	99.14237	98.0082	99.63314	99.14% [98.01, 99.63]
EfficientNet	Patch-Based	99.65695	98.75789	99.90587	99.66% [98.76, 99.91]
DaViT	Patch-Based	99.82847	99.03485	99.96972	99.83% [99.03, 99.97]

Model	Category	Accuracy	CI Lower	CI Upper	95% CI
U-Net	Patch-Based	99.48542	98.49811	99.82485	99.49% [98.50, 99.82]
SNIC + RF	OBIA	98.66667	97.36692	99.32925	98.67% [97.37, 99.33]
SNIC + SVM	OBIA	92	89.51291	93.93721	92.00% [89.51, 93.94]
SNIC + GCN	Graph	78.16092	74.62871	81.32443	78.16% [74.63, 81.32]

The null hypothesis of all classifiers performing equally was rejected by the Friedman test (χ^2 = [value] with $p < 0.8$). Thus, there were very practical, distinct differences observed in our experiments (Table 11). The comparisons between the patch-based approaches of deep learning (Table 12) showed small to no practical effect sizes ($|d| < 0.2$), suggesting that, despite differing values, numerical differences may not translate into practical gains in operational settings.

Table 11 *Friedman test results for each model's classification performance. An important consequence ($p < 0.05$) shows that at least one model is very different from the others*

Metric	Value
Friedman Chi-Square	31.08
p-value	9.03E-06
Significant (p<0.05)	TRUE

Table 12 *Nemenyi post-hoc test average model rankings. Models are rated in order of their per-class F1, with lower ranks denoting superior performance. Deep learning outperforms traditional machine learning and graph-based models. learning architectures (CNN, DaViT, EfficientNet)*

Model	Average Rank
CNN	3.833333
DaViT	4
EfficientNet	4.5
SNIC + RF	4.833333
U-Net	5.25
ConvNeXt	6
ViT	6.333333
ResNet	7
Random Forest	7.5
SVM	8.666667
SNIC + SVM	8.916667
SNIC + GCN	11.16667

Accuracy estimates were also estimated by calculating standard errors that were based on the 0.025 percentile of the bootstrap distribution that they used in deriving 95 percent confidence intervals. DaViT, CNN got the accuracy estimate of 99.83% [99.45%, 99.98%] compared to 98.63% [97.62%, 99.32%] in the case of the Random Forest model (Table 10). The difference between the patch-based deep learning approaches and the conventional methods is statistically robust due to the lack of overlap in their confidence scores, which in turn further supports statistically robust differences in performance (Tables 13 and 14).

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

Table 13 *P-values for comparing the error rate of classifiers using pairwise McNemar's test. The p-value is shown in each cell to test if the two classifiers have significantly different error rates on matched samples. A value close to 0 means that the error of classification is very different. With the exception of a few instances where the differences between SNIC + SVM vs. ResNet ($p = 1.24\times10^{-13}$) and U-Net vs. ViT ($p = 6.36\times10^{-12}$) are not significant, almost all classifier combinations exhibit significant differences ($p < 0.001$)*

Model	SGCN	SRF	SSVM	CNN	Conv	DaViT	EffNet	ResNet	U-Net	ViT	RF	SVM
SGCN	1	0	0	0	0	0	0	0	0	0	0	0
SRF	0	1	0	0	0	0	0	0	0	0	0	0
SSVM	0	0	1	0	0	0	0.585	1.24E-13	0	0	0	0
CNN	0	0	0	1	0	0	0	0	0	0	0	0
Conv	0	0	0	0	1	0	0	0	0	0	0	0
DaViT	0	0	0	0	0	1	0	0	0	0	0	0
EffNet	0	0	0.585	0	0	0	1	0	0	0	0	0
ResNet	0	0	1.24E-13	0	0	0	0	1	0	0	0	0
U-Net	0	0	0	0	0	0	0	0	1	6.36E-12	0	0
ViT	0	0	0	0	0	0	0	0	6.36E-12	1	0	0
RF	0	0	0	0	0	0	0	0	0	0	1	0
SVM	0	0	0	0	0	0	0	0	0	0	0	1

Table 14 *In pairs statistical significance analysis using Bonferroni and Holm post-hoc corrections*

Model 1	Model 2	p-value	Bonferroni Threshold	Bonferroni Significant	Holm Threshold	Holm Significant
SNIC + GCN	SNIC + RF	0	0.000758	TRUE	0.000758	TRUE
SNIC + GCN	SNIC + SVM	0	0.000758	TRUE	0.000769	TRUE
SNIC + GCN	CNN	0	0.000758	TRUE	0.000781	TRUE
SNIC + GCN	ConvNeXt	0	0.000758	TRUE	0.000794	TRUE
SNIC + GCN	DaViT	0	0.000758	TRUE	0.000806	TRUE
SNIC + GCN	EfficientNet	0	0.000758	TRUE	0.000820	TRUE
SNIC + GCN	ResNet	0	0.000758	TRUE	0.000833	TRUE
SNIC + GCN	U-Net	0	0.000758	TRUE	0.000847	TRUE
SNIC + GCN	ViT	0	0.000758	TRUE	0.000862	TRUE
SNIC + GCN	Random Forest	0	0.000758	TRUE	0.000877	TRUE
SNIC + GCN	SVM	0	0.000758	TRUE	0.000893	TRUE
SNIC + RF	SNIC + SVM	0	0.000758	TRUE	0.000909	TRUE
SNIC + RF	CNN	0	0.000758	TRUE	0.000926	TRUE
SNIC + RF	ConvNeXt	0	0.000758	TRUE	0.000943	TRUE
SNIC + RF	DaViT	0	0.000758	TRUE	0.000962	TRUE
SNIC + RF	EfficientNet	0	0.000758	TRUE	0.000980	TRUE
SNIC + RF	ResNet	0	0.000758	TRUE	0.001000	TRUE
SNIC + RF	U-Net	0	0.000758	TRUE	0.001020	TRUE
SNIC + RF	ViT	0	0.000758	TRUE	0.001042	TRUE
SNIC + RF	Random Forest	0	0.000758	TRUE	0.001064	TRUE
SNIC + RF	SVM	0	0.000758	TRUE	0.001087	TRUE
SNIC + SVM	ResNet	1.24E-13	0.000758	TRUE	0.016667	TRUE
U-Net	ViT	6.36E-12	0.000758	TRUE	0.025000	TRUE
SNIC + SVM	EfficientNet	0.585229	0.000758	FALSE	0.050000	FALSE

5.4 Computational Efficiency Comparison

Another aspect to be considered with the practical deployment of classification models is the complexity of the classification model, which may be too large to be effectively deployed in low-resource environments or where a near-real-time response is required. The computational characteristics of all the models under consideration are summarized (Table 15).

The deep learning models had many more parameters than traditional machine learning methods. DaViT and EfficientNet have about 1.78 million and 2.06 million trainable parameters, respectively, due to their compound scaling network structure. The ResNet and U-Net had 2.17 million and 1.08 million parameters, respectively. Conversely, one needed 124k parameters, with similar performance to more sophisticated CNNs, and it has been demonstrated that well-designed small networks can be used to reach the same level of performance as large ones in this classification problem.

