EXPLAINING A DEEP SPATIOTEMPORAL LAND COVER CLASSIFIER WITH ATTENTION AND REDESCRIPTION MINING

N. Méger[1,*], H. Courteille[1], A. Benoit[1], A. Atto[1], D. Ienco[2]

[1] Polytech Annecy-Chambery, LISTIC, Université Savoie Mont Blanc, F-74940 Annecy, France
(nicolas.meger, hermann.courteille, alexandre.benoit, abdourrahmane.atto)@univ-smb.fr
[2] INRAE, UMR TETIS, Université de Montpellier, F-34000 Montpellier, France - dino.ienco@inrae.fr

KEY WORDS: Explainable AI, Deep Learning, Land Cover Classification, Satellite Image Time Series, Attention, Redescription Mining, Grouped Frequent Sequential Patterns.

ABSTRACT:

Deep learning-based land cover classifiers learnt from Satellite Image Time Series (SITS) are known to reach high performances. In order to explain, at least partly, the rationale leading to each one of their decisions, attention-based architectures have been proposed to automatically weight the importance of predefined data components in the classification process. Though generated for each decision separately, the informational content conveyed by such explanations can remain insufficient to end-users because of the complex nature of SITS. Moreover, getting a general perspective about the way a classifier works requires merging all explanations for each class and relating them to its mode of operation, which is not always straightforward. A preliminary and complementary approach for automatically identifying the data features detected by a pixel-wise deep spatiotemporal land cover classifier and explaining its behavior at the class level is therefore proposed in this paper. Classified pixels are first described using interpretable features coming under the form of data mining patterns. A redescription mining technique is then employed to automatically select, for each class, the features matching the different activation level configurations of the layer that is assumed to capture the aforementioned patterns. Experiments based on a Sentinel-2 time series and a deep spatiotemporal neural network implementing a channel-separated processing as well as a channel-based attention mechanism show the interest of such a combined approach.

1. INTRODUCTION

Land Cover Classification (LCC) is a task that has been benefiting from the last advances in deep learning for several years (Vali et al., 2020). Many works such as (Pelletier et al., 2019), (Rußwurm and Körner, 2020) or (Censi et al., 2021) shows that high-performance deep land cover classifiers can be learnt from *Satellite Image Time Series (SITS)*. Moreover, the *black-box* nature of these *Deep Neural Networks (DNNs)* has been being addressed using different methods designed to gain insights into the rationale leading to their decisions (Campos-Taberner et al., 2020). Among them, *Saliency Masks (SMs)* have been proven to be effective at explaining each one of the outcomes. These masks can be produced by recent explanation methods such as Grad-CAM (Selvaraju et al., 2017) and SHAP (Lundberg and Lee, 2017). Other explanation methods can be established by integrating specific operators to both enhance model relevance and provide insights on the predictions, for example using *attention*-based architectures. Works such as (Garnot et al., 2020), (Rußwurm and Körner, 2020), (Ienco et al., 2020), (Censi et al., 2021) or (Courteille et al., 2021) follow this direction. These architectures weight the importance of predefined data components in the classification process. To our knowledge, no proposal taking into account all SITS dimensions has emerged so far. In addition, no explanation regarding the results supplied by attention mechanisms themselves is provided. Finally, as decisions are explained independently from each other, getting a general perspective about the way a classifier works requires merging all explanations for each class and relating them to its mode of operation, which is complex.

A preliminary and complementary approach for identifying automatically the dataset features detected by a pixel-wise deep spatiotemporal land cover classifier and explaining its functioning at the class level is therefore proposed in this paper. Each classified pixel is first described using interpretable features coming under the form of data mining patterns. These features, built using all SITS dimensions, are detailed in Section 2. Each pixel is then also characterized by the activation levels of the layer that is assumed to summarize a large part of the processing performed by the network and capture high level concepts similar to the aforementioned intepretable features. