Orthrus: multi-scale land cover mapping from satellite image time series via 2D encoding and convolutional neural network

Azza Abidi, Dino Ienco, Ali Ben Abbes, Imed Riadh Farah

Orthrus: Multi-Scale land cover mapping from satellite image time series via 2D encoding and Convolutional Neural Network

Azza Abidi[1,2*], Dino Ienco[2,3], Ali Ben Abbes[1], Imed Riadh Farah[1]

[1]RIADI Laboratory, National School of Computer Science, Manouba, 2010, Tunisia.
[2]UMR TETIS, INRAE, University of Montpellier, Montpellier, 34000, France.
[3]INRIA, Montpellier, 34000, France.

*Corresponding author(s). E-mail(s): azza.abidi@etu.umontpellier.fr;
Contributing authors: dino.ienco@inrae.fr; ali.benabbes@yahoo.fr;
imedriadh.farah@mse.uma.tn;

Abstract

With the advent of modern Earth Observation (EO) systems, the opportunity of collecting Satellite Image Time Series (SITS) provides valuable insights to monitor spatio-temporal dynamics. Within this context, accurate Land Use/Land Cover (LULC) mapping plays a pivotal role in supporting territorial management and facilitating informed decision-making processes. However, traditional pixel-based and object-based classification methods often face challenges to effectively exploit spectral and spatial information. In this study, we propose Orthrus, a novel approach that fuses multi-scale information for enhanced LULC mapping. The proposed approach exploits several 2D encoding techniques to encode times series information into imagery. The resulting image is leveraged as input to a standard Convolutional Neural Network (CNN) image classifier to cope with the downstream classification task. The evaluations on two real word benchmarks namely, Dordogne and Reunion-Island, demonstrated the quality of Orthrus over state-of-the-art techniques from the field of land cover mapping based on SITS data. More precisely, Orthrus exhibits an enhancement of more than 3.5 accuracy points compared to the best competing approach on the `Dordogne` benchmark, and surpasses the best competing approach on the `Reunion-Island` dataset by over 3 accuracy points.

Keywords: Deep Learning, Pixel-object classification, Convolutional Neural Networks (CNN), Multivariate Time-Series , classification, 2D Encoding representation, Land use land cover

1 Introduction

Land Use and Land Cover (LULC) maps serve as a fundamental information for territory management, land use planning, and supporting a wide range of applications [1–3]. It plays a pivotal role in monitoring and comprehending the dynamic changes occurring on the Earth's surface over time. Leveraging remote sensing techniques, accurate LULC mapping can be effectively achieved, facilitating crucial monitoring and understanding of these surface changes. The availability of precise and up-to-date land cover information is paramount for various purposes such as urban settlement monitoring, forest assessment, characterization of agricultural practices [4, 5], and vegetation monitoring [6].

Recently, Satellite Image Time Series (SITS) analysis has emerged as a pivotal approach in remote sensing to support LULC classification [4, 7–9]. By utilizing the regular and high-frequency image acquisition capabilities of missions like Sentinel [10], researchers can leverage SITS data to gain comprehensive insights into the Earth's surface. This analytical approach involves examining a collection of satellite images captured over time thanks to their image acquisition frequency that offers a wealth of valuable data to study temporal changes and analyze environmental phenomena.

However, the process of LULC analysis faces several challenges, especially when dealing with multivariate SITS. Selecting an appropriate classification method that can effectively differentiate between land cover types is critical [11, 12]. The choice of the classification algorithm, feature extraction techniques, and training sample selection methods can significantly impact the accuracy and consistency of LULC classification results [12, 13].

Typically, two major classification strategies for SITS data have been proposed for LULC mapping: pixel-based classification [8] and object-based classification [14]. The pixel-based classification approach focuses on individual pixels and their multispectral values to determine land cover classes [15]. It effectively extracts relevant features from multivariate SITS data and preserves the spectral properties of pixels [4, 8]. On the other hand, the object-based classification involves an initial preprocessing step that segments the image into objects or regions. These objects represent groups of neighboring pixels and are delineated based on their spectral and spatial characteristics [14, 16]. The classification then considers the internal variability and contextual information of these objects, such as shape, neighborhood relationship, smoothness, and compactness [5].

Despite their effectiveness, these approaches face challenges in accurately characterizing land cover classes. Pixel-based classification struggles to classifying heterogeneous landscapes due to possible mixed information. Moreover, it overlooks contextual information and spatial relationships among neighboring pixels, making it difficult to differentiate similar land cover types or identify transition areas [17]. Regarding the Object-based classification, it encounters challenges when applied to heterogeneous landscapes. In such areas with diverse land cover types, objects may exhibit a range of spectral signatures, reducing the spectral separability between different classes [18]. Additionally, particular objects variety (in terms of shape and size) need to be carefully considered during classification. Small objects may lack sufficient spectral information for accurate classification, while large objects may encompass multiple land cover types, leading to potential misclassification [18].

Recent advancements in SITS analysis have introduced innovative techniques to address these challenges and improve the accuracy of land cover mapping [19]. Researchers have embraced recent advancements in computer vision and applied them to address the complex and unique nature of SITS data, including its multivariate characteristics that demands specialized techniques for processing and analysis [20, 21]. A key approach involves encoding time series signals into 2D images [22]. By capitalizing on the inherent temporal information in SITS data and incorporating contextual and spatial relationships, this approach aims to enhance the accuracy and dependability of land cover mapping [23]

Various techniques have been proposed to encode multivariate time series data as images, including Gramian Angular Summation Field (GASF), Gramian Angular Difference Field (GADF), Markov Transition Field (MTF) and Recurrence Plots (RP) [24, 25]. Each technique is applied to the univariate TS, resulting in multiple images that are concatenated and analyzed using a Convolutional Neural Network (CNN). The use of a 2D encoding approach in LULC classification has yielded notable enhancements, allowing for the identification of subtle changes that may not be easily distinguishable in the original time series signal [11]. This approach significantly facilitates the analysis of such type of data over time and improves the precision of classifiers in the LULC classification process [7].

Enlightened by previous research literature, our proposed framework aims to combine the pixel and object scales by employing a 2D encoding technique for multivariate SITS data analysis in the context of land cover mapping. The goal is to leverage the benefits of multi-scale information fusion and introduce a new pixel-object-based image classification framework that leverages 2D encoded multivariate SITS data through CNNs. By combining the pixel and object scales via 2D encoding, we aim to overcome the limitations associated with each approach while capitalizing on their individual strengths. At the pixel scale, we can capture fine-scale spectral details that characterize the land surface. This allows for precise analysis and detection of subtle variations in land cover. On the other hand, with object scale information, we consider groups of pixels that form meaningful spatial entities, allowing for the use of contextual information that can capture larger-scale patterns.

The proposed research makes contributions to the field of LULC using SITS and multivariate data. These contributions are outlined below:

- An innovative combination of multi-scale (pixel-level / object-level) information to advance the land cover mapping task from SITS data.
- The use of 2D encoding techniques to encode multivariate time series data to images, allowing the reuse of established computer vision frameworks.
- An extensive experimental evaluation. The proposed methodologies, including the multi-scale analysis and 2D encoding of encoding time series are rigorously tested and evaluated using two real-world study sites with distinct land cover characteristics.

The rest of this paper is structured as follows: Section 2 presents related works on time series encoding techniques and pixel-object classification. Section 3 provides a brief background on time series encoding techniques. In Section 4, the proposed approach is described in details with the used study sites. While Section 6 show the experimental results and discuss the findings of the comparative analysis. And Section 7 concludes the paper and suggests possible future works.

2 Related works

In recent years, considerable efforts have been devoted to enhance LULC classification leveraging multivariate SITS data. In this section, we will delve into the latest advancements in SITS-based LULC classification, encompassing approaches that leverage both spatial and contextual information, as well as those that operate independently of contextual considerations. Additionally, we will explore a broader approach for general time series classification, incorporating techniques that use 2D encoding.

2.1 SITS-based LULC classification

SITS-based LULC classification is a process that aims to categorize LULC classes using SITS data. The research activities in this field have seen significant growth, with many studies dedicated to explore and advance this area of study [8, 26, 27].

2.1.1 Spatial Context-independent SITS-based LULC Classification

A well-established family of methods for LULC classification based on SITS data only leverages pixel-level information where no spatial context is considered. Researchers have explored a wide range of approaches and methodologies in this field. Pelletier et al. [8] introduced a temporal 1D-CNN (TempCNN) approach that manages the temporal dimension for SITS classification by means of temporal convolutions. The proposed technique yielded promising results, demonstrating its potential to capture temporal dependencies and enhancing the accuracy of land cover mapping, particularly for Sentinel-2 SITS data.

Significant advancements have been made in the realm of DL models for SITS data processing such as the Lightweight Temporal Attention Encoder [28, 29] (LTAE). Based on LTAE, a novel approach that effectively captures temporal dependencies within sequential data has been proposed by Zhang et al. [29] and referred as the Global–Local Temporal Attention Encoder (GL-TAE). This approach consists of two parallel sub-modules of LTAE for extracting global temporal attention and the Lightweight Convolution (LConv) for extracting local temporal attention. By combining both global and local temporal features, GL-TAE achieves better performances on two public SITS datasets compared to methods that solely focus on global or local temporal features demonstrating the quality of incorporating hybrid global-local temporal attention features.

Indeed, while the use of time series signal has demonstrated its effectiveness across various applications, previous approaches predominantly consider only pixel-level information [26, 30]. However, in challenging and complex landscapes, it becomes evident that including spatial context is essential for an accurate analysis [4, 31].

2.1.2 Incorporating Spatial Context in SITS-based LULC Classification

The use of spatial context, from various scales or levels of detail, into SITS-based LULC classification is of paramount importance to support a better characterization of the underlying land cover classes, especially in complex and heterogeneous landscapes. Abidi et al. [4] presented a method for mapping LULC that exploits convolutional operators on the temporal dimension of SITS data. The suggested approach involves the assimilation of multivariate data from both pixel and object levels into the classification process. This is achieved through a two-branch CNN. Finally, a late fusion stage is implemented to combine the per-encoder features by means of summation.