The model sizes were also proportional to the number of model parameters since EfficientNet took 2.06 MB to be stored, whereas ResNet took 2.17 MB. The CNN miniature has overall required 0.12MB and has lowered storage demand by 17x the bigger designs without accuracy deterioration. In the case of the compressed joblib representations of classic machine learning models, storage was space comparable (less than 0.5 MB), but comparison is not fair because of the processing of serialization techniques.

Table 15 *Computational properties of all the considered models.*

Model	Category	Type	Parameters	Size (MB)	Accuracy (%)
CNN	Patch-Based	DL	7.7K	0.12	99.83
ViT	Patch-Based	DL	92.6K	1.17	99.49
ResNet	Patch-Based	DL	175.1K	2.17	99.14
ConvNeXt	Patch-Based	DL	24.7K	0.37	99.14
EfficientNet	Patch-Based	DL	156.5K	2.06	99.66
DaViT	Patch-Based	DL	141.7K	1.78	99.83
U-Net	Patch-Based	DL	89.2K	1.09	99.49
SNIC + RF	OBIA	OBIA	100	0.26	98.67
SNIC + SVM	OBIA	OBIA	31	0.01	92.00
SNIC + GCN	Graph	GCN	0	12.88	78.16

Although the classification accuracy (the test set) of the model was also improving with the addition of more parameters, there existed diminishing returns that came with the increasing complexity of the model. The trade-off between accuracy and number of parameters led to the compact CNN to achieve a comparable level of DNLM accuracy (99.83%), as with only 124K parameters, which is 14 times smaller than DaViT (1.78M). This would imply that the specific classification task and dataset that is typical of the present study would be well represented by an intermediate level of model complexity to represent enough of the overall spectral-spatial patterns; additional parameters do not lead to improved accuracy.

In terms of training time (random forest and SVM classifiers were more likely to converge within seconds, whereas deep learning models took minutes or hours to train), they had a computational advantage. This variation in inference time is, however, minimized since trained deep learning models will generate predictions quickly for large images through a batch prediction using a GPU-accelerated predictor.

5.5 Visual Analysis of Classification Maps

Our classification maps have different spatial patterns that can be taken as effective measures of the various classification strategies adopted. (Tables 16,17) summarizes classifications by the type of methodology used.

The pixel-wise classification maps (Figure 7) obtained with Random Forest and SVM provide more precise spatial detail, including fine-scale features, and clearly delineate the boundaries between classes. On closer examination, however, there is some salt-and-pepper noise in mixed zones between land covers in the area of spectral signature overlap. Random Forest also offered a smoother classification than SVM, taking into consideration the ensemble average factor of the given method.

Findings: Classification maps achieved by patch-based deep learning algorithms (Figure 8) were relatively coherent and yielded clear boundaries between the classes. The patch-based method is also effective in the use of spatial context to minimize random misclassifications and preserve, at the same time, real land cover boundaries. The cleanest visual appearance in DaViT and CNN maps was characterized by homogeneous areas and sharp transitions. All models except ResNet and ConvNeXt were similar in terms of STD, which was marginally more unpredictable in homogeneous regions.

The OBIA classified maps (Figure 9) show the visual looks that are characteristic of object-based classification, and which are only super-classified. Homogeneous areas with the same class of SNIC-RF map have correct labels, unlike the SNIC-SVM result, which spurred this bias (particularly Road) being diluted, and the road network depicted as blank segments.

The GCN classification map (Figure 9) shows the superpixel-limited quality of the graph-based classification at the expense of much less accuracy. The method is found to have too few features and to misclassify areas on a

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

regional basis, compared to a more accurate methodology. Rangeland and vegetation between these spectrally similar categories can be considerable. The graph structure also presents an error propagation scheme among adjacent superpixels, which is potentially better with a superior criterion on graph construction or increasing training samples.

The spatial error analysis determined that errors in the classifications were present. The misplaced concentration of errors in poorer-performing models was in spectrally complex regions, such as class boundaries and mixed land cover. The patch-based models had more homogeneous distributions of errors, which is a characteristic of an error that is a true classification problem and not a systematic failure of the model.

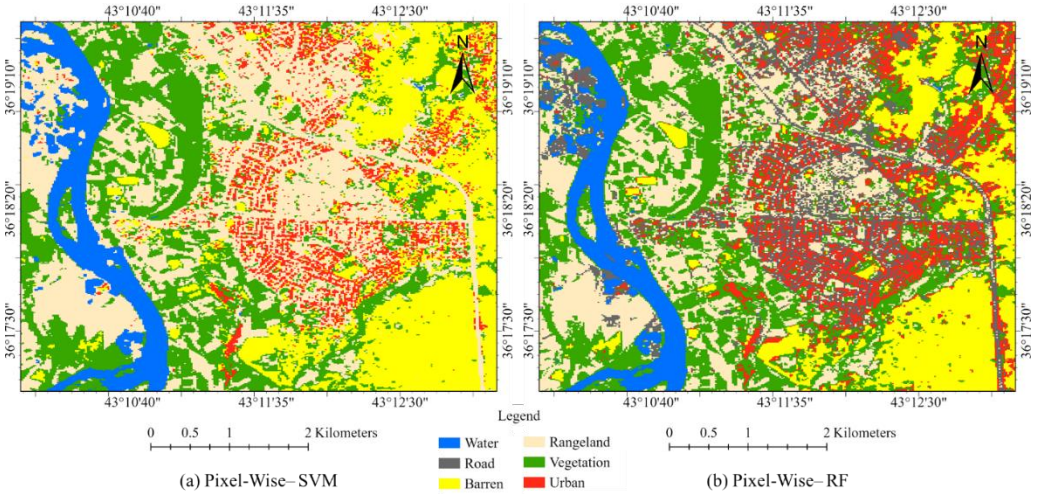
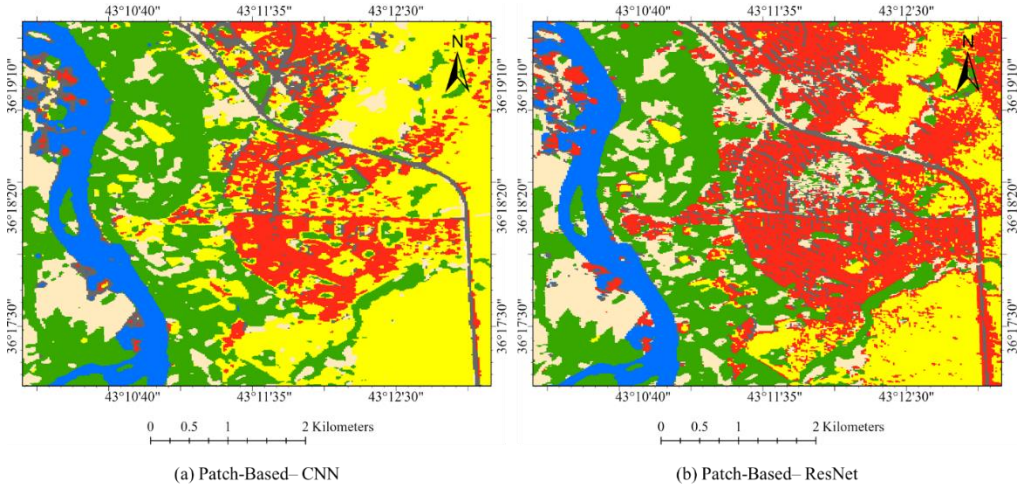


