ADVANCES IN GEODESY AND GEOINFORMATION
(formerly GEODESY AND CARTOGRAPHY)
Vol. 74, no. 1, article no. e62, 2025

https://doi.org/10.24425/agg.2025.150692

Original paper

Part of Topical collection:
"Advancements in Applied Geoinformatics"

Global-scale explainable AI assessment for OBIA-based classification using Deep Learning and Machine Learning methods

Taskin Kavzoglu, Yakup Kaan Uzun, Emre Berkan, Elif Ozlem Yilmaz*

Gebze Technical University, Kocaeli, Turkey
e-mail: kavzoglu@gtu.edu.tr; ORCID: http://orcid.org/0000-0002-9779-3443
e-mail: y.uzun2020@gtu.edu.tr; ORCID: http://orcid.org/0009-0007-2556-8363
e-mail: e.berkan2019@gtu.edu.tr; ORCID: http://orcid.org/0009-0009-9861-6113
e-mail: eoyilmaz@gtu.edu.tr; ORCID: http://orcid.org/0000-0002-6853-2148

*Corresponding author: Elif Ozlem Yilmaz, e-mail: eoyilmaz@gtu.edu.tr

Received: 2024-10-15 / Accepted: 2025-02-19

Abstract: Over the past decade, object-based image analysis (OBIA) has gained prominence as a widely adopted method for generating land use/land cover (LULC) maps. This study aims to evaluate the performance of various classification algorithms within the OBIA framework using SPOT-6 satellite imagery. The research methodology involved segmenting the images with the multi-resolution segmentation (MRS) algorithm, followed by the application of convolutional neural networks (CNN), random forest (RF), and support vector machine (SVM) algorithms for classification. The study was conducted in the Perpignan province, located in the Pyrénées-Orientales region of France. After the segmentation stage, CNN, RF, and SVM classifiers were employed to classify the image segments based on both spectral and spatial attributes. The accuracy of the resulting thematic maps was assessed using standard metrics, including overall accuracy (OA), the Kappa coefficient (KC), and the $F-$score (FS). Of the three classifiers, CNN achieved the highest overall accuracy at 91.28%, outperforming SVM, which attained an OA of 90.50%, and RF, which recorded an OA of 87.28%. Additionally, this study explored the integration of explainable artificial intelligence (AI) techniques, specifically the Shapley Additive Explanations (SHAP) algorithm, to enhance the interpretability of the machine learning models. This approach fosters greater trust, accountability, and acceptance in decision-making processes. By leveraging SHAP values, the study provides deeper insights into the decision-making processes of the CNN, SVM, and RF classifiers, ultimately enhancing the transparency and comprehensibility of these models.

Keywords: Support Vector Machine, Convolutional Neural Network, Random Forest, XAI, Object-Based Image Analysis

1. Introduction

Land use/land cover (LULC) maps are of critical importance and required for various applications, including land use mapping and monitoring, ecological assessment of vegetation communities, agricultural and forest management, evaluating land capability, and understanding interactions between the land surface and the atmosphere (Townshend, 1992; Chughtai et al., 2021; Wang et al., 2023). The creation of these maps has been enhanced by the development of remote sensing technology and image processing approaches (Pandey et al., 2021). Remote sensing technology is extensively utilized for the accurate extraction of LULC information due to its capability of periodically capturing synoptic data over large geographical areas (Moharram and Sundaram, 2023; Kavzoglu et al., 2024). The dynamic nature of LULC requires continuous monitoring and assessment; therefore, accurate predictions are essential to support improved planning and management through various methods (Singh et al., 2017; Panda et al., 2024). However, the inherent variability and complexity of remotely sensed data often lead to inefficiencies when traditional pixel-based techniques are employed. These methods are limited in their ability to capture high-level characteristics, as they fail to account for the spatial context associated with neighboring pixels (Gamanya et al., 2009). Moreover, pixel-based approaches have been reported to lose effectiveness for high- and very high-resolution images, as the spatial resolution of satellite imagery improves (Kavzoglu and Tonbul, 2018). These limitations suggest that object-based image analysis (OBIA) provides a more effective approach for analyzing high spatial resolution imagery. Unlike pixel-based methods, OBIA analyzes image objects – clusters of pixels – as the primary unit of analysis (Kavzoglu and Tonbul, 2017; Colkesen and Kavzoglu, 2019). In this context, OBIA leverages the spatial and hierarchical relationships of image segments to address more complex LULC classifications. Rather than focusing on individual pixels, OBIA evaluates groups of pixels, enabling a more comprehensive consideration of spatial, textural, and contextual features. A critical step in the OBIA framework is the image segmentation process, where regions corresponding to distinct land cover classes are divided into homogeneous sections to provide more meaningful representations of the landscape. The primary advantage of OBIA lies in its ability to integrate textural and contextual information with spectral data, significantly enhancing the accuracy of LULC maps. The overall success of OBIA, however, is contingent upon the quality of the segmentation process and the careful selection of its parameters (Kavzoglu, 2017).

Machine learning techniques, including random forest (RF) and support vector machines (SVM), have been widely employed in remote sensing research (Colkesen and Kavzoglu, 2018; Arfa and Minaei, 2024; Chowdhury, 2024). RF is particularly effective in managing complex relationships within high-dimensional datasets, while SVM is recognized for its ability to construct optimal decision boundaries between different classes in such feature spaces (Sheykhmousa et al., 2020). More recently, deep learning-based methods, especially convolutional neural networks (CNNs), have been increasingly adopted in OBIA applications since they offer significant advantages in the analysis of complex, high-resolution images due to their capacity for automatic feature extraction (Kavzoglu and Yilmaz, 2022). Recent research reveals that the integration of CNNs with

the OBIA framework has demonstrated superior classification performance, as these models are well-suited for handling segmented image objects and accurately classifying various LULC types (Kavzoglu et al., 2018; Azeez et al., 2022; Ameslek et al., 2024).

In addition to these advancements, Explainable Artificial Intelligence (XAI) methods like SHapley Additive exPlanations (SHAP) have been recently introduced to provide critical insights into the decision-making processes of both deep learning and traditional machine learning models. SHAP analysis, which is based on game theory, leverages Shapley scores to generate interpretable explanations of how individual features contribute to classification outcomes (Lundberg et al., 2020). This transparency enhances the interpretability of classification models and fosters greater confidence in their results. The incorporation of XAI techniques has made a significant contribution into the reliability, trustworthiness, and accountability of classification results derived from the analysis of remotely sensed data.

This study aims to assess the effectiveness of OBIA in the interpretation of high-resolution imagery, while also evaluating the performance of several machine and deep learning algorithms – namely CNN, SVM, and RF – in the classification process. The CNN model was employed to explore the high-accuracy potential of deep learning approaches, while SVM and RF served as benchmark methods, representing robust machine learning techniques. The primary objective was to investigate the ability of each classifier to learn and accurately classify spatial and spectral information within the image data. In addition, the study incorporated an XAI method, specifically the SHAP analysis technique, to enhance the reliability and transparency of the classification results. The proposed framework provides a comprehensive evaluation of the decision-making processes of each classifier, exposing the influence of individual features on the classification outcomes. In essence, the methodology employed involves the utilization of the SHAP technique to enhance the interpretability of the classifiers utilized in the OBIA-based LULC classification framework. The implementation of SHAP analysis facilitates the identification and ranking of the features that exert the most significant influence on the classification decisions, thereby offering a more profound insight into the decision-making processes employed by the various classifiers (CNN, SVM, and RF). The main objective of this study is to investigate the effectiveness of integrating SHAP analysis into the OBIA workflow, utilizing classifiers that operate on different principles, to enhance model transparency but also provide insights into feature prioritization, which is an important challenge in remote sensing applications.