Mohammadi et al.[32] addressed the challenge of learning more discriminative feature representations for crop mapping using SITS data considering spatial context. The researchers proposed to supervise intermediate layers of a 3D Fully Convolutional Neural Network (FCN) with two middle supervision methods: Cross-entropy loss Middle Supervision (CE-MidS) and a novel middle supervision method called Supervised Contrastive loss Middle Supervision (SupCon-MidS). These methods enhance feature discrimination and clustering throughout the network, improving its overall performance.

Censi et al. [27] proposed neural network architecture known as the attentive spatial-temporal graph CNN. This architecture is especially tailored to model, simultaneously, both the temporal and the spatial

information. The former employing 1D CNN approach and the latter by graph neural. The graph neural network module explicitly allows to exploit the spatial context around the sample to classify. The main objective of this research study is to leverage simultaneously both the information carried out by the temporal dynamics associated to the SITS data as well as the spatial context induced by the neighborhood information, with a specific focus on object-based image classification.

Derksen et al. [31] introduced a framework named Pixel Based Corner Match (PBCM) that quantitatively measures the geometric precision of classification maps. The framework utilizes corner detection and matching to analyze the impact of spatial contextual features when dense validation data is not available. The study assesses the proposed framework to classify Sentinel-2 image time series with context-dependent classes. It evaluates three spatial support shapes to test their ability to enhance classification performances while maintaining geometric precision.

2.2 General approaches for time series classification with 2D encoding techniques

Unlike traditional 1D representations, which treat time series data as a linear sequence, 2D encoding methods represent an alternative way to capture relationships between timestamp via a 2D embedding structure. These representations aim to encode time series data by leveraging adjacent information and their spatial connections. Several works have been proposed to explore the benefits of 2D encoding techniques for general time series classification in various application domain. However, it is important to highlight that there has been relatively limited research that have used the 2D encoding strategies to support LULC mapping [22].

There have been studies that explored the use of combined encoding techniques to enhance the performances of general time series classification. Jiang et al. [33] proposed a novel method that transforms univariate time series into 2D images by leveraging a combination of encoding techniques, including GAF, MTF, and RP. Different CNN models were use for classification tasks using as input the image representations obtained from the encoded TS. The results of the evaluation demonstrated the competitive performance of the proposed method, as it outperformed the benchmark models.

Velasco et al. [19] conducted a study in decision-making processes in the maritime industry. They analyzed five time series image encoding approaches, including GASF, GADF, MTF, RP, and Markov Transition Matrix (MTM). Their analysis demonstrated the ability of these approaches to identify fault patterns that may not be discernible when considering the original time series data alone.

Lee et al. [34] introduced an image encoding scheme for dealing with practical Received Signal Strength Indicator (RSSI) data containing artificial Gaussian noise. The scheme considers image encoding to find appropriate features from RSSIs while maintaining the characteristics of serialized time series data using a transition of image from time series data to 2D data using MTF method. The results show a significant enhancement of about 46.2% compared to methods that solely rely on CNNs without encoding.

Considering the analysis of remote sensing satellite image time series data, Menini et al. [35] highlighted the significance of applying univariate time series representations in the context of classifying eucalyptus regions. In the proposed framework, images represented via the RP method were used to illustrate the time series associated to the image pixels. Experimental results indicated the effectiveness of the proposed method compared to methods that simply depend on the use of the original time series data.

Abidi et al. [7] introduced a novel framework that exploits the transformation of multivariate Sentinel-2 SITS data from its native 1D signal form into the realm of 2D images, employing a range of encoding techniques like GASF, GADF, MTF, and RP. These methods are combined into a single image fed into a CNN for the final classification.

The exploration of 2D encoding techniques for LULC classification remains relatively limited, leaving a gap in our understanding of their potential for this task [35]. Similarly, the integration of pixel and object-level information in this context is an area that has seen sparse investigation [4, 17]. This approach, however, holds particular promise for LULC classification in regions characterized by complex landscapes [7]. In this context, our proposed framework seeks to bridge these gaps by combining multi-scale pixel-object level information and 2D encoding techniques for LULC classification with the aims to leverage well-established classification frameworks from the computer vision community. This research presents a new approach for LULC classification using 2D image encoding, highlighting the need to leverage multiple sources of information for accurate and effective LULC characterization. The proposed method offers a a novel perspective for LULC classification and aligns with the growing trend of utilizing multi-scale information.

3 Background

In this section we provide a brief mathematical foundation of the encoding methods we will leverage for our framework, including an explanation of how the images are generated via these procedures.

3.1 Gramian Angular Field (GAF)

A Gramian Angular Field (GAF) is a 2D representation used to capture orientation information in an image, particularly for analyzing texture and visual features [24]. To create a GAF, we start with a time series (TS) denoted as $(X = x_1, x_2, ..., x_n)$ with n real-valued observations. The first step involves normalizing this time series so that all values fall within the range $[0, 1]$, resulting in $\overline{x}$.

This normalization process is based on a mathematical operation that uses a parameter known as $\emptyset$. $\emptyset$ corresponds to the angle, akin to a timestamp in polar coordinates. When $\overline{x}_i$ lies between -1 and 1, $\emptyset$ represents the inverse cosine of $\overline{x}_i$. In simpler terms, it's a way to convert the values of $\overline{x}_i$ into angles within the polar coordinate plane denoted as N. The radius is represented by 'r,' and 'x' is the variable.

The Gramian Angular Field [24] (GAF) can be created using two different methods: the Gramian Angular Summation Field (GASF), that involves in essence trigonometric calculations that consider the sum of angles ($\emptyset_i$ and $\emptyset_j$) to capture temporal dynamics and and the Gramian Angular Difference Field (GADF) which is similar to obtaining the GADF representation. However, it is determined by calculating the sine of the difference of angles for each point.

Figure 1 shows a graphical representation of the GADF and GASF techniques. For this particular example, the time series under consideration is the sequence of cosine function values at 5 000 equally-spaced points within the interval $[0, 4\pi]$ X represents a sequence of time series data in a dataset.

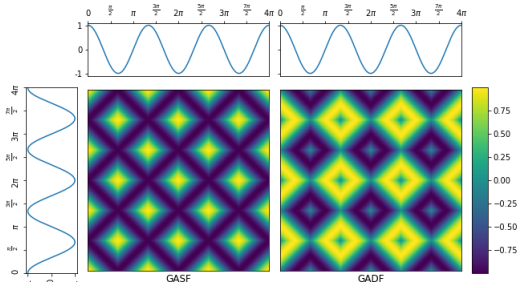

Fig. 1 A visualization of the encoding schemes of GADF and GASF techniques. The colors closer to yellow in the Figure represent high values, while colors closer to dark blue and blue represent lower values.

3.2 Markov Transition Field (MTF)

MTF (Markov Transition Field) is a technique commonly employed in signal processing and image analysis to capture spatial and temporal dependencies within a sequence of data [24]. It is basically constructed as a matrix where each element w_{ij} represents the frequency with which a point in quantile q_j follows a point in quantile q_i. This matrix visually represents the transition patterns between different quantiles without delving into mathematical equations.

Figure 2 shows a graphical representation of the technique. In this case, the time series being studied is the sequence of cosine function values at 5 000 equally-spaced points within the interval $[0, 4\pi]$.The representation demonstrates that the cosine function is periodic with a period π.

3.3 Recurrence plot (RP)

The RP is a visual representation of the distances between trajectories that are extracted from the original time series [25]. The RP process starts by generating a 2D phase space trajectory from the time series. Then, the R-matrix is calculated based on the proximity of the states in the phase space. The resulting R-matrix consists of only 0s and 1s due to a threshold parameter, without going into mathematical detail.

4 Methodology

This section introduces Orthrus, a novel framework for land use land cover mapping that leverages multi-scale information and 2D encoding techniques to cope with satellite image time series information. The

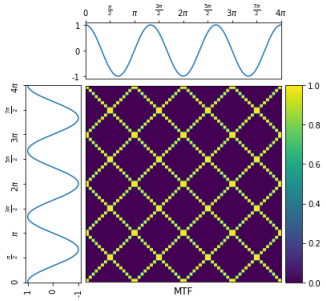

Fig. 2 A visualization of the encoding schemes of MTF. The colors closer to yellow in the figure represent high values, while colors closer to dark blue and blue represent lower values.

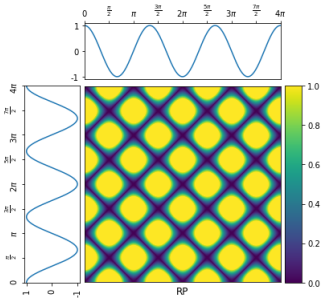

Fig. 3 A visualization of the encoding schemes of the RP technique. The colors closer to yellow in the figure represent high values, while colors closer to dark blue and blue represent lower values.

approach is depicted in Figure 4, which sketches the flowchart of the process. The proposed framework consists of two stages: the first stage performs 2D Encoding using various techniques for each of the multi-scale information, while the second stage uses a dual branch Convolutional Neural Network to classify the encoded time series data. This framework effectively captures both the pixel-level and object-level features of multivariate time series data, improving the accuracy of classification tasks.