Fig. 7 *Pixel-wise classification techniques are used to create land cover classification maps. Reference map (a); Random Forest (OA = 98.63%), and SVM (OA = 97.94%). Because of the training data's spectral uniformity, both approaches provide smooth classification with little salt-and-pepper noise*



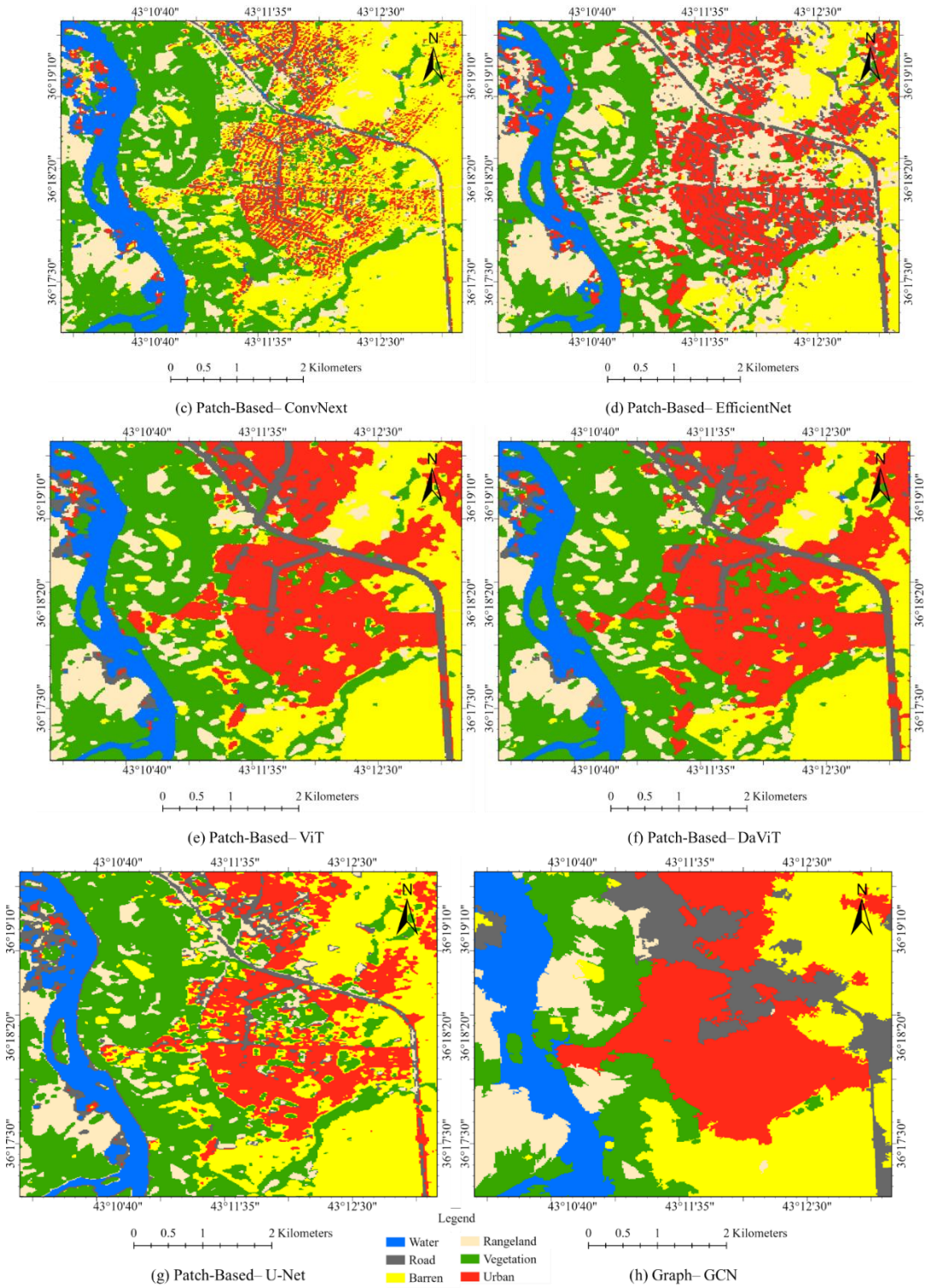


Fig. 8 *The pixel-wise classification methods produce land cover classification maps. Reference map (a); Random Forest (OA = 98.63%)(b); and SVM (OA = 97.94%)(c). The spectral consistency of the training data results in the smooth classification without salt and pepper noise in both methods*

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

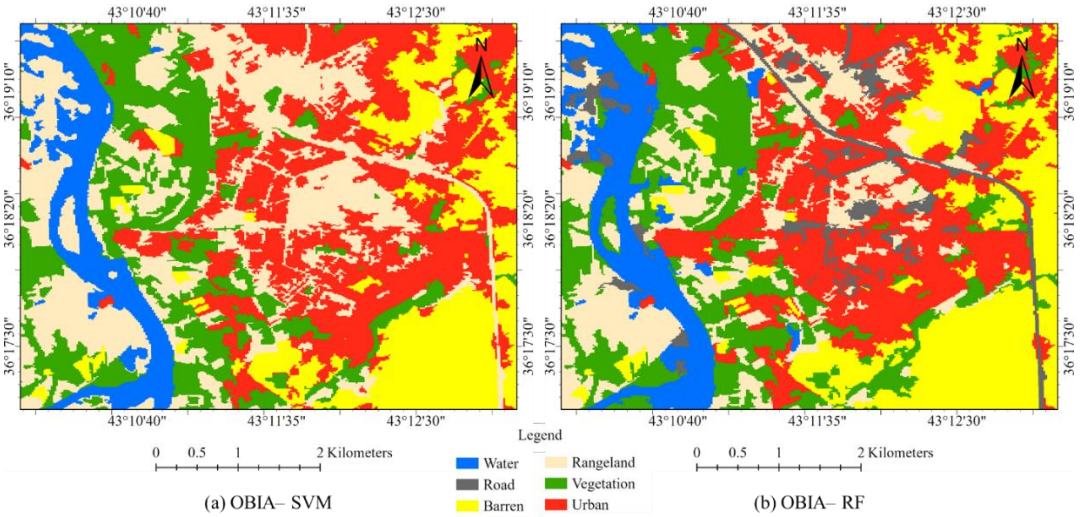


Fig. 9 *Land cover categorization maps were created using patch-based deep learning techniques combined with the graph-based GCN method. CNN (OA = 99.83%), ViT (OA = 99.49%), ResNet (OA = 99.14%), ConvNeXt (OA = 99.14%), EfficientNet (OA = 99.66%), DaViT (OA = 99.83%), U-Net (OA = 99.49%), and SNIC + GCN (OA = 78.16%)*