2. Study area and dataset

The province of Perpignan, located in the Pyrénées-Orientales region of southern France, was selected as the study area for this research (Fig. 1). It is situated at the foot of the Pyrenees and agriculture, tourism, and trade are the main forms of the economic foundation of activities in the Perpignan province. The optical image of SPOT-6 covering the study site of about 47 km^2 was downloaded from Airbus Sample Imagery (https://intelligence.airbus.com/imagery/sample-imagery/). The image was acquired on October

3, 2023, and it contains a panchromatic band with a spatial resolution of 1.5 m and four spectral bands with a spatial resolution of 6 m (Table 1). The SPOT imagery was already pre-processed by the data providers to downsample the resolution of all spectral bands to 1.5 m using panchromatic bands. All bands were then merged into a multi-layer image, including spectral bands at red, green, blue, and near-infrared (NIR) wavelengths.

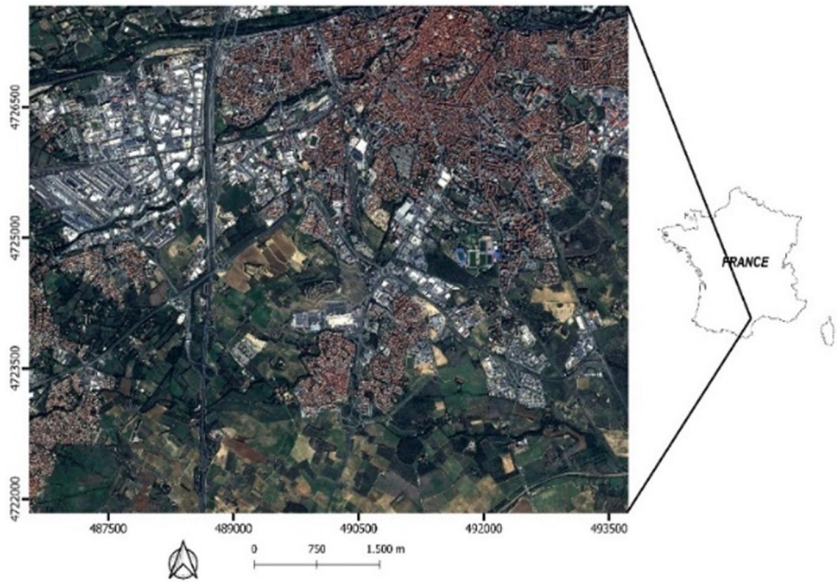

Fig. 1. The location of the study area in France

Table 1. Specifications of the SPOT-6 imagery

Band number	Band name	Wavelength (nm)	Resolution (m)
B1	blue	450–520	6
B2	green	530–590	6
B3	red	625–695	6
B4	near infrared	760–890	6
B5	panchromatic	450–745	1.5

Considering the distinctive characteristics associated with the study area, a total of six LULC classes were identified as road, forest, grassland, soil, red roof, and white roof. Subsequently, a total of 600 training samples and 300 test samples were collected for each class, to evaluate the performance of the classification algorithms. The creation of this dataset entailed the allocation of points to segments, thereby ensuring the representation of LULC classes. The samples were collected from homogeneous areas to facilitate spectral separation, and the creation of the dataset was meticulously performed to ensure adequate representation of the LULC classes.

3. Methodology

The OBIA approach was used to generate LULC maps from SPOT images using deep learning (i.e., CNN model) and machine learning (i.e., SVM and RF) methods. The framework of OBIA generates more significant and uniform data structures by aggregating adjacent pixels into image segments (Tonbul and Kavzoglu, 2020a; 2020b; Kavzoglu et al., 2024). It can be used to generate more detailed LULC maps, especially for intricate regions. The satellite image was segmented using the multi-resolution segmentation (MRS) approach, and LULC maps were generated using CNN, SVM, and RF methods. Furthermore, the performance of the segment-based classifiers was evaluated by computing the $F-$score values, Overall Accuracy (OA), and Kappa Coefficient. Subsequently, the XAI method of SHAP was employed to assess the contribution of each spectral band and object feature, also elucidating the model's decision-making mechanism. In other words, the SHAP analysis was used to clarify the attributes on which classification models grounded their conclusions and assess the influence of segment properties on categorization. The computer hardware used for the processes in this study is a PC with AMD Ryzen 7 5700X 8-core processor, 64 GB of RAM, and NVIDIA GeForce GTX 4700 Super graphics card. The stages of the methodology adopted in this study are shown as a flowchart in Figure 2.

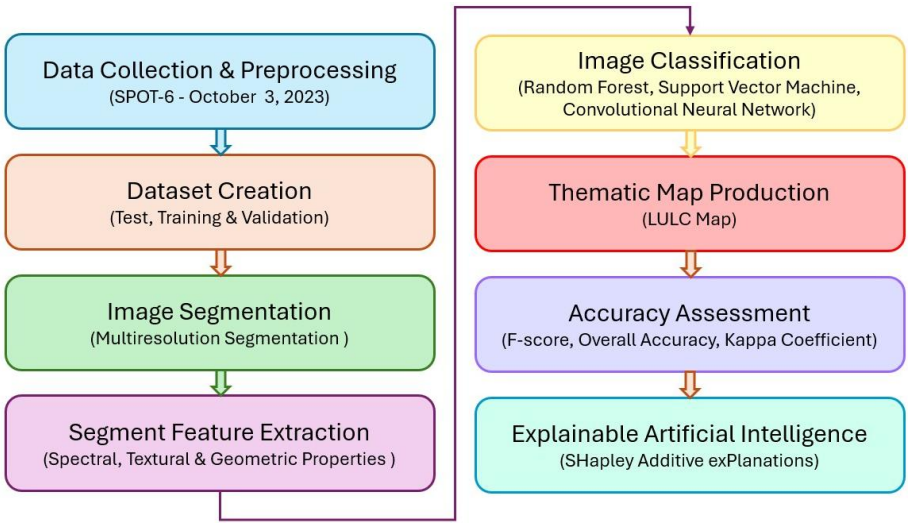

Fig. 2. The workflow of the study

3.1. Multiresolution Segmentation

The Multiresolution Segmentation (MRS) algorithm provides a comprehensive view by analyzing the objects or areas at different resolution levels. First, the image was downsampled to multiple resolution levels within a pyramidal structure. Each level represents a larger but less detailed version of the image. The image was then divided

into sections with similar characteristics at each resolution level, often comparing pixels based on color, brightness, and texture. These sections were transformed into segments through zoning, in which similar regions are combined or separated. It should be noted that the segments represent specific elements at each resolution level and are organized hierarchically to show connections between segments at different levels of detail (Yilmaz and Kavzoglu, 2024). The MRS algorithm is characterized by three significant parameters: shape, compactness, and scale. The scale parameter is employed to control the dimensions of image objects, with a large scale resulting in large image segments. The shape parameter facilitates the separation of spectral classes, while the compactness parameter primarily affects the quality of image object boundaries. The values of shape and compactness parameters range from 0 to 1. It can be said that the selection of these parameters has a significant impact on the classification accuracy. In addition, the Estimation of Scale Parameter 2 (ESP2) tool, developed by Drăguţ et al. (2014), was employed to determine the optimal scale parameter for the MRS algorithm. It considers the characteristics of objects and structures at different scales to improve the effectiveness of the algorithm and optimize the segmentation results.