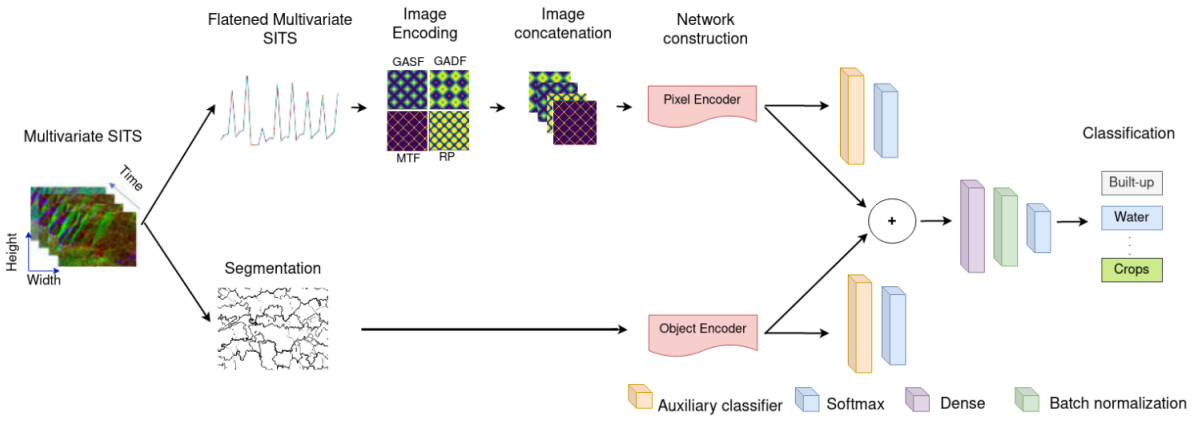

Fig. 4 Flowchart of the proposed method: 2D encoded multivariate SITS for multi-scale land cover mapping based on the fusion of pixel-level and object-level time series

.

4.1 Multivariate SITS Encoding

The 2D representation of multivariate SITS is achieved through four encoding techniques : GADF, GASF, MTF, and RP, as illustrated in Figure 5. These techniques aim to convert the characteristics of the SITS data into an image format that facilitates their subsequent analysis.

The encoding phase begins by flattening the multivariate SITS all at once. This process is carried out on the complete SITS, resulting in a flattened representation. The aim of flattening is to convert the SITS

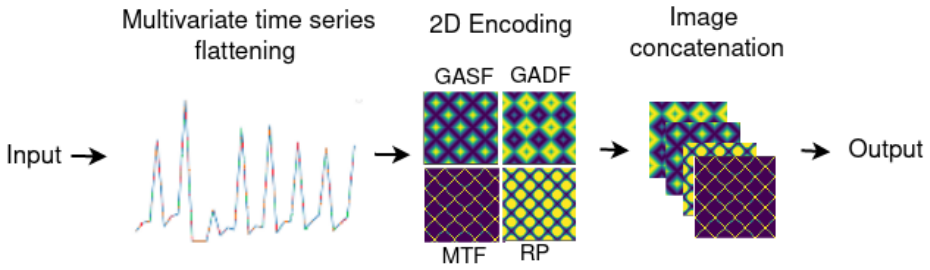

Fig. 5 Graphical representation of the proposed multivariate SITS encoding process.

into a two-dimensional representation, facilitating easier processing and analysis. This step ensures that all variables and observations are treated equally, regardless of the original dimensionality of the SITS. By flattening the SITS, it becomes easier to extract relevant information, identify patterns, and perform subsequent analyses. The flattening process is conducted on the complete multivariate SITS simultaneously, rather than encoding each band value independently. In other words, the values of the four bands composing the original S2 multivariate SITS for each time instant t_i.

This approach considers the relationship between the bands and provides more comprehensive information about the spectral properties of the scene. By analyzing all the bands together, it is possible to extract more meaningful and accurate information about the underlying dynamics. Considering a multivariate SITS with dimensions (pixels, time, number of bands), denoted as X. In order to flatten the time and number of bands dimensions and preserve the radiometric information, we can reshape the time series X into a 2D matrix denoted as X_{flat}, where the rows correspond to individual pixels and each row contains the concatenated time and band information for that pixel. The resulting flattened SITS would have dimensions (pixels, time $\times$ number of bands). Through this concatenation process, a condensed portrayal of the complete multivariate SITS data is generated at each time instance, maintaining the temporal coherence of the data. The combination of the temporal and spectral information achieved by the flattening operation allows to leverage the temporal context for each pixel across various time steps [7].

The encoding process is subsequently applied to the flattened images using the techniques introduced in section 3. This 2D encoding offers context-aware features that can substantially enhance visual recognition, once 2D encoding has been performed, for the downstream classification task.

We obtained four individual matrices as a result, each with dimensions similar to a matrix. These outputs are then combined to construct a new SITS representation, encompassing information from all four encoding techniques. The purpose of this concatenation is to leverage the complementarity of using the four encoding techniques, as each technique captures different aspects and features of the multivariate SITS data, considering that:

- The GADF representation has a sinusoidal structure that makes it ideal for measuring regular seasonal patterns of vegetation, particularly growth. [24]
- GASF is able to capture quicker fluctuations that exist in complex patterns with uneven structures. This makes it for evaluating the probability of a woodland region being modified into an urban area [24].
- MTF representations are used to encode dynamic information regarding state transition probabilities. These probabilities represent the likelihood of a system transitioning from one state to another over time [24].
- RP representations can provide valuable insights into the recurrence patterns of multivariate SITS. By analyzing these representations, it is possible to identify areas within a city that experience regular annual growth as well as those that exhibit seasonal deforestation. [25]

4.2 Dual Branch multi-scale information fusion

Here, we will describe the Dual Branch multi-scale information fusion approach, detailing with the feature extraction and fusion technique as well as the model training process.

4.2.1 Feature Extraction and Fusion Technique

We use CNNs as feature extractors for both pixel-level and object-level encoders. CNNs have achieved impressive results in computer vision tasks by learning complex patterns and hierarchical representations from images [36]. Specifically, we utilize a CNN-1D architecture as the feature extractor for the object

branch[7], while for the pixel branch, we employ the ResNet-50 architecture as illustrated in Figure 6. The choice of ResNet-50 is motivated by its widespread acceptance and extensive usage in the field.

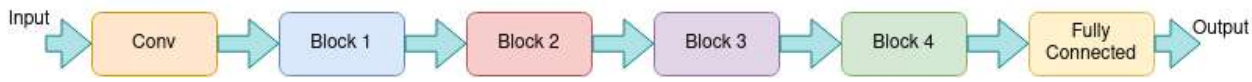

Fig. 6 Graphical representation of the different blocks of the used ResNet-50 classifier.

The feature extraction process involves inputting the 2D encoded data, as obtained from the previously outlined procedure, into the Dual Branch architecture illustrated in Figure. 4. The pixel-level encoder captures fine-grained information from individual pixels, such as spectral characteristics. Concurrently, the object-level encoder extracts higher-level spatial and contextual features by considering the relationships between neighboring pixels into the object. Once the feature extraction process is complete, the extracted features from both encoders are used to perform the final classification.

At this stage, the late fusion technique comes into the picture. We merge the features obtained from the pixel-level and object-level encoders using the element-wise sum operation. By combining the pixel-level and object-level features, the network can exploit a more spatially-aware representation, facilitating the understanding of complex spatial patterns and temporal changes over time [7]. Table 1 provides more details about the architecture of Orthrus.

Orthrus	
Pixel Encoder	ResNet-50
Object Encoder	CNN-1D
$Sum([feat_p, feat_o])$	
Dense (256, activation = ReLu)	
BatchNormalization()	
SoftMax	

Table 1 Architectural details of Orthrus

4.2.2 Model Training Process

To enhance the discriminative power of the information at different scales, we have leveraged auxiliary classifiers, a widely adopted technique in multi-sensor data fusion for Earth observation data analysis [37, 38]. These classifiers serve as output layers for each encoder, aiming to extract complementary and discriminative information from both the pixel and object scales. During the classification process, the auxiliary classifiers are jointly trained alongside the fused classifier to identify the land cover class associated with each pixel. As demonstrated in previous research studies, this mechanism has the ability to enhance the discriminative capacity of the extracted representations improving the final model's accuracy and effectiveness. The loss function related to Orthrus can be formalized as follows:

$$\begin{aligned} L = & CE(Y, CL(feat)) \\ & + \alpha(CE(Y, OUT_p(feat_p)) + CE(Y, OUT_o(feat_o))) \end{aligned} \tag{1}$$

Giving CE as the standard cross-entropy loss, Y the ground truth land cover class, $CL(.)$ is a neural network classifier composed of two fully connected layers and $feat$ represents the fused representation related either to the retrieved feature from pixel or object encoders (respectively $feat_p$ and $feat_o$). The hyper-parameter *alpha* modulates the effect of the per-encoder importance in the learning process. OUT_p (resp. OUT_o) is a fully connected (output) layer, with as many neurons as the number of the classes, associated with a SoftMax activation that performs classification exploiting the pixel-level $feat_p$ (resp. object-level $feat_o$) as input.

5 Satellite image time series and ground truth data

5.1 Satellite Image Time Series data

The method assessment is conducted on two datasets covering distinct geographic regions: *Dordogne*, a sub-region of the Dordogne department situated in the southwestern part of France, and *Reunion-Island*, a French overseas department located in the Indian Ocean. Specifically, For the *Dordogne* dataset, a total of 21 Sentinel-2[1] images were utilized. These images were acquired between January and December 2016, covering a spatial extent of 5 578 × 5 396 pixels, which corresponds to 3 010 Km^2. On the other hand, the *Reunion-Island* dataset included 23 Sentinel-2 images obtained between March and December 2017, covering an area of 6 656 × 5 913 pixels, corresponding to 3 010 Km^2. All the images used in the study were obtained from Theia[2] pole. The satellite images are preprocessed in surface reflectance using the MACCS-ATCOR Joint Algorithm [39] designed by the National Centre for Space Studies. Only four radiometric bands at 10m spatial resolution were considered: B2, B3, B4 and B8 correspond to Blue, Green, Red and Near-Infrared. To fill gaps induced by cloud phenomenon, a linear multi-temporal interpolation was performed across each band. More precisely, for invalid (cloudy) pixels, linear interpolation is performed by considering precedent and successive cloud-free acquisition dates following the operational schema proposed in [40]. This approach allows for estimating surface reflectance values at any specific date, thus providing satellite image time series that are regularly sampled and temporally aligned.

5.2 Ground Truth data

The Ground Truth (GT) data for the *Dordogne* study site was obtained through the *Registre Parcellaire Graphique*[3] (RPG) for 2014, and the Topographic database (BD-TOPO) from the French National Geographic Institute (IGN). Figure 7 and Figure 8 illustrate the two study sites, along with the related GT polygons. Concerning the *Reunion-Island* study site, it was formed using a variety of sources, (i) the RPG reference data for 2014 and (ii) a high spatial resolution (VHSR) SPOT 6/7 image visually interpreted by a domain specialist with territorial expertise to identify natural and urban areas.