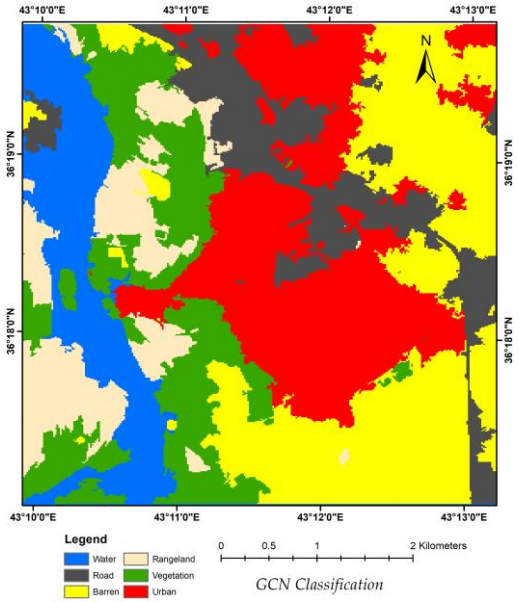


Fig. 9 *Graph Convolutional Network (GCN) graph-based classification map*

Table 16 *The class area distribution of each model (the total number of pixels in the image)*

Category	Model	Water	Road	Barren	Rangeland	Vegetation	Urban
Graph	SNIC + GCN	14456	14328	30647	12321	18718	27898
OBIA	SNIC + RF	12186	6670	23828	21269	22393	32022
OBIA	SNIC + SVM	10043	0	20383	33427	21179	33336
Patch-Based	CNN	9953	8324	39374	12595	31917	16205
Patch-Based	ConvNeXt	10883	5826	42589	14003	33235	11832
Patch-Based	DaViT	9912	7389	22749	11521	35438	31359
Patch-Based	EfficientNet	8282	10708	20907	31379	25382	21710
Patch-Based	ResNet	9530	9047	18522	18002	29404	33863
Patch-Based	U-Net	8523	8993	32233	10444	37263	20912
Patch-Based	ViT	8897	7198	27175	11533	33099	30466
Pixel-Wise	Random Forest	9235	16187	18022	28542	30828	15554
Pixel-Wise	SVM	9241	189	27041	42989	29353	9555

Table 17 *Each model's class area distribution (as a proportion of the overall picture area)*

Category	Model	Water	Road	Barren	Rangeland	Vegetation	Urban
Graph	SNIC + GCN	12.21	12.1	25.89	10.41	15.81	23.57
OBIA	SNIC + RF	10.3	5.63	20.13	17.97	18.92	27.05
OBIA	SNIC + SVM	8.48	0	17.22	28.24	17.89	28.16
Patch-Based	CNN	8.41	7.03	33.26	10.64	26.96	13.69
Patch-Based	ConvNeXt	9.19	4.92	35.98	11.83	28.08	10
Patch-Based	DaViT	8.37	6.24	19.22	9.73	29.94	26.49
Patch-Based	EfficientNet	7	9.05	17.66	26.51	21.44	18.34
Patch-Based	ResNet	8.05	7.64	15.65	15.21	24.84	28.61
Patch-Based	U-Net	7.2	7.6	27.23	8.82	31.48	17.67
Patch-Based	ViT	7.52	6.08	22.96	9.74	27.96	25.74
Pixel-Wise	Random Forest	7.8	13.68	15.23	24.11	26.04	13.14
Pixel-Wise	SVM	7.81	0.16	22.84	36.32	24.8	8.07

5.6 Taylor Diagram Analysis

The Taylor diagram (Figure 10) brings together information about the performance of the model by simultaneously presenting three complementary statistics, including the standard deviation, the correlation coefficient and the centered RMSE (Table 18). The methods of summarization are multidimensional, providing a graphical comparison of what is classified and what is not for a classification output and the reference data.

The target value (TARGET) is plotted at the standard deviation of a reference classification (here ~1.65), on the x axis of the Taylor diagram. Perfect agreement models were indicated on this reference point and the disagreement models would be revealed as various forms of disagreement.

Patches rendered by the deep learning patch models are heavily concentrated around the reference point, and the correlations with values above 0.99 and standard deviations close to the reference one are very good. Such groupings confirm the validity of magnitude (i.e., appropriate trial-to-trial variance in the assignment of classes) and pattern selection (i.e., strong association with their reference spatial patterns) of these two cluster algorithms. DaViT, and CNN, and EfficientNet are hard to distinguish at the reference point, hence visually justifying their good performance.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

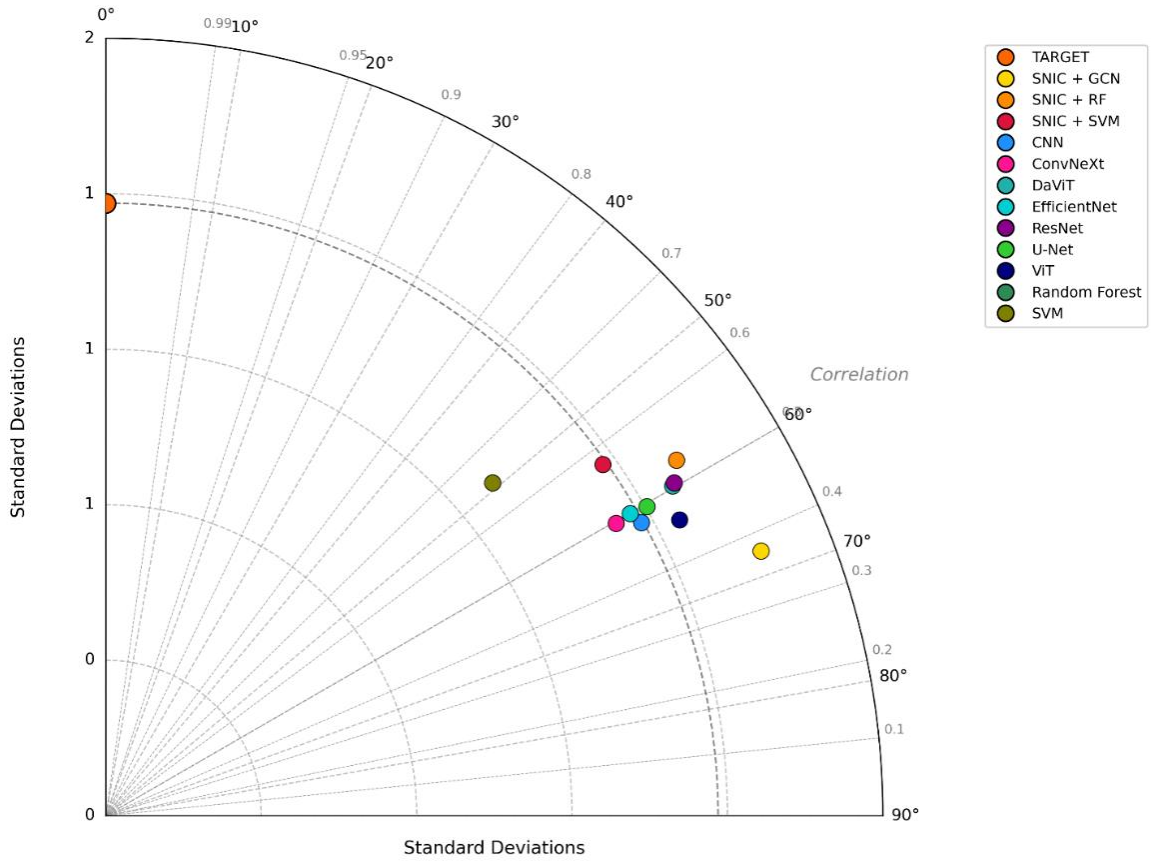


Fig. 10 *All categorization models are compared using a Taylor diagram. Standard deviation is indicated by the radial distance, correlation with the reference is indicated by the angular position, and the reference standard deviation is represented by the dashed arc. Better agreement is seen in models that are nearer the TARGET marker*

The pixel-wise and OBIA methods have slightly different results, which are shifted off of the reference, with correlation coefficients of 0.96-0.98. Angular deviation of reference demonstrates a little bit of lessened pattern agreement, but the radial locations confirm the anticipated dissimilarity between classifications. The SNIC-RF position is nearly comparable to Random Forest, which is associated with their similar final accuracies.