3.2. Random Forest

The random forest (RF) algorithm, a popular algorithm applied to a wide variety of problems, uses several trees to train and predict the problem under consideration. It is recognized for its robustness and effectiveness in managing complex relationships in high-dimensional datasets. The algorithm, proposed by Breiman (2001), includes multiple decision trees, and the output class is established by the classification method produced by the individual trees, the so-called 'forest' in that the output is established by the individual trees. Each decision tree operates independently using recursive partitioning to iteratively divide the feature space into smaller and more homogeneous segments guided by feature values. This allows input data points to be classified efficiently (Sheykhmousa et al., 2020; Kavzoglu and Bilucan, 2023). This approach can enhance the generalization and predictive capability of a model by employing diverse training sample subsets to augment the variance across the classification models. In contrast to other machine learning techniques, RF employs an ensemble method, demonstrating superior accuracy relative to most individual methods. RF exhibits excellent accuracy, resistance to overfitting, a smaller number of parameters, and reliable performance. On the other hand, an important advantage of the RF algorithm is that feature importances can be easily estimated during the training process, which makes it a good choice as a feature selection method (Kavzoglu et al., 2024). The algorithm is reported to yield commendable outcomes in remote sensing applications (Belgiu and Drăguţ, 2016; Franklin and Ahmed, 2017; Linhui et al., 2020; Tonbul et al., 2022).

3.3. Support Vector Machine

SVM is an effective and frequently employed supervised machine learning technique. In contrast to other supervised learning techniques, SVM can provide effective classification outcomes with a limited training dataset (Mantero et al., 2005; Pal and Foody, 2012;

Kavzoglu et al., 2024). It tackles the specific problem: while the classes are separated by lines, the algorithm identifies a hyperplane that minimizes the generalization error while maximizing the separation between the two groups (Cortes and Vapnik, 1995; Mountrakis et al., 2011). The objective of SVM training is to identify an "optimal separating hyperplane" that effectively separates the dataset into its distinct, predefined classes. In other words, SVM attempts to identify the optimal hyperplane that separates data points belonging to different classes in the feature space by maximizing the margin, which is the distance between the nearest data points of each class (Sheykhmousa et al., 2020). Utilizing the kernel approach, SVM transforms non-linear data into a higher-dimensional space, where they can establish a linear separation between the two classes. The transformation of data into a higher-dimensional space generally results in its dispersal, thereby facilitating the identification of a linear division. Thus, non-linear separations within the original data space can acquire a linear nature within the higher-dimensional space (Kavzoglu et al., 2024). In other words, the SVM algorithm is characterized by its ability to effectively manage high-dimensional feature spaces and make decisions based on non-linear boundaries. On the other hand, for nonlinearly distinct classes, a technique involving the projection of input data onto an extensive feature space with high dimensions using kernel functions was introduced, which transformed the problem into a linear classification problem within that space (Pal and Mather, 2005). It should be also mentioned that the success of the algorithm brought about the recent introduction of various variants, including relevance vector machines, twinSVM, and deepSVM (Kavzoglu et al., 2024).

3.4. Convolutional Neural Network

Convolutional neural networks (CNNs) are of the most popular deep learning models, effectively capturing spatial and spectral features in image data (Song et al., 2019). CNN-based models can capture the distinctive features of the dataset, enabling them to differentiate without relying on complex, human-generated rules. The basic architecture of the model comprises convolutional and pooling layers that extract local patterns from images (Kavzoğlu and Yilmaz, 2022). They can vary in dimensions, although they universally consist of input, hidden, and output layers. The input layer accepts a one-dimensional matrix containing a feature value for each element of the input data in the 1D CNN model. Each convolutional layer consists of convolutional filters, with parameters adjusted by back-propagation methods. The dataset undergoes convolution with a sequence of trainable filters that thoroughly traverse the dataset, producing feature maps. Consequently, the dataset's most significant attributes are retrieved by convolutional processing. This approach can consistently reduce data size, leading to a drop in the total number of parameters and computational costs. Consequently, the issue of overfitting could be mitigated (Kavzoglu et al., 2021). The activation function, the ReLU function in this study, is applied after the convolutional layers. The ReLU imparts a non-linear architecture to the model by nullifying negative values, thereby accelerating the computational process. Optimization techniques are employed to reduce the loss function in the model training process. These techniques allow dynamic adjustment of the learning rate, resulting in

fast and reliable performance in gradient-based optimization processes (Kattenborn et al., 2021). The CNN model designed for the problem under consideration in this study is conceptually shown in Figure 3. This model comprising successive Conv1D layers, each of which contains a ReLU activation function, was performed. These layers include 32, 64, and 128 filters with kernel sizes of 3 and 1. Each convolutional layer was optimized for feature extraction by supporting MaxPooling1D. The model is then connected to a flattened layer, followed by fully connected (dense) layers of 256, 128, and 64 neurons. The model also employs the SoftMax activation function for 6 output classes in the final layer to perform the classification task. The model was constructed using the Adam optimization algorithm and categorical cross-entropy loss.

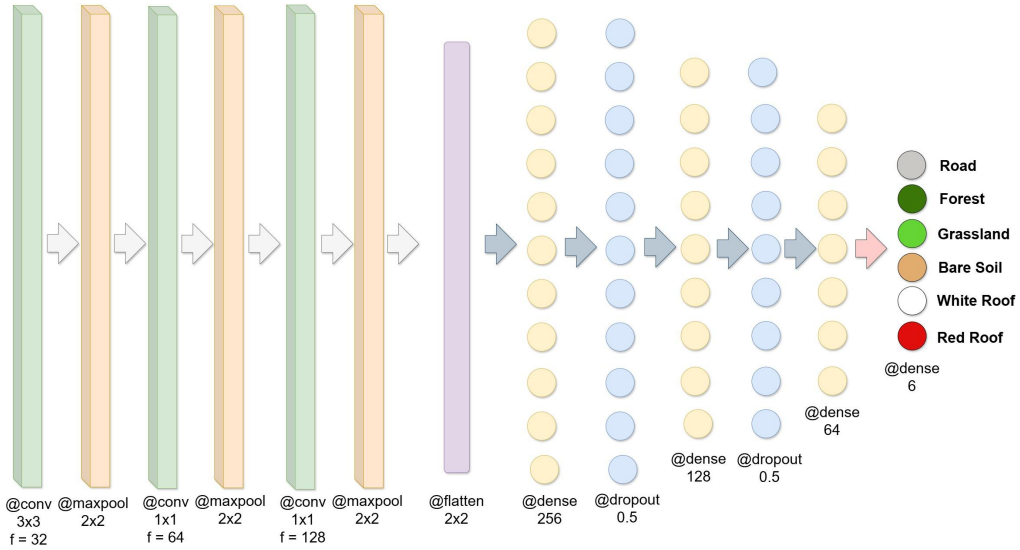

Fig. 3. The CNN model designed for the dataset used in this study

3.5. SHapley Additive eXPlanations

SHapley Additive exPlanations (SHAP) analysis is employed to interpret the predictions made by the machine learning models (Lundberg and Lee, 2017; Kavzoglu and Teke, 2022). The method utilizes Shapley values from basic game theory to equitably allocate the influence of each characteristic. Positive and negative numbers reflect each feature's contribution to each category for each data point. A higher mean absolute Shapley value across all data points indicates the feature's overall increased significance. This technique is more reliable than standard measurements. It also accounts for the interactions among the characteristics. Consequently, in addition to offering a comprehensive overview of the algorithm, the SHAP approach provides localized insights into the contribution of each attribute to sample-specific attribution. In other words, it provides an objective and interpretable explanation of how each feature influences the classification result using Shapley scores. The SHAP methodology is model-agnostic and can be used to elucidate the outputs

of various deep learning and machine learning approaches. Furthermore, this approach uses various graphical representations to gain insight into the decision-making processes of classification algorithms, thereby increasing their reliability and interpretability.