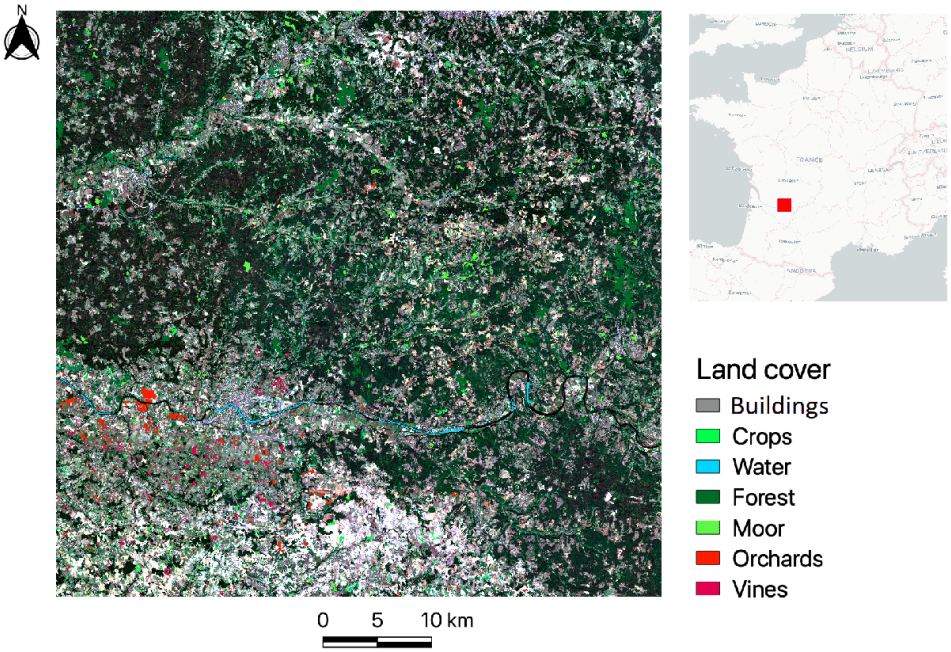

Fig. 7 Location of the *Dordogne* study site; The RGB composite is a Sentinel-2 image.

The GT data for both study sites is stored in GIS vector file format that includes a set of polygons, each with a unique land cover class label. These vector files were then transformed to raster format at the spatial resolution of 10m. Table 2 and Table 3 provide the GT information for the *Dordogne* and the *Reunion-Island* study sites.

[1] https://sentinel.esa.int/web/sentinel/missions/sentinel-2
[2] http://theia.cnes.fr
[3] RPG represents a part of the European Land Parcel Identification System (LPIS), which is handled by the French Agency for Services and Payment

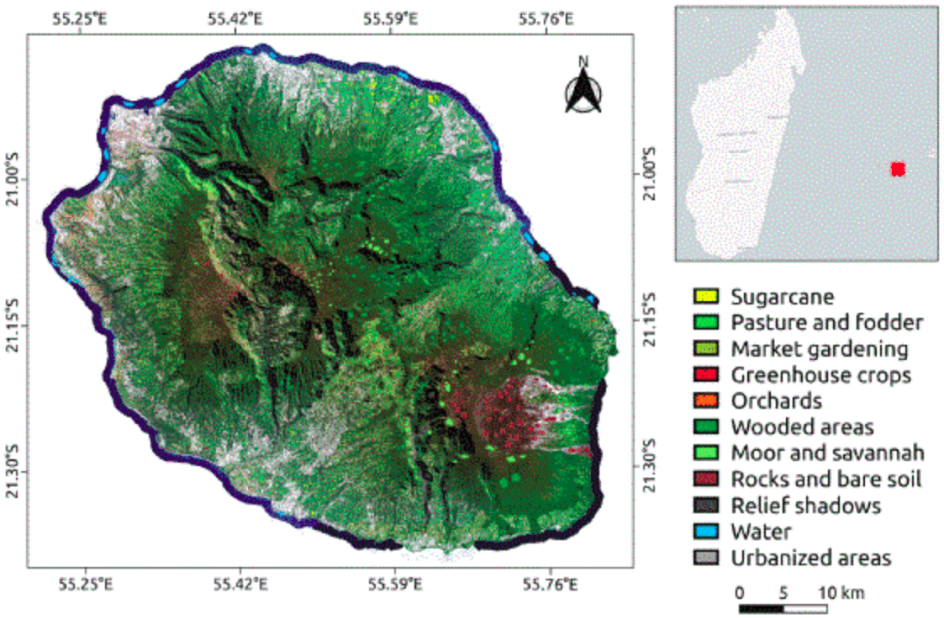

Fig. 8 Location of the *Reunion-Island* study site; The RGB composite is a SPOT 6/7 image upscaled at 10-m of spatial resolution.

Study Site	Class	Label	# Objects
Dordogne	0	Built up	849
	1	Crops	1 554
	2	Water	1 217
	3	Forest	2 703
	4	Moor	1 108
	5	Orchards	1 099
	6	Vines	1 389
		Total	9 919

Table 2 Characteristics of the *Dordogne* site GT

Study Site	Class	Labels	# Objects
Reunion-Island	0	Sugar cane	2 190
	1	Pasture and fodder	1 565
	2	Market gardening	1 284
	3	Greenhouse crops	339
	4	Orchards	1 563
	5	Wooded areas	2 741
	6	Moor and Savannah	2 169
	7	Rocks and bare soil	1 687
	8	Relief shadows	560
	9	Water	873
	10	Urbanized areas	1 540
		Total	16 511

Table 3 Characteristics of the *Reunion-Island* site GT

5.3 Data Preprocessing

The pixel values in each band of the dataset were normalized in the range [0,1]. This normalization process ensures that the pixel values are scaled proportionally, facilitating consistent comparison and interpretation across different spectral bands or images. This is also done in order to avoid numerical issues related to the use of neural network based approaches since theses approaches may be sensible to input values spanning wide ranges. The segmentation preprocessing step was carried out using the Simple Linear Iterative Clustering [41] method (SLIC). This technique is extremely fast and can be used to process large scale remote sensing images. It strives to produce segments that are similar in size and with reasonably compact shape. Several scale parameters were tested and evaluated during the preliminary experiments to assess their impact on

superpixel size and compactness. Therefore, a scale value of 3 was chosen through various trial and error processes and visual inspections, also following the findings reported in [4], to ensure that the final segments closely matched the land cover units in the study areas. Lastly, three descriptors are derived per object, for each band at each time stamp: mean, median, and standard deviation. The mean is the average value of all the pixels in a given object. The median is the 50^{th} percentile value of all the pixels of a specific band. While the standard deviation is a measure of how much the pixels deviate from the mean value. These features or descriptors are commonly used in image processing and can provide valuable information regarding the radiometric information of an object in an image. Furthermore, combining these features can provide a more complete and robust set of features that can support the object-based image classification process. An example of the SLIC segmentation result on a Sentinel-2 satellite image is shown in Figure 9.

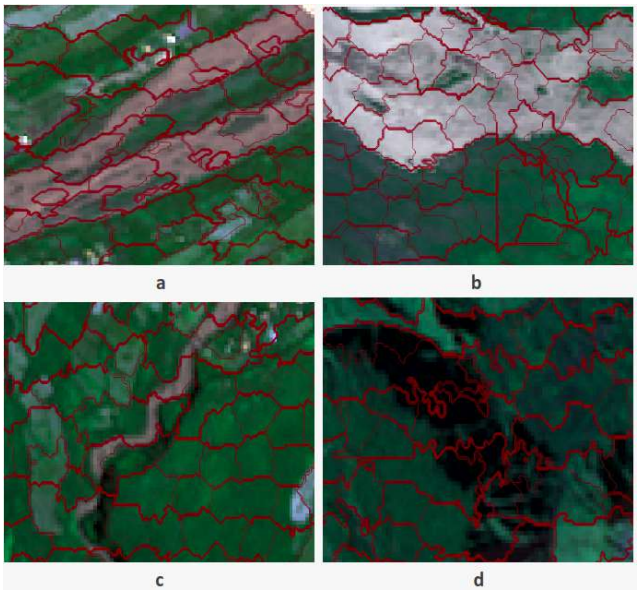

Fig. 9 SLIC segmentation of a Sentinel-2 multivariate SITS on heterogeneous class types. Images A and B represent homogeneous segment groups while images C and D represent an under/over-segmentation respectively that imply segments which not able to identify the texture due to the extensive spectral fluctuation.

The encoding was performed on the flattened multivariate SITS signal. For that, the data shape for each satellite time series sample was transformed from (23,4) to (92) for the *Dordogne* study site (resp. from (21,4) to (84) for the *Reunion-Island*), leading to images with dimensions (92,92) for the *Dordogne* study site (resp. (84,84) for the *Reunion-Island*). The resulting 2D encoded multivariate SITS were then resized to 32×32 images. This downsampling stage was done in order to reduce computational complexity, facilitating efficient processing while preserving essential information.

6 Experimental evaluation

6.1 Experimental Settings

Each dataset has been divided into three partitions: training, test, and validation, with object proportions of 50%, 30%, and 20%, respectively. This split ratio was chosen to adhere to common practices in recent remote sensing studies[18, 27, 37], aiming to create a balance between the amount of data used for training, testing, and validation. To prevent a possible spatial bias in the evaluation procedure, we enforce that pixels related to the same GT polygon be allocated solely to one of the data partitions (training, validation, or test). Training data are used to train the model, whereas validation data are used, along the iterative process, to select the model parameters that should generalize the best. Finally, the selected model, the one that performed the best on the validation set, was used to perform prediction on the test set.

In our proposed approach, we harness the power of the ResNet-50 model, renowned for its use of residual connections. These connections allow gradients to flow back to previous layers during training, improving the learning capabilities of the model. The architecture consists of five stages, with each stage containing a convolutional and an identity block. Each block consists of three convolutional layers. The five stages are then connected with 2D global average pooling. A fully connected layer further processes these features,

resulting in a 50-layer residual neural network. The number of filters grows as the blocks progress (128, 128, 256, 256, 512), while the kernel size varies from [1,1] to [3,3].