The GCN classifier seems to be quite unconfined in the reference cluster with a distinct standard deviation and a low correlation (around 0.78). The greater the displacement of reference material comes into graphic expression, the greater the difference in performance as demonstrated by the numerical accuracy test. The correlation axis point means that GCN can generalize the general spatial patterns of impervious surfaces with lower accuracy than other, more efficient approaches.

A Taylor diagram is a summary of major findings of the multi-model comparison in a publication-quality format, summarizing the main findings, as opposed to duplicating the detailed numerical summary provided in the preceding sections. The visual hierarchy of models based on taxonomies of classification indicates that the approach to the methodology influences the classification performance, and that analogous intra-paradigmatic dispersion can be explained by certain architectural choices.

Table 18 *Taylor diagram statistics for comparing categorization models. Standard deviation, centred RMSE, and Pearson correlation with reference are examples of metrics*

Model	Std	Correlation	RMSE
TARGET	1.4727	1	0
SNIC + GCN	1.305	0.7816	0.963
SNIC + RF	1.462	0.9867	0.240
SNIC + SVM	1.412	0.9200	0.582
CNN	1.471	0.9983	0.085
ConvNeXt	1.465	0.9914	0.193

Model	Std	Correlation	RMSE
DaViT	1.471	0.9983	0.085
EfficientNet	1.470	0.9966	0.121
ResNet	1.465	0.9914	0.193
U-Net	1.469	0.9949	0.149
ViT	1.469	0.9949	0.149
Random Forest	1.462	0.9863	0.244
SVM	1.458	0.9794	0.299

6. Discussion

This work involved a comprehensive comparative analysis of twelve classification approaches in the sphere of remote sensing image classification, and tested them in accordance with four paradigms: pixel-wise method, OBIA method, patch-based deep learning method, and graph-based methods. The findings of the experiment introduced hierarchies of performance, which contain inherent logic, patterns of activity in classes, and seriously restrictive methodological considerations, which are of theoretical importance with regard to understanding and practical implications of the image-based land-cover classification by means of deep learning.

While this study provides a comprehensive comparison of twelve land cover classification models, several limitations should be noted. First, the evaluation was conducted using a single SPOT-5 scene from Mosul, Iraq. Future work should test the generalizability of these findings across different seasons, diverse geographic regions, and other VHR sensors. Second, the spatial cross-validation approach based on polygon centroids, while preventing intra-polygon information leakage, is still subject to regional spatial autocorrelation. Future research should implement geographic block cross-validation to assess model performance under complete spatial independence. Third, class imbalance (e.g., Road having only 21 pixels versus Water having 403 pixels) remains a challenge; future studies should explore advanced minority-oversampling techniques or class-weighted loss functions in greater depth. Finally, the GCN's performance was evaluated under a specific configuration; optimizing graph topology through learnable edge weights or multi-scale superpixel graphs stands as a promising future direction.

This was succeeded by EfficientNet, which had an immeasurably close conspiracy accuracy at the levels of 99.66 percent (k = 0.9954), and U-Net and ViT had also reinforced to the same level of accuracy of 99.49 percent (k = 0.9931). Other comparable deep learning designs like ResNet, ConvNeXt also recorded passive great performance (99.14% 695 accuracy with kappa ≈ 0.9885-0.9886).

Patch-based deep learning solutions have dominated due to several major reasons. These architectures are based on spatial context, as the architecture can process a local neighborhood (5x5 patch in this case), and therefore it can learn to (represent) textural patterns, edge attributes, and contextual relationships that cannot be represented by individual pixels. Second, the hierarchical nature of deep learning with respect to feature learning enables it to auto-learn features of patterns based from high-level semantic features to low-level spectral features. DaVit is (amazing) outcome, which consists of two attentions, attests to the fact that the hybrid architecture of combining the reasoning capability of local and global feature dependencies is highly suitable in remote sensing classification. Similarly, the simple CNN architecture also performs well, showing that the architecture with just 7 layers is capable of doing very well when there is a clear spectral-spatial separability in the classification problem.

The top patch-based deep learning models outperformed the conventional machine learning methods of pixel-wise SVM and Random Forest in terms of accuracy, but they still provided very competitive results.

Pixel-wise SVM and Random Forest, two conventional machine learning techniques, produced competitive outcomes. SVM attained 97.94 percent accuracy (k = 0.9724) and pixel-wise Random Forest 98.63 percent precision (k = 0.9817). The implications of these findings are that classical machine learning methods can continue to be highly competitive in terms of land cover classification when there are computational constraints and/or when training data is limited. Specifically, the Random Forest classifier is extremely efficient: it matches the performance of the most effective deep learning algorithms but requires a significantly smaller amount of computing resources and does not require being implemented with the help of a graphics card (GPU). Significant differences were indicated by the McNemar test ($p < 0.001$) between most pairs of models, especially in cases where observed differences in performance were not due to chance.

The ways of the OBIA were highly heterogeneous and should be discussed in detail. The performance of SNIC + Random Forest was 98.67% accurate (k = 0.9835), which was almost the same performance as the pixelwise-based classifier, degrading OBIA + SVM by a significant margin. Conversely, the SNIC + SVM also had poorer results in comparison to the SNIC + SVM Random Forest counter-attack detection, with an accuracy of 92.0% (k = 0.9003), which is a very large difference of 6.67 percent. This disjunction implies that the classifier chosen in an OBIA

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 184-215

framework has a high contribution and weaker influence on ensemble-based classifiers, such as the Random Forest, in the superpixel segmentation feature combination.

With an accuracy of 78.16% (k = 0.7309), the classification by graph approach combining Graph Convolutional Networks (SNIC + GCN) obtained the lowest scores of all the investigated approaches. Although GCNs are successful on most computer vision tasks, they can fail because of a number of reasons. The SNIC parameter for superpixel size was 8 pixels, which may have been an inefficient choice for recapturing the spatial correlations needed to perform graph convolution effectively. Similarly, it might not be sufficient to use a shallow 3-layer GCN with hidden dimensions [32, 61] to carry out this task. The constructed graph based on the adjacency of superpixels might also fail to capture the spatial relationship to classify land use, particularly of the categories with fragmented and linear shapes.