SHAP analysis enables the evaluation of the impact of each feature on the model, thereby facilitating the interpretation of black box models. In instances of multicollinearity in linear models, Shapley regression values are employed to ascertain feature importance. To calculate these values, the model is retrained with feature subsets, and the impact of adding or removing a feature on prediction is measured. In the mathematical formula, F is used to represent all features, S is used to indicate specific feature subsets, i is used to specify the selected feature, and f is used to represent the model. Furthermore, x is used to denote the input data (feature values) that is given to the model in the SHAP analysis (Kavzoglu et al., 2024).

$$\emptyset_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S(x_S) \right]. \tag{1}$$

The SHAP summary plot is a visual tool that provides an explanation of the decisions made by the classifiers by displaying the calculated values. The y-axis of the graph represents the object features that were utilized in the classification process, while the x-axis shows the calculated SHAP values. In essence, SHAP values signify the impact of a feature on the prediction of the classification model. On the horizontal axis, features located to the right of zero indicate a positive contribution (where the model's prediction positively affects the feature in question). At the same time, those positioned to the left of zero demonstrate a negative contribution (where the model's prediction negatively affects the feature). To illustrate this, as the SHAP value on the x-axis of the graph moves away from zero, the magnitude of the effect of that feature on the prediction increases. When the color codes are examined, a low feature value is shown as blue, and a high feature value is shown as red. It is further noted that the position of features in these graphs, with those at the top having the most significant impact on model estimation, while those at the bottom play a less substantial role.

3.6. Accuracy assessment

In the domain of remote sensing, a range of metrics are employed to evaluate the accuracy of classification outputs (Foody, 2020; Kavzoglu et al., 2024). These metrics in this context are generally classified into two categories: global and class-based accuracy metrics. Global metrics (e.g., OA and Kappa Coefficient) provide an Overall Accuracy rating for the entire thematic map, while class-based metrics (e.g. $F-$score, user's and producer's accuracy) calculate the accuracy of individual LULC classes within the thematic map. The most common technique for calculating metrics involves the use of a confusion matrix, a tool that allows the calculation of the relationship between classified data and the reference data, from which the aforementioned metrics are generated. The calculation of OA involves the ratio of correctly classified data to all predictions. The Kappa Coefficient

is employed as a reliable metric because it eliminates the effect of random guesses. Additionally, the $F-$score metric is utilized for class-based evaluation in thematic maps. The calculation of this metric necessitates the values of the producer's and user's accuracy. Producer's accuracy calculates how many of all instances of a class are correctly classified, while User's accuracy shows how many of the instances predicted as a class belong to that class. The $F-$score value is calculated by taking the harmonic mean of these two metrics (Vizzari, 2022). All the metrics mentioned above are calculated according to the following formulas:

$$\text{Overall Accuracy} = \frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}}, \tag{2}$$

$$\text{Kappa Coefficient} = \frac{\text{Observed agreement} - \text{Expected agreement by chance}}{1 - \text{Expected agreement by chance}}, \tag{3}$$

$$F-\text{score} = 2 * \frac{\text{Producer's Accuracy} * \text{User's Accuracy}}{\text{Producer's Accuracy} + \text{User's Accuracy}}. \tag{4}$$

4. Results

In the applications of OBIA-based classifications, segmentation, which is a fundamental stage of OBIA and facilitates the identification of essential objects, is conducted before classification. The MRS segmentation approach was applied using specialized eCognition software. The effectiveness of segmentation algorithms is primarily dependent on image attributes and meticulous parameter selection. Therefore, the first step was to determine the most effective scale parameter for the MRS algorithm using the ESP2 tool. As shown in Figure 4, the scale parameter with the highest optimal performance was detected as 36. Following the recommendations in the literature (e.g., Kavzoglu and Tonbul, 2018), the shape and compactness parameters were set to 0.1 and 0.5, respectively. When the MRS algorithm was applied to the SPOT image with the estimated parameters, a segmented image (a total of 200,692 segments), including image objects representing homogeneous regions, was obtained. The segments were then analyzed to calculate various features, which were introduced as inputs to the OBIA processes. The process involved the evaluation of various characteristics of each object, including its area, boundary length, brightness properties, HSI transformation-based features, object length, maximum difference, GLCM contrast, dissimilarity, entropy, and homogeneity. In addition, features unique to spectral bands were considered, including mean and minimum pixel values, ratios, and standard deviations (Table 2). Calculating these features is crucial in the classification process because it allows algorithms to effectively identify and categorize objects by determining their spectral, morphological, and textural features. For instance, geometric features provide information about the size and shape of objects, while spectral features allow the object belonging to the LULC class to be spectrally separated from other classes. On the other hand, textural features examine the consistency in the distribution of pixels belonging to the LULC class. For the application of RF and SVM methods, the

optimal parameters were determined through a process of trial and error for machine learning classifiers. Within the RF algorithm, the number of decision trees was set at 100, with a depth of 5 for each tree. A linear kernel was utilized for SVM.

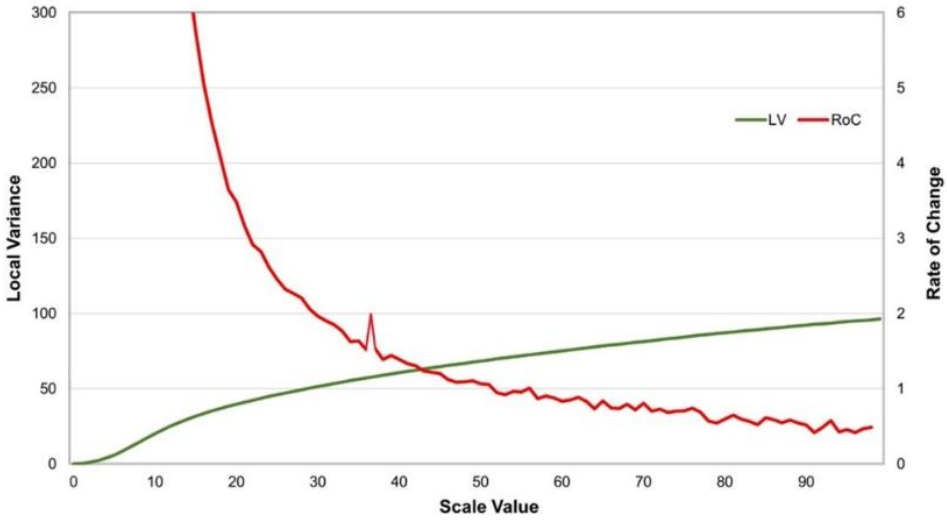

Fig. 4. Estimation of scale parameters identified using ESP2 for the SPOT-6

Table 2. Spectral, geometric and textural characteristics of the segments used to study

Feature type	Feature name	Band name			
		Red	Green	Blue	NIR
Spectral	Ratio	✓	✓	✓	✓
	HSI transformation	✓	✓	✓	×
	Minimum	✓	✓	✓	✓
	Maximum	✓	✓	✓	✓
	Mean	✓	✓	✓	✓
	Standard deviation	✓	✓	✓	✓
	Brightness	✓			
	Maximum difference	✓			
Geometric	Border length	✓			
	Area	✓			
Textual (GLCM)	Homogeneity	✓			
	Dissimilarity	✓			
	Contrast	✓			
	Entropy	✓			

In the CNN training stage, the model was trained on 64 mini-batches for 800 epochs and evaluated using the accuracy metric. In addition, model training was performed using trial-and-error hyperparameters. A learning curve was generated to evaluate the training period (Figure 5). Looking at the loss and accuracy curves of the training and validation processes of the model, it can be stated that the model did not overfit. In the left-hand plot showing the training and validation losses in Figure 5, a rapid decrease is observed in both curves at the beginning and then stabilizes at a fixed point. The difference between the training and validation losses was minimal. This indicates that the model fits the validation data well, which indicates that overlearning was not performed in the training stage. In the accuracy graph on the right, it can be seen that both the training and validation accuracy curves are close to each other and are close to 1. This indicates that the model achieves high performance on both training and validation data and exhibits strong generalizability. In summary, the model performance was high, and there was no overfitting.