The classification performances were evaluated using the Accuracy (global accuracy) and F1 Score metrics. Considering that the model performances may vary depending on the data split, all metrics were averaged over five random train/validation/test splits using the previously described strategy. GASF, GADF, MTF have been implemented as described in [24], whereas the RP representation was implemented as defined in [25]. Experiments were carried out on a workstation equipped with an Intel(R) Xeon(R) CPU E5-2667 v4 @ 3.20 GHz with 256 GB of RAM and four TITAN X GPU. All neural network-based algorithms were implemented with the Python TensorFlow library and trained using the ADAM[42] optimizer over 100 epochs with a learning rate of 10^{-4}. For the machine learning algorithms, we utilized those available through the Python Scikit-learn package [43]. A batch size of 32 was utilized. Table 4 provides a summary of the experimental settings.

Table 4 Experimental settings details

Experimental Settings	Details
Data Partitioning	Training: 50% , Testing: 30% and Validation: 20%
Model evaluation	Selected model performance assessed on the test set.
Classification Metrics	Accuracy (global accuracy) and F1 Score.
Data splits	Metrics averaged over five random train/validation/test splits.
Implementation Details	- GASF, GADF, MTF implemented as described in [24]. - RP representation implemented as defined in [25].
Neural network algorithms	Python TensorFlow and PyTorch libraries ADAM optimizer over 100 epochs learning rate of 10^{-4} Batch size of 32

6.2 Competing approaches

To assess the behavior of our method, Orthrus, we consider several state-of-the-art approaches as competitors. These competitors were selected from both the remote sensing field, with a particular emphasis on multi-scale approaches, and the general literature related to multivariate time series classification. Here the list of the adopted competitors:

- Two branch One dimensional CNN (TwoBCNN), a DL approach proposed as a recent method for LULC mapping that integrates pixel- and object-level information in the form of multivariate SITS data [4]. This method combines multi-scale multi-temporal remote sensing data without any additional transformation on the original time series signal. This competitor is of particular interest in order to evaluate the behaviour of Orthrus on handling multi-scale remote sensing data, which is central to our study objectives.
- Temporal Convolutional Neural Network (TempCNN) presented in [8]. This technique is built on a one dimensional convolutional neural network, where the convolutional operators are employed across the time dimension. This allows for an examination of how our method performs relative to a method that focuses solely on the temporal aspects of the data, providing insights into the effectiveness of our approach in leveraging both spatial and temporal information.
- InceptionTime, this method has been introduced in the field of multivariate time series classification [44, 45]. It is based on the Inception-v4 architecture and it utilizes a combination of deep CNN models for time series classification. We propose two different version of the InceptionTime approach. A first version, referred as InceptionTime$_{EF}$, adopts an early fusion strategy w.r.t. the way the multi-scale data are managed. More precisely, we combine both pixel-level and object-level information together before feeding them to the model. The second version, referred as $TwoB - InceptionTime_{LF}$, deploys a late fusion strategy, with a dual-branch architecture, similarly to what proposed in [4], in order to manage the multi-scale information with a per-scale encoder. Despite differences in the specific techniques used, both methods aim to leverage CNN architectures for processing multivariate time series data, although with varying fusion strategies. This point allows for a meaningful comparison in terms of performance metrics and effectiveness in handling complex data structures such as multi-scale information.
- MultiRocket, a recent technique for multivariate time series classification that combines random convolutions with multiple pooling and transformation operators to increase the variety of the generated features [46]. Then, these features are integrated through a linear classifier. Similarly to what proposed

for InceptionTime, also for the MultiRocket competitor we propose two different variants. The first one, $MultiRocket_{EF}$, exploits an early fusion strategy to cope with multi-scale information, following the procedure previously described. The second one, denote as $MultiRocket_{LF}$, adopts a late fusion strategy where a classifier is learnt per source (pixel-level or object-level) and then the two classification results (at pixel and object levels) are combined via a majority voting mechanism. While MultiRocket focuses on generating diverse features through random convolutions and pooling operations, our method encodes temporal information into 2D images, offering a different approach to addressing the challenge of integrating temporal and multi-scale information for the downstream classification task.

- ConvTran is the current state-of-the-art for multivariate time series classification [47]. This method is based on the Transformer [48] backbone and it integrates both absolute and relative position encoding strategies, namely time Absolute Position Encoding (tAPE) and computationally efficient Relative Position Encoding (eRPE), respectively. These encoding techniques are specifically designed for time series data and they boost the model's ability to capture long-short temporal relationships. Additionally, ConvTran combines convolution-based input encoding with transformers, resulting in enhanced position and data embedding for time series analysis. Furthermore, we employed ConvTran for classifying both pixel-based and object-based data using both early and late fusion strategies (respectively $ConvTran_{EF}$ and $ConvTran_{LF}$).

Additionally, we conducted an ablation study to comprehensively evaluate the effectiveness of our method with the aim to asses the influences of the different object-level information we are considering for the downstream classification. To this end, we employed two distinct variants, more precisely, Orthrus_{M} indicates the model that only uses the mean descriptor as feature while, Orthrus_{MM} indicates the model that exploits both the mean and the median descriptor as features. This ablation study allows to assess the influence of the different object-level features we have considered on the performance of our method.

6.3 Results

According to the experimental design we have previously introduced, in the below sections we report both quantitative and qualitative results in order to characterize and assess the behaviour of Orthrus.

6.3.1 Quantitative Results

We report and analyze the results obtained by all the competing approaches, with respect to both overall and per-class performances. Table 5 and Table 6 illustrate the average quantitative performances of the different competitors on the *Dordogne* and *Reunion-Island* study sites, respectively.

Method	Accuracy	F1-score
$\text{InceptionTime}_{EF}$	86.09 ± 2.42	86.23 ± 2.25
MultiRocket_{EF}	85.45 ± 2.45	84.78 ± 2.16
Resnet50_{EF}	85.28 ± 2.76	85.86 ± 2.84
TempCNN_{EF}	86.74 ± 2.38	86.54 ± 2.33
ConvTran_{EF}	87.67 ± 1.89	87.48 ± 1.85
TwoB-InceptionTime	87.66 ±1.82	87.85 ± 1.89
TwoBCNN	86.44 ± 2.34	86.78 ± 2.21
$\text{MultiRocket}_{MajVot}$	83.97 ± 3.31	86.78 ± 2.28
ConvTran_{LF}	88.13 ± 1.34	89.52 ± 1.56
Orthrus_{M}	89.23 ± 1.78	88.89 ± 1.67
Orthrus_{MM}	91.07 ± 1.69	91.33 ± 1.58
Orthrus	**91.84** ± 1.58	**92.36** ± 1.49

Table 5 Accuracy and F-Measure performances of all the competing approaches on the *Dordogne* study site. Results averaged over five splits.

When considering state-of-the-art competitors, it's worth noting that those employing an early fusion strategy, including $InceptionTime_{EF}$, $MultiRocket_{EF}$, $Resnet50_{EF}$, and $TempCNN_{EF}$, systematically achieve lower performance results compared to their late fusion counterparts. This difference in performance can be attributed to the ability of late fusion approach to exploit more effectively the diverse information embedded in multi-scale data, resulting in improved classification accuracy.

Method	Accuracy	F1-score
InceptionTime$_{EF}$	85.96 ± 2.67	86.22 ± 2.47
MultiRocket$_{EF}$	83.75 ± 2.75	79.66 ± 2.33
Resnet50$_{EF}$	84.88 ± 2.85	85.14 ± 2.71
TempCNN$_{EF}$	86.27 ± 2.07	85.88 ± 2.11
ConvTran$_{EF}$	87.43 ± 1.78	87.19 ± 1.84
TwoB-InceptionTime$_{LF}$	87.17 ± 2.40	87.21 ± 2.47
TwoBCNN	87.12 ± 1.91	87.67 ± 1.98
MultiRocket$_{LF}$	86.45 ± 2.03	88.33 ± 2.46
ConvTran$_{LF}$	88.05 ± 1.54	88.94 ± 1.38
Orthrus$_{M}$	88.35 ± 1.77	88.56 ± 1.68
Orthrus$_{MM}$	90.33 ± 1.53	90.82 ± 1.61
Orthrus	**91.35** ± 1.68	**91.84** ± 1.84

Table 6 Accuracy and F-Measure performances of all the competing approaches on the *Reunion-Island* study site. Results averaged over five splits.

Additionally, it's worth mentioning that competitors based on neural network architectures (TwoB-InceptionTime and TwoBCNN) outperformed the state of the art MultiRocket approach from general multivariate time series classification that only reaches an accuracy of 86.45%. This finding underscores the importance of considering specialized architectures tailored to the specific characteristics of the SITS data, rather than relying solely on generic approaches

When the Orthrus ablations are analyzed, Orthrus$M$ and Orthrus$(Mean, Median)$, it becomes evident that Orthrus$(Mean, Median)$, that combines information from both mean and median features, exhibits better performances. This improvement is of particular interest as it emphasizes the synergistic effects of integrating different statistical measures, enhancing the ability of the model to capture the nuances present in the data.

However, the proposed method, which goes a step further by including standard deviation as an additional feature, stands out as the best performing approach with respect to all the considered evaluation metrics. This behavior underscores the importance of considering complementary feature sets that capture both central tendency and variability in the data, thereby providing a more robust representation of the object-level spatial context information. Recognizing the value of such comprehensive feature sets enhances our understanding of the practical implications of feature engineering in optimizing model performance for land cover mapping applications.