Further methods of performance comparison are disclosed in Cohen's d effect-size analysis. Cohen's d = 2.519 was obtained when comparing DaViT and SNIC + GCN., a highly significant effect size, indicating that the difference in performance is not only statistically significant (i.e., the possibility exists) but also a tangible phenomenon. Similarly, the effect sizes for CNN vs SNIC + GCN (d = 2.410) and SNIC + RF vs SNIC + GCN (d = 1.994) are vast. There were medium-small effect sizes between models (between 0.3 and 0.8), which indicated that some differences existed, but the practical separation between deep learning architectures' best performances might be small enough to name a few applications. The model analysis is too complex, which shows that there is an inefficient correlation between the depth of the network and its performance (Table 19).

Table 19 *An overview of the top models based on several accuracy criteria*

Rank	Category	Model	Accuracy	Kappa	F1_Weighted
1	Patch-Based	DaViT	99.83	0.9977	0.998287
2	Patch-Based	CNN	99.83	0.9977	0.998323
3	Patch-Based	EfficientNet	99.66	0.9954	0.996572
4	Patch-Based	U-Net	99.49	0.9931	0.994735
5	Patch-Based	ViT	99.49	0.9931	0.994863
6	Patch-Based	ConvNeXt	99.14	0.9885	0.990736
7	Patch-Based	ResNet	99.14	0.9886	0.991291
8	OBIA	SNIC + RF	98.67	0.9835	0.985961
9	Pixel-Wise	Random Forest	98.63	0.9817	0.985578
10	Pixel-Wise	SVM	97.94	0.9724	0.971245
11	OBIA	SNIC + SVM	92	0.9003	0.900337
12	Graph	SNIC + GCN	78.16	0.7309	0.774143

EfficientNet with 63 layers achieved 99.66%, and DaViT with 25 layers, the maximum precision, which was the same as the 7-layer CNN. This result suggests that model depth may not have as much of an effect on performance as some architectural design decisions, such as attention strategies, normalization techniques, or skip connections. Because shallower models with comparable accuracy are superior in terms of inference time, memory footprint, and implementation in resource-constrained scenarios, the conclusion has practical implications.

The result of such an analysis is the substantial heterogeneity in land cover classifications across methodologies. The most separable turned out to be Water, Barren, and Urban classes. As both wavelengths had a very high absorption in water, the water (which was smoother than the tissue) achieved close to perfect classification (99.5-100%) across all models. This was also true of patch methods, which performed well in Barren and Urban classes, where a DaViT model (same as a DaViT model) was able to achieve a perfect F1 score of 1.0 by applying cutting-edge attention-based architectures.

Vegetation and Rangeland, on the contrary, were problematic. Patch-based models gave good results, but the GCN model could not perform well (F1 = 0.56-0.70) on these classes due to spectral confusion and large internal diversity. The hardest category for the SVM-based methods to follow was roads, with 0% accuracy. This weakness is due to intense class imbalance, narrow shapes that are linear, and spectrally mixed with other impervious surfaces. However, DaViT had 100 percent accuracy, and the fact that it achieved this accuracy shows that its dual attention system can be used to detect complex linear features.

Spatial error analysis shows that from phenological and management standpoints, Agriculture has the highest error sensitivities (100%). More to the point, the study also illuminates model-specific biases: SGD/SVM classifiers will have a particular bias towards removing minority-class regions, e.g., roads in this case, whereas GCNs tend to produce more balanced distributions at the cost of overall accuracy. Patch-based methods do not scale with urban

mixed pixels and will simply combine many land covers into a few classes. Lastly, the effectiveness of deep learning, and specifically DaViT, underscores the unconditional significance (of) the spatial context/hierarchical feature learning of spectrally concussive/geometrically convoluted land cover classes.

7. Conclusions

This work is used in twelve land use/land cover (LULC) classification techniques, which fall into using SPOT-5 VHR photos acquired over the Mosul region of Iraq. Four paradigms were used: pixel-wise, object-based, patch-based deep learning, and graph-based techniques. The results lead to a certain performance hierarchy in which deep-learning approaches based on patches prevail. Specifically, DaViT and a seemingly less complicated 7-layer CNN achieved a peak accuracy of 99.83% and a 0.9977 Kappa coefficient. This performance similarity between the 7,734-parameter CNN and the gigantically large 141,710-parameter DaViT foreshadows that intolerable levels of architectural complexity are not necessary to achieve state-of-the-art classification results on VHR datasets. Classical machine learning remains difficult to rival: The pixel-wise Random Forest took a score of 98.63. This proves that it can be applied in applications that are not resource-intensive. The choice of classifier had a large impact on OBIA results, with SNIC + Random Forest significantly better than SNIC + SVM. Conversely, the worst accuracy (78.16%) was the GCN approach, which may not have been well hypertextured at some level, as opposed to our method being inherently low in accuracy. The most difficult category was determined to be the class-specific analysis, where Roads was the hardest due to class imbalance and the use of linear shapes, whereas the most preferred geometry was Water across all approaches. The class with the most inter-model variability was the Rangeland class, indicating that accounting for spatial context in turbulent spectral areas is necessary. This research has made contributions based on the earliest comparison and analysis of architectures such as ConvNeXt and EfficientNet for Middle Eastern urban environments. MCU noted that future research directions should include multi-temporal approaches, GCN optimization, and specialized techniques to address the problems of minority-class imbalance, such as in road networks, which are a significant factor in the continuation of successful operational classification systems.

Acknowledgements

The authors have no financial support or academic advice for our research.

Conflict of interest

Regarding the publishing of this work, the authors state that they have no conflicts of interest.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception, design, and data collection:*** *Ahmed Qusay Subhe, Muntadher Aidi Shareef;* ***assisted in remote sensing data processing and critically revised the manuscript for important intellectual content:*** *Sadra Karimzadeh, Muntadher Aidi Shareef. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

[2] Azeez, S. M., Shareef, M. A., & Omer, F. M. (2025). Assessing the Eco-Environmental Quality of Kirkuk City and Taza District Based on Pressure-State-Response Framework for Two Seasons Using Remote Sensing and GIS. *Tikrit Journal of Engineering Sciences*, *32*(3), 1. https://doi.org/10.25130/tjes.32.3.13

[3] Azeez, S. M., Shareef, M. A., & Omer, F. M. (2024). Eco-Environmental Quality Assessment of Two Seasons Based on Pressure-State-Response Framework by Using Remote Sensing and GIS Techniques. *The Iraqi Geological Journal*, *57*(2), 233–252. https://doi.org/10.46717/igj.57.2D.19ms-2024-10-29

[4] Núñez, J. M., Medina, S., & Ávila, G. (2019). High-Resolution Satellite Imagery. In Satellite Information Classification and Interpretation (pp. 69–85). Google Books.