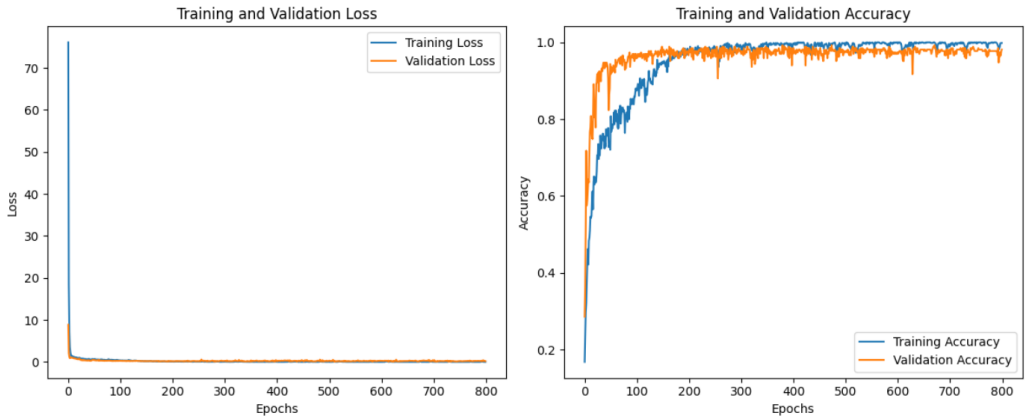

Fig. 5. Learning curves of the segment-based CNN model

When examining the results in terms of map-based accuracy metrics, the CNN model obtained the highest OA of 91.28%, followed closely by SVM with 90.50%, and a lower OA of 87.28% was achieved with the RF algorithm. Considering the estimated Kappa Coefficient values, similar results were observed. The CNN model provided the highest agreement, with 0.90, while the SVM model had a lower agreement value of 0.89, followed by RF with 0.85. These results demonstrate that the CNN model exhibits superior performance, especially in distinguishing spectral and spatial details; however, the SVM model appeared as a strong alternative. The performance of the RF model was inferior when handling more complex classification tasks. In particular, high accuracy levels were achieved by the CNN and SVM models in classes with distinct spectral features, including red roof and white roof. Another important finding was related to the relatively poor performances of CNN and SVM for more heterogeneous classes, particularly road and soil classes. Overall, the results demonstrate that the robustness of the CNN model in OBIA-based LULC classification, with similar performance of the SVM model.

Classification performances estimated for the three classifiers using the OBIA approach are shown in Table 3 where comparisons can be made using statistical metrics such as $F-$score, OA, and Kappa Coefficient. The results showed that the CNN model achieved the highest $F-$score of 97.00% for the forest class, followed by the white-roof (95.95%) and red-roof (94.02%) classes, while the $F-$score of the CNN was relatively low at 80.68% for the road class. The SVM model outperformed the CNN model with 82.02% for this class but achieved lower $F-$score s for the forest (94.44%) and grassland (88.50%) classes. The RF model generally performed worse, especially for road (74.48%), grassland (84.15%), and soil (81.44%) classes, but it achieved the highest $F-$score of 96.23% for the red roof class. Based on the analysis of the results, the $F-$score values obtained from the classification process indicated that the road class exhibited a low value across all classifiers. The road class typically encompasses structures characterized by narrow and long geometries, which can be subdivided into non-homogeneous segments during the segmentation process. It is posited that the presence of objects such as trees in certain areas and the suboptimal resolution of satellite images, which hinders effective classification, contributes to the observed metrics.

Table 3. Accuracy assessment for the thematic maps produced by CNN, SVM and RF methods

$F-$score (%)	CNN	SVM	RF
Road	80.68	82.02	74.48
Forest	97.00	94.44	93.09
Grassland	91.10	88.50	84.15
Soil	88.71	86.38	81.44
Red roof	94.02	95.49	96.23
White roof	95.95	95.96	94.27
Overall acc. (%)	91.28	90.50	87.28
Kappa coefficient	0.90	0.89	0.85

The thematic maps (Fig. 6) show the LULC classifications of the CNN, SVM, and RF algorithms. The map produced by CNN generally exhibits a more balanced distribution among the LULC classes. In particular, the forest and grassland areas were identified clearly. Red and white roof classes were also well represented in urban areas. The road class was separated alongside the main arteries, indicating that the performance of the model was reasonable. However, confusion was observed in the determination of the boundaries of the bare soil and grassland classes in some parts of the study site. On the other hand, the thematic map produced by SVM showed clearer class boundaries compared to CNN, in particular, a clear distinction between the forest, grassland, and bare soil classes. However, it appears that the classifier was not effective in the delineation of red and white roofs in urban areas. Although the road class was clearer than that of CNN, the narrow roads in rural areas were not detected accurately. The SVM algorithm was mostly effective for the determination of urban and rural areas. On the other hand, the map produced by RF included more erroneous predictions than the other two models and