Figure 10 and Figure 11 report, respectively, the per class F-Measure obtained by the different competing methods on the *Dordogne* and the *Reunion-Island* study sites. As a general remark, we can note that Orthrus demonstrates a clear improvement in the classification precision of all land cover classes compared to its ablation Orthrus$_{Mean,Median}$, as illustrated in Figure 10. In addition, a reduction in confusion among the `1.Water` and `6.Crops` land cover classes is obtained by the proposed Orthrus. The use of the standard deviation statistic to describe object-level information seems to be a key factor in achieving these results. This is probably due to the fact that this feature provides more information about the distribution of pixel values within an object, which leads to additional knowledge in order to better discriminate between the different land cover classes. Considering these improvements, it is important to highlight that the enhanced precision in classifying all land cover classes, especially the reduction in confusion between the `1.Water` and `6.Crops` classes, may have practical implications in downstream applications related to a more sustainable management of natural resources and agricultural areas.

For an in-depth investigation of models behaviour, we also examine the confusion matrices of each method regarding both study sites.

Concerning the *Dordogne* site, Figure 12 underlines an improvement in the precision of classification results using the proposed method, which resulted in an increase of more than 30% in the classification accuracy of the `0.Built-up` land cover class compared to InceptionTime$_{EF}$. More in detail, this improvement has led to a reduction in the confusion between `0.Built-up` and `6.Vines` land cover classes that can exhibit similar spatial patterns. These results clearly highlight the ability of our approach to effectively leverage multi-scale information to discriminate between land cover classes characterized by subtle differences in visual appearance and spatial patterns.

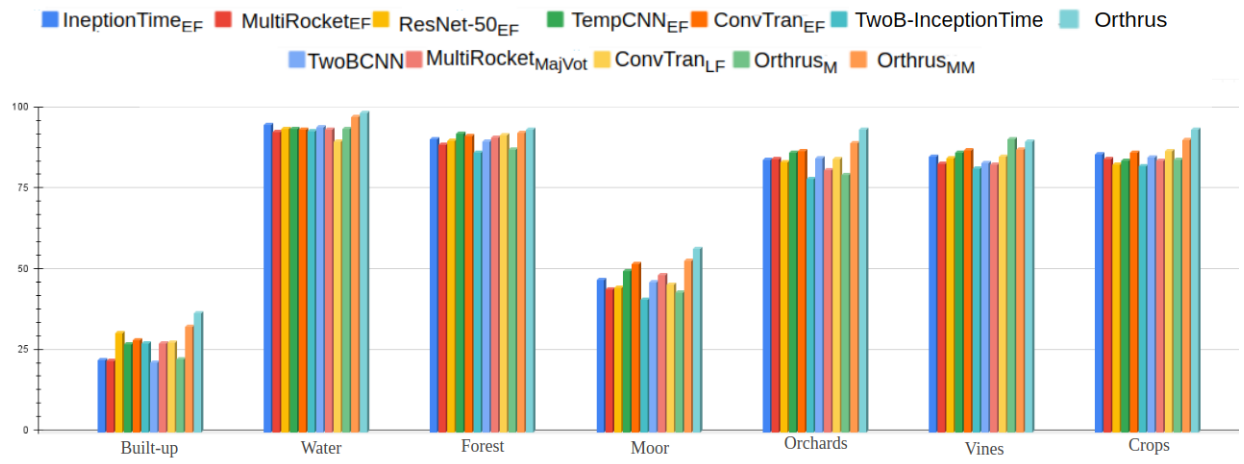

Fig. 10 Per class F-Measure performances of the different competing methods considering the *Dordogne* study site.

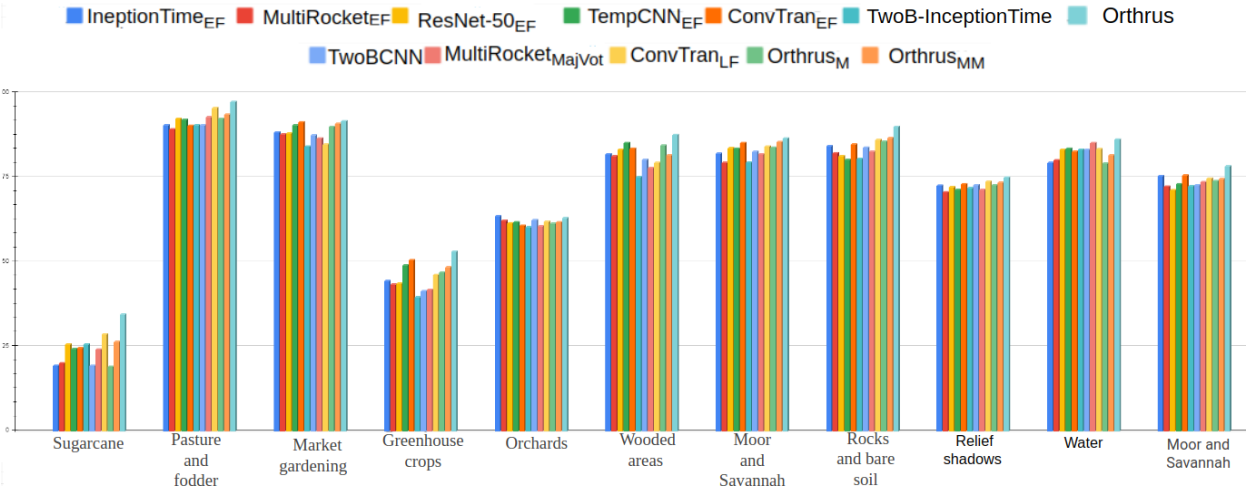

Fig. 11 Per class F-Measure performances of the different competing methods considering the *Reunion-Island* study site.

Concerning the *Reunion-Island* (Figure 13), all the methods exhibit comparable behavior. This is especially clear in the confusions between `3.Greenhouse crops` and `10.Urbanized areas` land cover classes. This confusion arise due to the high visual similarity the two land cover classes shared. In fact, they may include features such as asphalt surfaces in the urbanized areas and impermeable materials in greenhouse construction, which can make challenging to differentiate between them. However, these confusions are notably reduced by MultiRocketLF InceptionTime$_{EF}$ and ConvTran$_{LF}$ that highlight the effectiveness of multi-scale approaches. Finally, Orthrus exhibit a more clear diagonal structure with respect to all the other competing approaches.

6.3.2 Qualitative Results

Here, we adopt the Uniform Manifold Approximation and Projection [49] (UMAP) technique to visually explore the representation learned by MultiRocket$_{EF}$, Resnet50$_{EF}$, TwoBCNN, Orthrus$_{M}$, Orthrus$_{MM}$ and Orthrus, after randomly selecting 300 samples per class from the test set on the `Dordogne` study site. Figure 14 depicts the two dimensional representation derived by UMAP. We observe a clear enhancement in the differentiation of land cover classes as object-level information is integrated. The figure highlights that early fusion strategies, (a) MultiRocket$_{EF}$ and (b) Resnet50$_{EF}$, show lower discriminatory power compared to their late fusion counterparts and Orthrus. This is due to the inherent differences in fusion techniques. Early fusion combines pixel and object-level information at the initial stages of the classification process, potentially leading to a loss of spatial context and nuanced features present in the object-level information. In contrast, Orthrus integrates pixel and object-level information after separate processing, preserving the distinctive characteristics of each information. Moreover, between Orthrus$_{M}$, Orthrus$_{MM}$ and Orthrus,

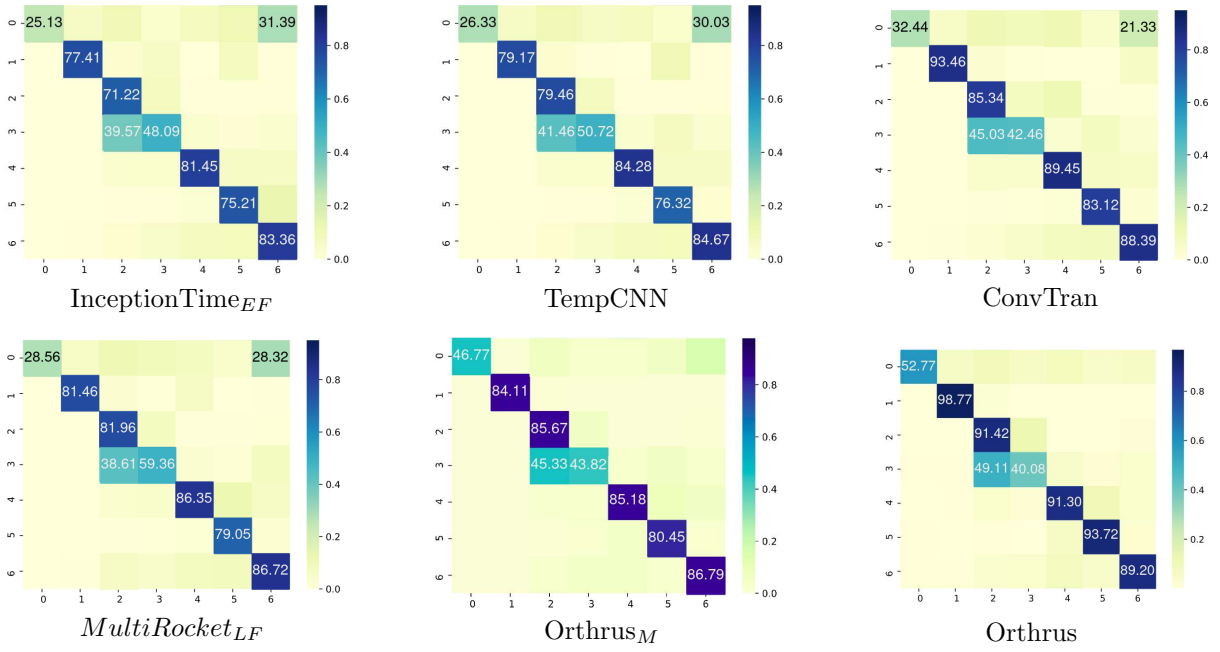

Fig. 12 Confusion matrices of the classification results of several competing approaches on the *Dordogne* study site.