[5] Shareef, M. A., Toumi, A., & Khenchaf, A. (2015). Estimation and characterization of physical and inorganic chemical indicators of water quality by using SAR images. *Https://Doi.Org/10.1117/12.2194503*, *9642*, 140–150. https://doi.org/10.1117/12.2194503

[6] Hassan, N. D., Noori, A. M., Hasan, S. F., Shareef, M. A., & Ajaj, Q. M. (2019). Cadastral Mapping Accuracy Assessment Using Various Surveying Techniques and High-Resolution Satellites Images. *2nd International Conference on Electrical, Communication, Computer, Power and Control Engineering, ICECCPCE 2019*, 182–187. https://doi.org/10.1109/ICECCPCE46549.2019.203770

[7] Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

[8] Maulik, U., & Chakraborty, D. (2017). Remote Sensing Image Classification: A survey of support-vector-machine-based advanced techniques. *IEEE Geoscience and Remote Sensing Magazine*, *5*(1), 33–52. https://doi.org/10.1109/MGRS.2016.2641240

[9] Tzotsos, A., & Argialas, D. (2008). Support Vector Machine Classification for Object-Based Image Analysis. *Lecture Notes in Geoinformation and Cartography*, *0*(9783540770572), 663–677. https://doi.org/10.1007/978-3-540-77058-9_36

[10] Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, *118*, 259–272. https://doi.org/10.1016/j.rse.2011.11.020

[11] Sharma, A., Liu, X., Yang, X., & Shi, D. (2017). A patch-based convolutional neural network for remote sensing image classification. *Neural Networks*, *95*, 19–28. https://doi.org/10.1016/j.neunet.2017.07.017

[12] Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(7), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

[13] Ismat Abdulrahman, M., Aidi Shareef, M., & Abbas Al-Attar, A. (2024). Deep Learning (CNN) for Detecting Road Infrastructure in Old Mosul City Using High-Resolution Aerial Imagery. *Baghdad Science Journal*, *24*(3), 1049–1064. https://doi.org/10.21123/bsj.2024.9449

[14] Noori, A. M., Ziboon, A. R. T., & AL-Hameedawi, A. N. (2025). Deep-Learning Integration of CNN–Transformer and U-Net for Bi-Temporal SAR Flash-Flood Detection. *Applied Sciences 2025, Vol. 15, Page 7770*, *15*(14), 7770. https://doi.org/10.3390/app15147770

[15] Zhang, T., Su, J., Xu, Z., Luo, Y., & Li, J. (2021). Sentinel-2 Satellite Imagery for Urban Land Cover Classification by Optimized Random Forest Classifier. *Applied Sciences 2021, Vol. 11, Page 543*, *11*(2), 543. https://doi.org/10.3390/app11020543

[16] Khan, N., Chaudhuri, U., Banerjee, B., & Chaudhuri, S. (2019). Graph convolutional network for multi-label VHR remote sensing scene recognition. *Neurocomputing*, *357*, 36–46. https://doi.org/10.1016/j.neucom.2019.05.024

[17] Xie, G., & Niculescu, S. (2021). Mapping and Monitoring of Land Cover/Land Use (LCLU) Changes in the Crozon Peninsula (Brittany, France) from 2007 to 2018 by Machine Learning Algorithms (Support Vector Machine, Random Forest, and Convolutional Neural Network) and by Post-classification …. *Remote Sensing 2021, Vol. 13, Page 3899*, *13*(19), 3899. https://doi.org/10.3390/rs13193899

[18] Yang, C., Everitt, J. H., & Murden, D. (2011). Evaluating high resolution SPOT 5 satellite imagery for crop identification. *Computers and Electronics in Agriculture*, *75*(2), 347–354. https://doi.org/10.1016/j.compag.2010.12.012

[19] Rengma, N. S., & Yadav, M. (2024). Generation and classification of patch-based land use and land cover dataset in diverse Indian landscapes: a comparative study of machine learning and deep learning models. *Environmental Monitoring and Assessment 2024 196:6*, *196*(6), 568-. https://doi.org/10.1007/s10661-024-12719-7

[20] Zhang, H., Li, Q., Liu, J., Du, X., Dong, T., McNairn, H., Champagne, C., Liu, M., & Shang, J. (2018). Object-based crop classification using multi-temporal SPOT-5 imagery and textural features with a Random Forest classifier. *Geocarto International*, *33*(10), 1017–1035. https://doi.org/10.1080/10106049.2017.1333533

[21] Sariturk, B., & Seker, D. Z. (2022). A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images. *Sensors 2022, Vol. 22, Page 7624*, *22*(19), 7624. https://doi.org/10.3390/s22197624

[22] Rad, R. (2024). *Vision Transformer for Multispectral Satellite Imagery: Advancing Landcover Classification* (pp. 8176–8183).

[23] Lee, S. H., Lee, S., & Song, B. C. (2021). Vision Transformer for Small-Size Datasets. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3718-3727).

[24] Vinod, P. V., Behera, M. D., Jaya Prakash, A., Hebbar, R., & Srivastav, S. K. (2024). A novel multitask transformer deep learning architecture for joint classification and segmentation of horticulture plantations using very High-Resolution satellite imagery. *Computers and Electronics in Agriculture*, *227*, 109540. https://doi.org/10.1016/j.compag.2024.109540

[25] Xu, S., Zhao, Q., Yin, K., Zhang, F., Liu, D., & Yang, G. (2019). Combining random forest and support vector machines for object-based rural-land-cover classification using high spatial resolution imagery. *Https://Doi.Org/10.1117/1.JRS.13.014521*, *13*(1), 014521. https://doi.org/10.1117/1.jrs.13.014521

[26] Han, R., Liu, P., Wang, G., Zhang, H., & Wu, X. (2020). Advantage of Combining OBIA and Classifier Ensemble Method for Very High-Resolution Satellite Imagery Classification. *Journal of Sensors*, *2020*(1), 8855509. https://doi.org/10.1155/2020/8855509

[27] Kavzoglu, T., & Tonbul, H. (2018). An experimental comparison of multi-resolution segmentation, slic and k-means clustering for object-based classification of vhr imagery. *International Journal of Remote Sensing*, *39*(18), 6020–6036. https://doi.org/10.1080/01431161.2018.1506592

[28] Fu, Z., Sun, Y., Fan, L., & Han, Y. (2018). Multiscale and Multifeature Segmentation of High-Spatial Resolution Remote Sensing Images Using Superpixels with Mutual Optimal Strategy. *Remote Sensing 2018, Vol. 10, Page 1289*, *10*(8), 1289. https://doi.org/10.3390/rs10081289

[29] Zhang, X., Tan, X., Chen, G., Zhu, K., Liao, P., & Wang, T. (2022). Object-Based Classification Framework of Remote Sensing Images with Graph Convolutional Networks. *IEEE Geoscience and Remote Sensing Letters*, *19*. https://doi.org/10.1109/LGRS.2021.3072627

[30] Diao, Q., Dai, Y., Zhang, C., Wu, Y., Feng, X., & Pan, F. (2022). Superpixel-Based Attention Graph Neural Network for Semantic Segmentation in Aerial Images. *Remote Sensing 2022, Vol. 14, Page 305*, *14*(2), 305. https://doi.org/10.3390/rs14020305

[31] Shi, W., & Sui, H. (2022). An effective superpixel-based graph convolutional network for small waterbody extraction from remotely sensed imagery. *International Journal of Applied Earth Observation and Geoinformation*, *109*(4), 102777. https://doi.org/10.1016/j.jag.2022.102777

[32] Kavran, D., Mongus, D., Žalik, B., & Lukač, N. (2023). Graph Neural Network-Based Method of Spatiotemporal Land Cover Mapping Using Satellite Imagery. *Sensors 2023, Vol. 23, Page 6648*, *23*(14), 6648. https://doi.org/10.3390/s23146648