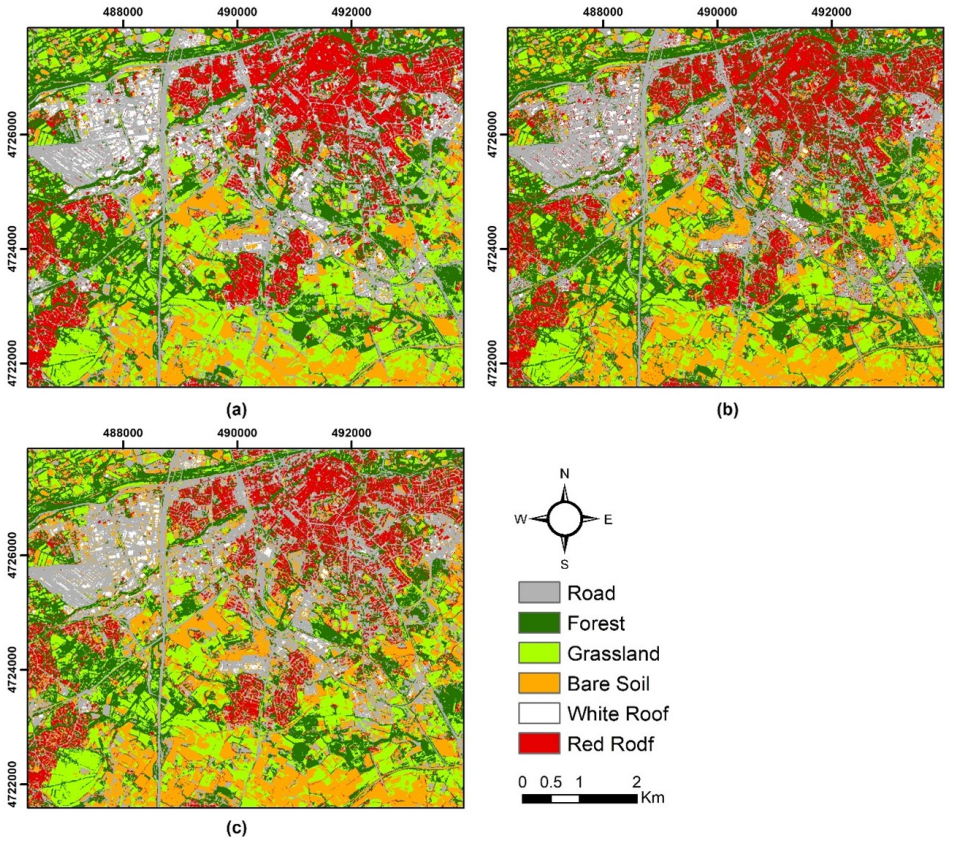

Fig. 6. Thematic maps generated using (a) CNN, (b) SVM and (c) RF methods

exhibited a significant margin of error, particularly in distinguishing between grassland and bare soil classes. Although the classification of roof types in urban areas (red and white roofs) appeared to be more accurate than the other models, it is recognized that the RF model had difficulty in distinguishing the classes from each other using the employed features. There is also considerable complexity in the identification of road class, except for the main roads, which were mixed with other classes. To highlight the differences in the classification maps, three subsets were extracted from the produced thematic maps (Fig. 7). As can be seen from the figure, the RF method performed worse than CNN and SVM, especially in terms of clarity and discrimination of class boundaries. While the CNN and SVM models provided more detailed and accurate results in LULC classification, the RF model had difficulty distinguishing classes; hence it produced lower prediction accuracies.

SHAP summary plots obtained by the three classifiers are shown in Figure 8, which shows the global impact of the main features that influence the decisions of each classifier model. In the CNN model (Fig. 8a), object features such as “Mean-NIR”, “Mean-Blue” and “Mean-Green” were the most influential in the model’s decision. At the same

time, GLCM-based metrics (contrast, dissimilarity, entropy) also played important roles; however, their effects were limited. In the SVM model (Figure 8b), "Mean-NIR", "Mean-Red" and "Mean-Green" had the highest SHAP values and were the most decisive factors for classification accuracy. However, it was observed that spatial features such as "GLCM_contrast" and "Border_len" also had a significant impact on the classification

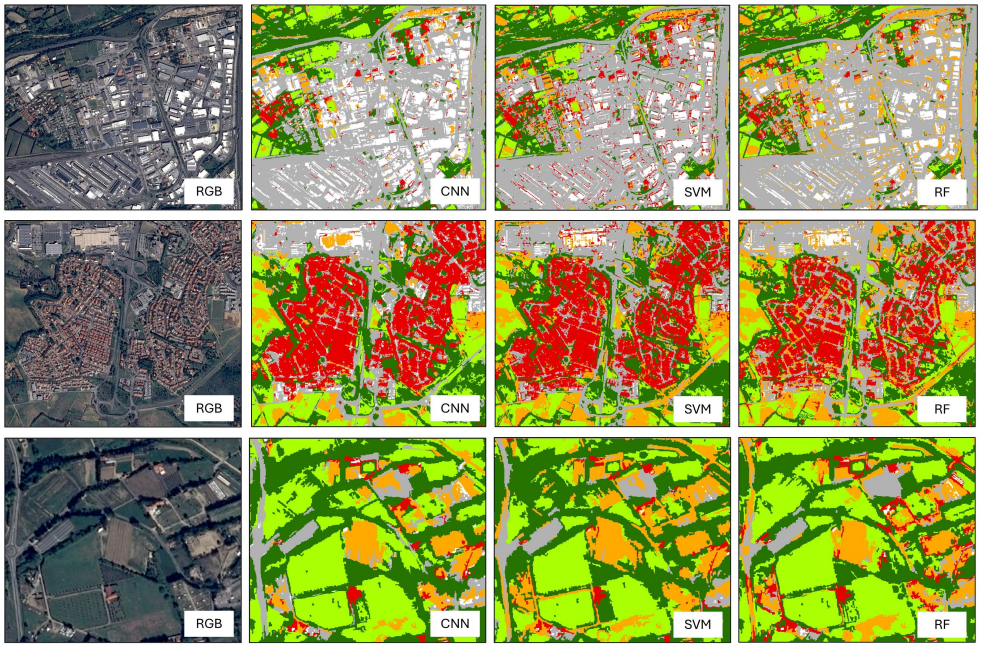

Fig. 7. Subset examples of thematic maps generated using CNN, SVM, and RF classification algorithms

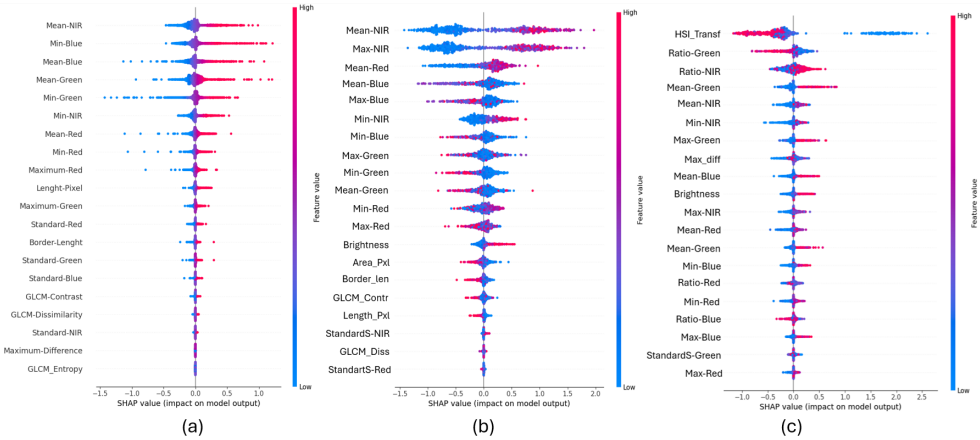

Fig. 8. SHAP summary plots for (a) CNN, (b) SVM, and (c) RF classifiers

result. In the RF model (Figure 8c), more complex spectral features such as "HSI_Transf" and "Ratio-Green" are observed, whereas traditional spectral bands such as "Max-Green" and "Mean-NIR" were also decisive. This comparison between the three classifiers showed that the CNN was better at evaluating the spectral information of the objects, the SVM worked with more limited but distinct features, and the RF considered complex spectral ratios. In general, these graphs provided a visual representation of which features were more important for the models constructed with object features while deepening the model insight in terms of explainability.