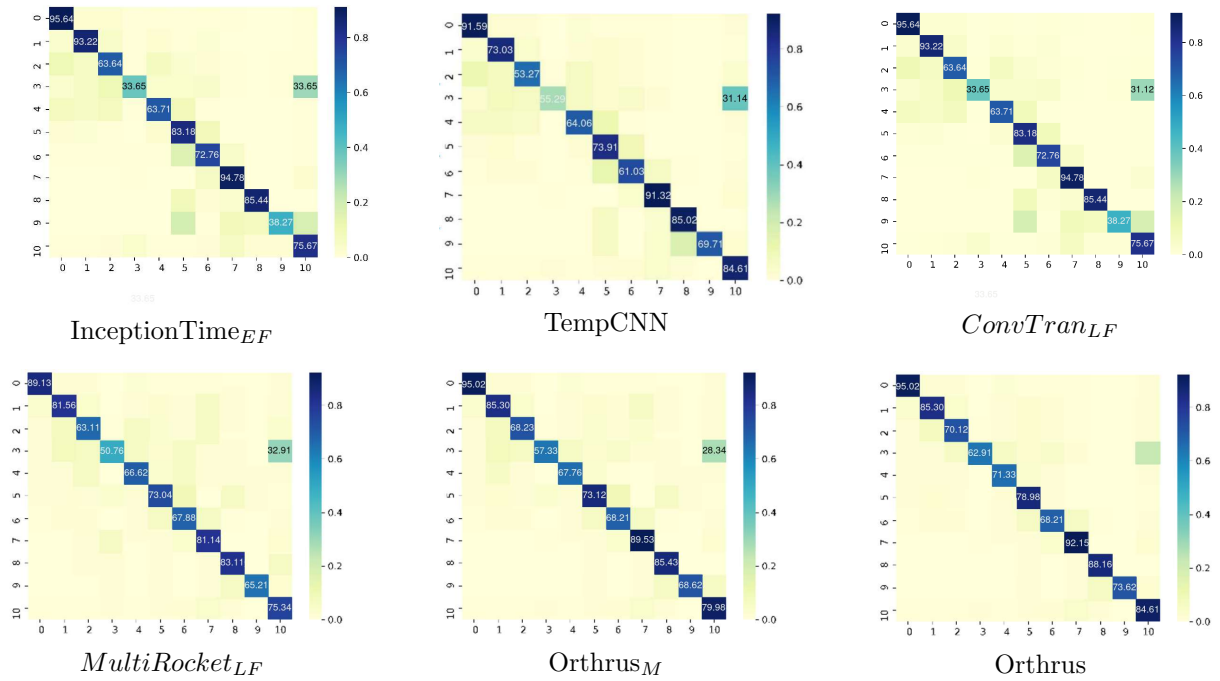

Fig. 13 Confusion matrices of the classification results of several competing approaches on the *Reunion-Island* study site.

the latter, which utilizes mean, median, and standard deviation information at the object level, provides better visual results. This is probably due to the inclusion of median and standard deviation statistics that provide additional insights and variability in the object-level representations, leading to a more clear cluster structure compared to the ones exhibited by the competitors. The qualitative findings well align with the quantitative results detailed in Table 5.

Figure 15 and Figure 16 illustrate land cover map extracts generated by Orthrus, OrthrusMM, MultiRocketLF, and ResNet-50, for the *Dordogne* and *Reunion-Island* sites, respectively.

As general remark for both study sites, we can note that Orthrus was able to reduce noise and produce smoother classification maps, with less fragmentation and better continuity of land cover classes. Regarding the *Reunion-Island*, as depicted in Figure 16, the land cover map obtained via Orthrus shows low salt and pepper errors. This is probably due to the use of the standard deviation statistic. In addition, we can

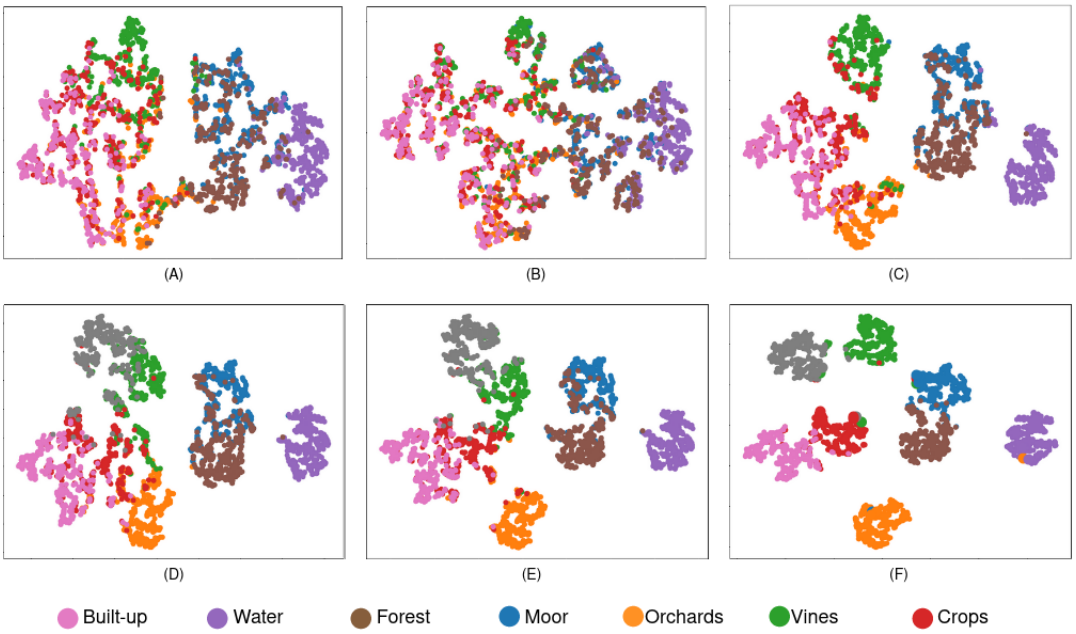

Fig. 14 Visualization of the embeddings learnt by: (a) MultiRocket$_{EF}$, (b) TwoBCNN, (c) ConvTran, (d) Orthrus$_{M}$, (e) Orthrus$_{MM}$ and (f) Orthrus on the *Dordogne* study site.

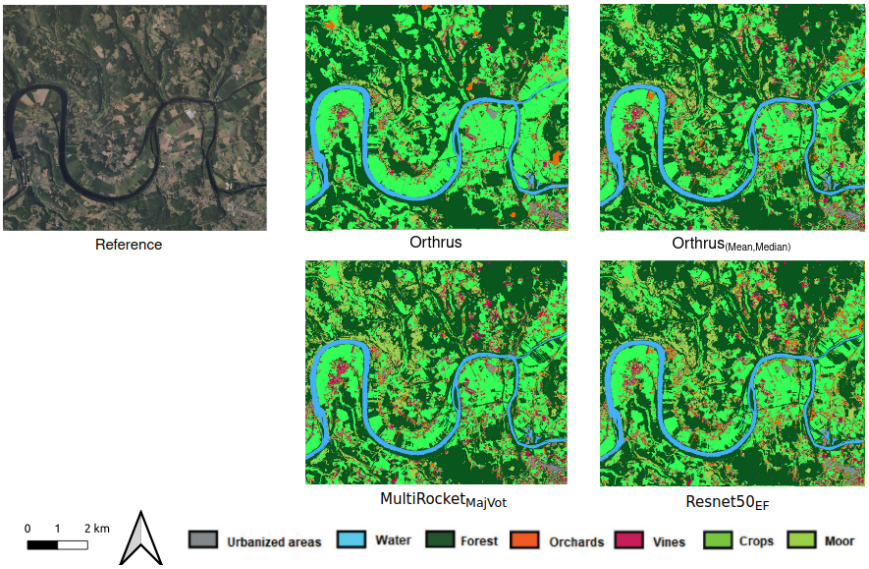

Fig. 15 Land cover map extracts by different competing methods on the *Dordogne* study site.

confirm that only Orthrus was able to detect a realistic amount of `4.orchards`, as compared to the other competing methods, according to field experts. Finally, we would underline that MultiRocket$_{LF}$ exhibits serious confusions between `6.Moor and Savannah` and both `5.wooded areas` and `1.Pasture and fodder` while our approach clearly reduce such confusions.

7 CONCLUSION

This study introduces a novel approach, Orthrus, for enhancing LULC mapping using multi-scale and multi-temporal remote sensing data. The approach combines pixel-level information and spatial context, provided by object-level features, by using 2D encoded image representations derived from various encoding techniques (GADF, GASF, MTF, and RP). These images are then analyzed using a two-branch CNN architecture that employs a late fusion strategy. To describe the object-level information, we exploit standard statistical features (mean, median, and standard deviation) to capture spatial patterns and relationships within the data. Both quantitative and qualitative evaluation results have demonstrated the effectiveness of our approach w.r.t. recent state-of-the-art approaches for SITS-based LULC mapping.

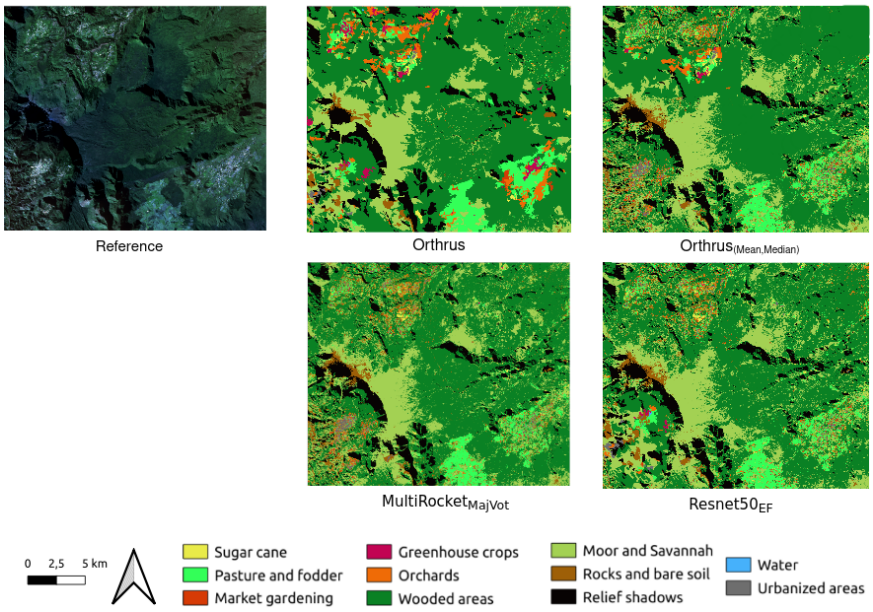

Fig. 16 Land cover map extracts by different competing methods on the *Reunion-Island* study site.