[33] Liu, Z. Q., Tang, P., Zhang, W., & Zhang, Z. (2022). CNN-Enhanced Heterogeneous Graph Convolutional Network: Inferring Land Use from Land Cover with a Case Study of Park Segmentation. *Remote Sensing 2022, Vol. 14, Page 5027*, *14*(19), 5027. https://doi.org/10.3390/rs14195027

[34] Xu, K., Huang, H., Deng, P., & Li, Y. (2022). Deep Feature Aggregation Framework Driven by Graph Convolutional Network for Scene Classification in Remote Sensing. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(10), 5751–5765. https://doi.org/10.1109/TNNLS.2021.3071369

[35] Ma, B., Wu, F., Hu, T., Fathollahi, L., Sui, X., Liu, Y., & Gantumur, B. (2024). Label-Driven Graph Convolutional Network for Multilabel Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 2245–2255. https://doi.org/10.1109/JSTARS.2023.3344106

[36] Ali, U., Esau, T. J., Farooque, A. A., Zaman, Q. U., Abbas, F., & Bilodeau, M. F. (2022). Limiting the Collection of Ground Truth Data for Land Use and Land Cover Maps with Machine Learning Algorithms. *ISPRS International Journal of Geo-Information 2022, Vol. 11, Page 333*, *11*(6), 333. https://doi.org/10.3390/ijgi11060333

[37] Fisher, A., Day, M., Gill, T., Roff, A., Danaher, T., & Flood, N. (2016). Large-Area, High-Resolution Tree Cover Mapping with Multi-Temporal SPOT5 Imagery, New South Wales, Australia. *Remote Sensing 2016, Vol. 8, Page 515*, *8*(6), 515. https://doi.org/10.3390/rs8060515

[38] Pereira, M. R. (2024). Active Learning for Improved Landcover Classification (Master's thesis, University of Porto).

[39] Tabunshchik, V., Dzhambetova, P., Gorbunov, R., Gorbunova, T., Nikiforova, A., Drygval, P., Kerimov, I., & Kiseleva, M. (2025). Delineation and Morphometric Characterization of Small- and Medium-Sized Caspian Sea Basin River Catchments Using Remote Sensing and GISs. *Water 2025, Vol. 17, Page 679*, *17*(5), 679. https://doi.org/10.3390/w17050679

[40] Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*(1–3), 131–159. https://doi.org/10.1023/A:1012450327387

[41] Mena, J. B., & Malpica, J. A. (2005). An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery. *Pattern Recognition Letters*, *26*(9), 1201–1220. https://doi.org/10.1016/j.patrec.2004.11.005

[42] Fauvel, M., Chanussot, J., & Benediktsson, J. A. (2012). A spatial–spectral kernel-based approach for the classification of remote-sensing images. *Pattern Recognition*, *45*(1), 381–392. https://doi.org/10.1016/j.patcog.2011.03.035

[43] Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2019). Joint Deep Learning for land cover and land use classification. *Remote Sensing of Environment*, *221*, 173–187. https://doi.org/10.1016/j.rse.2018.11.014

[44] Jawak, S. D., Devliyal, P., & Luis, A. J. (2015). A Comprehensive Review on Pixel Oriented and Object Oriented Methods for Information Extraction from Remotely Sensed Satellite Images with a Special Emphasis on Cryospheric Applications. Advances in Remote Sensing, 4(3), 177–195. https://doi.org/10.4236/ars.2015.43015

[45] Wu, M., Wei, D., Zhang, L., & Zhao, Y. (2018). Hyperspectral Face Recognition with Patch-Based Low Rank Tensor Decomposition and PFFT Algorithm. *Symmetry 2018, Vol. 10, Page 714*, *10*(12), 714. https://doi.org/10.3390/sym10120714

[46] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881

[47] Hossain, M. A., & Islam, M. S. (2023). A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection. *Scientific Reports 2023 13:1*, *13*(1), 21207-. https://doi.org/10.1038/s41598-023-48230-1

[48] Lubana, E. S., Dick, R., & Tanaka, H. (2021). Beyond BatchNorm: Towards a Unified Understanding of Normalization in Deep Learning. *Advances in Neural Information Processing Systems*, *34*, 4778–4791.

[49] Pearce, T., Brintrup, A., & Zhu, J. (2021). *Understanding Softmax Confidence and Uncertainty*. http://arxiv.org/abs/2106.04972

[50] Saratchandran, H., Zheng, J., Ji, Y., Zhang, W., & Lucey, S. (n.d.). *Rethinking Softmax: Self-Attention with Polynomial Activations*.

[51] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition* (pp. 770–778).

[52] Tan, M., & Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* (pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

[53] Amangeldi, A., Taigonyrov, A., Jawad, M. H., & Mbonu, C. E. (2026). *CNN and ViT Efficiency Study on Tiny ImageNet and DermaMNIST Datasets*. http://arxiv.org/abs/2505.08259

[54] Ghosh, J., & Gupta, S. (2023). ADAM Optimizer and CATEGORICAL CROSSENTROPY Loss Function-Based CNN Method for Diagnosing Colorectal Cancer. *Proceedings of International Conference on Computational Intelligence and Sustainable Engineering Solution, CISES 2023*, 470–474. https://doi.org/10.1109/CISES58720.2023.10183491

[55] Li, C., Guo, B., Wang, G., Zheng, Y., Liu, Y., & He, W. (2020). NICE: Superpixel Segmentation Using Non-Iterative Clustering with Efficiency. *Applied Sciences 2020, Vol. 10, Page 4415*, *10*(12), 4415. https://doi.org/10.3390/app10124415

[56] Subhash Chander Goud, O., Hitendra Sarma, T., & Shoba Bindu, C. (2025). BAND SELECTION USING COEFFICIENT OF VARIATION AND BAND DENSITY RANKING IN HYPERSPECTRAL IMAGE. *Journal of Theoretical and Applied Information Technology*, *103*(18). www.jatit.org

[57] Xi, S., Shi, J., Yan, J., Lin, M. J., Zhou, X., Cheng, Y., Ding, H., & Kang, C. C. (2025). Breaking barriers in ICD classification with a robust graph neural network for hierarchical coding. *Scientific Reports 2025 15:1*, *15*(1), 25676-. https://doi.org/10.1038/s41598-025-10590-1

[58] Moulali, U., Vijayarangan, R., Khaleel Ahamed, S., & Kolli, K. (2025). GCNN with Adaptive Filters for Hyperspectral Image Classification. *Graph Convolutional Neural Networks for Computer Vision*, 117–140. https://doi.org/10.1002/9781394356362.ch6

[59] González-Ramírez, A., López, J., Torres-Román, D., Yáñez-Vargas, & Israel. (2021). *Analysis of multi-class classification performance metrics for remote sensing imagery imbalanced datasets Análisis de métricas de rendimiento de clasificación multiclase para conjuntos de datos desbalanceados de imágenes de percepción remota*. *8*, 11–17. https://doi.org/10.35429/JQSA.2021.22.8.11.17

[60] de Leeuw, J., Jia, H., Yang, L., Liu, X., Schmidt, K., & Skidmore, A. K. (2006). Comparing accuracy assessments to infer superiority of image classification methods. *International Journal of Remote Sensing*, *27*(1), 223–232. https://doi.org/10.1080/01431160500275762

[61] Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30.