5. Discussion

In this study, OBIA-based LULC classification was performed using three machine learning classifiers having different working principles. Results indicated the superiority of the CNN model with the highest classification accuracies in terms of global and class-based accuracy metrics. The SHAP analysis revealed that CNN evaluated spectral information more effectively by assigning more importance compared to the other two models. Supporting the findings of this study, some recent studies have asserted the efficacy of deep learning-based models in discerning spatial and spectral relationships in remote sensing data (Zhang et al., 2018). On the other hand, the SVM classifier demonstrated its capacity to effectively utilize limited yet distinctive features, particularly for the road class characterized by thin and elongated geometries (Mountrakis et al., 2011). The RF model, while considering more complex spectral ratios, exhibited limited success in heterogeneous classes, as evidenced by its low performance in these categories (Pal, 2005). The observed low accuracy rates, particularly for the road class, may be attributed to the segmentation of the roads into non-homogeneous segments using the scale parameter of 36 due to their narrow and elongated geometric structures. Additionally, challenges related to the classification of some urban features, such as roads, have been identified as a prevalent issue in the existing literature (Blaschke, 2010; Hossain and Chen, 2019). To tackle this problem, it may be useful to use VHR satellite images, apply more robust hyperparameter optimization approaches, or optimize segmentation algorithms by finding the best parameter combination. As underlined by Kavzoglu et al. (2024), "In cases where the proportions of classes are distributed unevenly or are imbalanced compared to the actual land cover proportions, classification methods may exhibit bias". Therefore, the existence of imbalance negatively affects the classification process, as classifiers, especially machine learning algorithms, tend to prioritize the dominant class and overlook the rare classes. This can be also valid for the determination of scale parameter for spatially imbalanced samples, as in the case of road class in this study. In this study, SHAP analysis provided valuable insights into focusing on important features to increase classification accuracy. XAI methods, SHAP plots in this particular case, stand out as an important tool in strengthening the reliability and interpretability of classification models. The findings obtained from the study revealed that CNN and SVM models provided particularly high accuracy, while the RF model exhibited limited performance. These results underscore the pivotal role of appropriate model selection and effective feature utilization in remote sensing classification studies.

6. Conclusion

In this study, CNN, RF, and SVM algorithms were employed to classify the image segments or objects created by the MRS method. The results indicated that the CNN algorithm outperformed the other machine learning methods, achieving higher OA and exhibiting more robust classification performance. The model was particularly effective in classifying the road class, which can be considered as a difficult class to identify as it appeared as a thin slick form mostly formed by mixed pixels in the image. As indicated by the global map-based accuracies, the CNN classifier achieved the highest performance with 91.28%, followed by the SVM classifier with 90.50% and then the RF classifier with 87.28%. Overall, these findings highlight the effectiveness of the CNN model in accurately classifying various surface types. Furthermore, the SHAP method, providing a framework for elucidating the decision-making processes of models, and enhancing the understanding of the internal mechanisms that drive classification outcomes, was utilized to assess the influence of individual features in the classification process. SHAP summary plots revealed that object features (e.g., mean spectral value) derived from the NIR band were influential across all classifiers. Additionally, the SHAP analysis revealed that the CNN model evaluated spectral information more effectively, the SVM model focused on fewer but more distinct features, and the RF model incorporated more complex spectral ratios. Thus, the SHAP analysis contributed to a better understanding of classification judgments by ranking the importance of features that influence model decisions. The analysis helped to identify which components should be prioritized to improve classification accuracy and strengthen confidence in the decisions of the model. Overall, this research demonstrates the value of XAI approaches in enhancing the interpretability and reliability of machine learning models, particularly in the context of LULC classification using OBIA. The study underscores how the integration of SHAP analysis can improve model explainability, fostering greater trust in the classification results. Further research will concentrate on the evaluation of model performance through the utilization of the most effective features identified by SHAP analysis. Moreover, a comparative analysis of classifiers under varying segmentation parameters by concentrating more on rare classes will be conducted with a view to further optimizing the integration of XAI approaches in LULC classification.

Author contributions

Conceptualization and methodology: T.K., E.O.Y.; formal analysis and investigation: E.O.Y., Y.K.U., E.B.; writing original draft preparation: T.K., E.O.Y., Y.K.U., E.B.; writing – review and editing: T.K., E.O.Y., Y.K.U., E.B.; supervision: T.K.

Data availability statement

The datasets used in this study are available from the corresponding author.

Acknowledgements

This research was not funded by external sources.

References

Ameslek, O., Zahir, H., Latifi, H. et al. (2024). Combining OBIA, CNN, and UAV imagery for automated detection and mapping of individual olive trees. *Smart Agric. Technol.*, 9, 100546. DOI: 10.1016/j.atech.2024.100546.

Arfa, A., and Minaei, M. (2024). Utilizing multitemporal indices and spectral bands of Sentinel-2 to enhance land use and land cover classification with Random Forest and Support Vector Machine. *Adv. Space Res.*, 74(11), 5580–5590. DOI: 10.1016/j.asr.2024.08.062.

Azeez, O.S., Shafri, H.Z., Alias, A.H. et al. (2022). Integration of Object-Based Image Analysis and Convolutional Neural Network for the classification of high-resolution satellite image: a comparative assessment. *Appl. Sci.*, 12(21), 10890. DOI: 10.3390/app122110890.

Belgiu, M., and Drăguţ, L. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.*, 114, 24–31. DOI: 10.1016/j.isprsjprs.2016.01.011.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS J. Photogramm. and Remote Sens.*, 65(1), 2–16. DOI: 10.1016/j.isprsjprs.2009.06.004.

Breiman, L. (2001). Random Forests. *Mach. Learn.*, 45, 5–32. DOI: 10.1023/A:1010933404324.

Chowdhury, M.S. (2024). Comparison of accuracy and reliability of Random Forest, Support Vector Machine, Artificial Neural Network and Maximum Likelihood method in land use/cover classification of urban setting. *Environ. Chall.*, 14, 100800. DOI: 10.1016/j.envc.2023.100800.

Chughtai, A.H., Abbasi, H., and Karas, I.R. (2021). A review on change detection method and accuracy assessment for land use land cover. *Remote Sens. Appl.: Soc. Environ.*, 22, 100482. DOI: 10.1016/j.rsase.2021.100482.

Colkesen, I., and Kavzoglu, T. (2018). Selection of optimal object features in object-based image analysis using filter-based algorithms. *J. Indian Soc. Remote Sens.*, 46(8), 1233-1242. DOI: 10.1007/s12524-018-0807-x.

Colkesen, I., and Kavzoglu, T. (2019). Comparative evaluation of decision-forest algorithms for object-based land use and land cover mapping. In H.R. Pourghasemi and C. Gokceoglu (Eds.), Spatial Modeling in GIS and R for Earth and Environmental Sciences, 499-517. DOI: 10.1016/B978-0-12-815226-3.00023-5.

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.*, 20. 273–297.

Drăguţ, L., Csillik, O., Eisank, C. et al. (2014). Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS J. Photogramm. Remote Sens.* 88, 119–127. DOI: 10.1016/j.isprsjprs.2013.11.018.

Foody, G.M. (2020). Explaining the unsuitability of the Kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sens. Environ.*, 239, 111630. DOI: 10.1016/j.rse.2019.111630.

Franklin, S.E., and Ahmed, O.S. (2017). Object-based wetland characterization using Radarsat-2 Quad-Polarimetric SAR Data, Landsat-8 OLI imagery, and airborne lidar-derived geomorphometric Variables. *Photogramm. Eng. Rem. S.*, 83(1), 27–36. DOI: 10.14358/PERS.83.1.27.

Gamanya, R., De Maeyer, P., and De Dapper, M. (2009). Object-oriented change detection for the city of Harare, Zimbabwe. *Expert Systems with Applications*, 36(1), 571–588. DOI: 10.1016/j.eswa.2007.09.067.

Hossain, M.D., and Chen, D. (2019). Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS J. of Photogramm. and Remote Sens.*, 150, 115–134. DOI: 10.1016/j.isprsjprs.2019.02.009.

Kattenborn, T., Leitloff, J., Schiefer, F. et al. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.*, *173*, 24–49. DOI: 10.1016/j.isprsjprs.2020.12.010.

Kavzoglu, T. (2017). *Object-oriented random forest for high resolution land cover mapping using quickbird-2 imagery*. In P. Samui, S. Sekhar, and V. E. Balas (Eds.), Handbook of Neural Computation, 607–619. Academic Press. DOI: 10.1016/B978-0-12-811318-9.00033-8.