As a future direction, extending the encoding technique to multi-source remote sensing scenarios (e.g. integrate Sentinel-1 Synthetic Aperture Radar data in addition to the optical Sentinel-2 data) could allow to combine complementary information to further ameliorate the model performances for the downstream land cover mapping task. Furthermore, delving into the use of Vision Transformers (ViT) to enhance the exploitation of the 2D encoded time series may provide valuable insights and improvements to the whole pipeline.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

For both datasets, we utilised publicly available Sentinel-2 images, which can be accessed at http://theia.cnes.fr. Only the reference spatial database related to the Reunion-Island dataset is publicly available and can be found at https://doi.org/10.18167/DVN1/TOARDN. Regarding the data related to the Dordogne study site, the data can be shared upon request.

References

[1] Hosseiny, B., Abdi, A.M., Jamali, S.: Urban land use and land cover classification with interpretable machine learning – a case study using sentinel-2 and auxiliary data. Remote Sensing Applications: Society and Environment **28** (2022)

[2] Seyam, M.M.H., Haque, M.R., Rahman, M.M.: Identifying the land use land cover (lulc) changes using remote sensing and gis approach: A case study at bhaluka in mymensingh, bangladesh. Case Studies in Chemical and Environmental Engineering **7**, 100293 (2023)

[3] Haldar, S., Mandal, S., Bhattacharya, S., Paul, S.: Dynamicity of land use/land cover (lulc) an analysis from peri-urban and rural neighbourhoods of durgapur municipal corporation (dmc) in india. Regional Sustainability **4**(2) (2023)

[4] Abidi, A., Ben Abbes, A., Gbodjo, Y.J.E., Ienco, D., Farah, I.R.: Combining pixel- and object-level information for land-cover mapping using time-series of sentinel-2 satellite data. Remote Sensing Letters (2022)

[5] Ienco, D., Gbodjo, Y.J.E., Gaetano, R., Interdonato, R.: Weakly Supervised Learning for Land Cover Mapping of Satellite Image Time Series via Attention-Based CNN. IEEE Access (2020)

[6] Rhif, M., Abbes, A.B., Martínez, B., Farah, I.R.: Veg-w2tcn: A parallel hybrid forecasting framework for non-stationary time series using wavelet and temporal convolution network model. Applied Soft Computing **137** (2023)

[7] Abidi, A., D.Ienco, Abbes, A.B., Farah, I.R.: Combining 2d encoding and convolutional neural network to enhance land cover mapping from satellite image time series. Engineering Applications of Artificial Intelligence **122** (2023)

[8] Pelletier, C., Webb, G.I., Petitjean, F.: Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing **11** (2019)

[9] Chelali, M., Kurtz, C., P., A., Vincent, N.: Deep-star: Classification of image time series based on spatio-temporal representations. Computer Vision and Image Understanding **208-209** (2021)

[10] Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P.: Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment (2012)

[11] Civrizoglu, B.A.: A novel approach and application of time series to image transformation methods on classification of underwater objects. Gazi Journal of Engineering Sciences (2021)

[12] Wu, D., Junjie, X., Zhi, G.X., Huimin, Z.: An Enhanced MSIQDE Algorithm with Novel Multiple Strategies for Global Optimization Problems. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2022)

[13] Alshari, E.A., Gawali, B.W.: Development of classification system for lulc using remote sensing and gis. Global Transitions Proceedings **2**(1) (2021)

[14] L. Ma and, X.Z. M. Schmitt and: Uncertainty analysis of object-based land-cover classification using sentinel-2 time-series data. Remote Sens. **12**(22) (2020)

[15] De Giglio, M., Greggio, N., Goffo, F., Merloni, N., Dubbini, M., Barbarella, M.: Comparison of pixel- and object-based classification methods of unmanned aerial vehicle data applied to coastal dune vegetation communities: Casal borsetti case study. Remote Sensing **11**(12) (2019)

[16] Blaschke T. Hay G.J. Kelly M. Lang S. Hofmann P. Addink, E.F.R.Q.v.d.M.F.v.d.W.H.v.C.F.e.a.: Geographic object-based image analysis – towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing **87** (2014)

[17] Dawa, D., Jordi, I., Julien, M.: Geometry aware evaluation of handcrafted superpixel-based features and convolutional neural networks for land cover mapping using satellite imagery. Remote Sensing (2020)

[18] Wang, L., Sousa, W.P., Gong, P.: Integration of object-based and pixel-based classification for mapping mangroves with ikonos imagery. International Journal of Remote Sensing (2004)

[19] Christian, V., Iraklis, L.: Analysis of time series imaging approaches for the application of fault classification of marine systems. In: 32nd European Safety and Reliability Conference (2022)

[20] Xing, C., Huimin, Z., Shifan, S., Yongquan, Z., D.Wu, Huayue, ., Wuquan, D.: An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application. Expert Systems with Applications (2021)

[21] A.Sagheer, M.Kotb: Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. Scientific Reports (2019)

[22] N.Menini, A.Almeida, Rubens, L., Maire G., Santos Jefersson A., Helio, P., Marina, H., S., T.R.: A soft computing framework for image classification based on recurrence plots. IEEE Geoscience and Remote

Sensing Letters (2019)

[23] D.Dias, A.Pinto, U.Dias, Rubens, L., Maire, G.L., R.Torres: A multi-representational fusion of time series for pixelwise classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2020)

[24] W.Zhiguang, O.Tim: Imaging time-series to improve classification and imputation. IJCAI International Joint Conference on Artificial Intelligence (2015)

[25] J.Belaire-Franch, D.Contreras: Recurrence plots in nonlinear time series analysis: Free software. Journal of Statistical Software (2002)

[26] lung, Y.C., yi, Y.C., xuan, C.Z., wei, L.N., Senior, M.: Multivariate time series data transformation for convolutional neural network. Proceedings of the 2019 IEEE/SICE International Symposium on System Integration, SII 2019 (2019)

[27] A.Censi, D.Ienco, YJE.Gbodjo, R.Pensa, R.Interdonato, R.Gaetano: Attentive Spatial Temporal Graph CNN for Land Cover Mapping from Multi Temporal Remote Sensing Data. IEEE Access (2021)

[28] Fare, G.V.S., Loic, L.: Lightweight temporal self-attention for classifying satellite images time series. In: Lemaire, V., Malinowski, S., Bagnall, A., Guyet, T., Tavenard, R., Ifrim, G. (eds.) Advanced Analytics and Learning on Temporal Data. Springer, ??? (2020)

[29] W.Zhang, H.Zhang, Z.Zhao, Tang, P., Z.Zhang: Attention to both global and local features: A novel temporal encoder for satellite image time series classification. Remote Sensing **15** (2023)

[30] Ch.Zhang, P.Yue, D.Tapete, L.Jiang, Shangguan, B., Huang, L., Liu, G.: A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing **166** (2020)

[31] D.Derksen, J.Inglada, J.Michel: A metric for evaluating the geometric quality of land cover maps generated with contextual features from high-dimensional satellite image time series without dense reference data. Remote Sensing (2019)

[32] Mohammadi, S., Belgiu, M., Stein, A.: Improvement in crop mapping from satellite image time series by effectively supervising deep neural networks. ISPRS Journal of Photogrammetry and Remote Sensing **198** (2023)

[33] Jiang, W., Zhang, D., Ling, L., Liu, Y., Cao, J., Liu, G.: Time series classification based on image transformation using feature fusion strategy. Neural Processing Letters **54** (2022)

[34] Lee, H., Lee, J.: Convolutional model with a time series feature based on rssi analysis with the markov transition field for enhancement of location recognition. Sensors **23** (2023)

[35] Menini, N., Almeida, A.E., Lamparelli, R., Maire G., Santos J A., Pedrini, H., Hirota, M., Torres, R.d.S.: A soft computing framework for image classification based on recurrence plots. IEEE Geoscience and Remote Sensing Letters **16**(2) (2019)

[36] Mumuni, A., Mumuni, F.: Cnn architectures for geometric transformation-invariant feature representation in computer vision: a review. SN Computer Science **2**(5), 340 (2021)

[37] JE.Yawogan, D.Ienco, Leroux, L., R.Interdonato, Gaetano, R., Ndao, B.: Object-based multi-temporal and multi-source land cover mapping leveraging hierarchical class relationships. Remote. Sens. (2020)

[38] Interdonato, R., Ienco, D., Gaetano, R., Ose, K.: DuPLO: A DUal view Point deep Learning architecture for time series classificatiOn. ISPRS Journal of Photogrammetry and Remote Sensing **149** (2019)

[39] Olivier, H., Mireille, H., Villa, P.D., Dedieu, G.: A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of formosat-2, landsat, venμs and sentinel-2 images. Remote Sensing (2015)

[40] Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D..: Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sensing (2017)

[41] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11) (2012)

[42] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[43] Baranwal, A., Bagwe, B.R., Vanitha, M.: Machine learning in python. Handbook of Research on Applications and Implementations of Machine Learning Techniques **12** (2019)

[44] Kalita, I., Roy, M.: Inception time dcnn for land cover classification by analyzing multi-temporal remotely sensed images. In: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium (2022)

[45] Fawaz, H.I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.-, Petitjean, F.: Inceptiontime: Finding alexnet for time series classification. Data Min. Knowl. Discov. (2020)

[46] Chang, W.T., A.Dempster, Bergmeir, C., Webb, G.: MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. Data Mining and Knowledge Discovery (2022)

[47] Foumani, N.M., Tan, C.W., Webb, G.I., Salehi, M.: Improving position encoding of transformers for multivariate time series classification. Data Mining and Knowledge Discovery **38**, 22–48 (2024)

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[49] Leland, M., Healy, N. J.and Saul, Lukas, G.: UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software (2018)