Kavzoglu, T., and Tonbul, H. (2017). A Comparative Study of Segmentation Quality for Multi-Resolution Segmentation and Watershed Transform, Istanbul, Turkey. 8th International Conference on Recent Advances in Space Technologies (RAST), 19–22 June (pp. 113–117), Istanbul, Turkey. DOI: 10.1109/RAST.2017.8002984.

Kavzoglu, T., and Tonbul, H. (2018). An experimental comparison of multi-resolution segmentation, SLIC And K-Means Clustering for Object-Based Classification of VHR Imagery. *Int. J. Remote Sens.*, 39(18). 6020–6036. DOI: 10.1080/01431161.2018.1506592.

Kavzoglu, T., Tonbul, H., Yildiz Erdemir, M. et al. (2018). Dimensionality Reduction and Classification of Hyperspectral Images Using Object-Based Image Analysis. *J. Indian Soc. Remote Sens.*, 46, 1297–1306. DOI: 10.1007/s12524-018-0803-1.

Kavzoglu, T., Teke, A., and Yilmaz, E.O. (2021). Shared blocks-based ensemble deep learning for shallow landslide susceptibility mapping. *Remote Sens*., 13(23), 4776. DOI: 10.3390/rs13234776.

Kavzoglu, T., and Teke, A. (2022). Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using Random Forest, Extreme Gradient Boosting (Xgboost) and Natural Gradient Boosting (NGBoost). *Arab. J. Sci. Eng.*, 47(6), 7367–7385. DOI: 10.1007/s13369-022-06560-8.

Kavzoglu, T., and Yilmaz, E.Ö. (2022). Analysis of patch and sample size effects for 2D-3D CNN models using multiplatform dataset: hyperspectral image classification of ROSIS and Jilin-1 GP01 imagery. *Turk. J. Elec. Eng. Comp. Sci.*, 30(6), 2124–2144. DOI: 10.55730/1300-0632.3929.

Kavzoglu, T., and Bilucan, F. (2023). Effects of auxiliary and ancillary data on LULC classification in a heterogeneous environment using optimized random forest algorithm. *Earth Sci. Inform.*, 16(1), 415–435. DOI: 10.1007/s12145-022-00874-9.

Kavzoglu, T., Tso, B., and Mather, P.M. (2024). *Classification Methods for Remotely Sensed Data*. Third Edition, Boca Raton: CRC Press. DOI: 10.1201/9781003439172.

Linhui, L., Weipeng, J., and Huihui, W. (2020). Extracting the forest type from remote sensing images by Random Forest. *IEEE Sens. J.*, 21(16), 17447–17454. DOI: 10.1109/JSEN.2020.3045501.

Lundberg, S.M., and Lee., S.I. (2017). A unified approach to interpreting model predictions. In Proceedings of NIPS'17, December 4–9, Long Beach, CA, USA, 4765-4774.

Lundberg, S.M., Erion, G., Chen, H. et al. (2020). From Local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1), 56–67. DOI: 10.1038/s42256-019-0138-9.

Mantero, P., Moser, G. and Serpico, S.B. (2005). Partially supervised classification of remote sensing images through SVM-based probability density estimation. *IEEE Trans. Geosci. Remote Sens.*, 43, 559–570. DOI: 10.1109/TGRS.2004.842022.

Moharram, M.A., and Sundaram, D.M. (2023). Land use and land cover classification with hyperspectral data: a comprehensive review of methods, challenges and future directions. *Neurocomputing*, 536, 90–113. DOI: 10.1016/j.neucom.2023.03.025.

Mountrakis, G., Im, J., and Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.*, 66(3), 247–259. DOI: 10.1016/j.isprsjprs.2010.11.001.

Pal, M. (2005). Random Forest classifier for remote sensing classification. *Int. J. of Remote Sens.*, 26(1), 217–222. DOI: 10.1080/01431160412331269698.

Pal, M., and Mather, P.M. (2005). Support Vector Machines for classification in remote sensing. *Int. J. Remote Sens.*, 26(5), 1007–1011. DOI: 10.1080/01431160512331314083.

Pal, M., and Foody, G.M. (2012). Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE J. Sel. Top. Appl. Earth Obs.*, 5, 1344–1355. DOI: 10.1109/JSTARS.2012.2215310.

Panda, K.C., Singh, R.M., and Singh, S.K. (2024). Advanced CMD predictor screening approach coupled with cellular automata-artificial neural network algorithm for efficient land use-land cover change prediction. *J. Cleaner Prod.*, 449, 141822. DOI: 10.1016/j.jclepro.2024.141822.

Pandey, P.C., Koutsias, N., Petropoulos, G.P. et al. (2021). Land use/land cover in view of earth observation: data sources, input dimensions, and classifiers – a review of the state of the art. *Geocarto Int.*, 36(9), 957–988. DOI: 10.1080/10106049.2019.1629647.

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H. et al. (2020). Support Vector Machine versus Random Forest for remote sensing image classification: a meta-analysis and systematic review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 13, 6308–6325. DOI: 10.1109/JSTARS.2020.3026724.

Singh, S.K., Srivastava, P.K., Szabó, S. et al. (2017). Landscape transform and spatial metrics for mapping spatiotemporal land cover dynamics using earth observation data-sets. *Geocarto Int.*, 32(2), 113–127. DOI: 10.1080/10106049.2015.1130084.

Song, J., Gao, S., Zhu, Y. et al. (2019). A survey of remote sensing image classification based on CNNs. *Big Earth Data*, *3*(3), 232–254. DOI: 10.1080/20964471.2019.1657720.

Tonbul, H., and Kavzoglu, T. (2020a). Semi-Automatic building extraction from Worldview-2 imagery using Taguchi optimization. *Photogramm. Eng. Remote Sensing*, 86(9), 547–555. DOI: 10.14358/PERS.86.9.547.

Tonbul, H., and Kavzoglu, T. (2020b). A spectral band based comparison of unsupervised segmentation evaluation methods for image segmentation parameter optimization. *Int. J. Environ. Geoinformatics*, 7(2), 132–139. DOI: 10.30897/ijegeo.641216.

Tonbul, H., Colkesen, I., and Kavzoglu, T. (2022). Pixel-and object-based ensemble learning for forest burn severity using USGS FIREMON and Mediterranean condition dNBRs in Aegean ecosystem (Turkey). *Adv. Space Res.*, 69(10), 3609–3632. DOI: 10.1016/j.asr.2022.02.051.

Townshend, J.R.G. (1992). Land cover. *Int. J. Remote Sensing*, 13(6–7), 1319–1328. DOI: 10.1080/01431169208904193.

Vizzari, M. (2022). PlanetScope, Sentinel-2, and Sentinel-1 data integration for object-based land cover classification in Google Earth Engine. *Remote Sens.*, 14(11), 2628. DOI: 10.3390/rs14112628.

Wang, Y., Sun, Y., Cao, X. et al. (2023). A review of regional and global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing. *ISPRS J. Photogramm. Remote Sens.*, 206, 311–334. DOI: 10.1016/j.isprsjprs.2023.11.014.

Yilmaz, E.O., and Kavzoglu, T. (2024). Quality assessment for multi-resolution segmentation and segment-anything model using WorldView-3 imagery. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-4/W9-2024*, 48, 383–390. DOI: 10.5194/isprs-archives-XLVIII-4-W9-2024-383-2024.

Zhang, X., Chen, L., and Jia, X. (2018). Deep learning for remote sensing image classification: A comprehensive review. *IEEE Geosci. Remote Sens Mag.*, 6(3), 6–20. DOI: 10.1002/widm.1264.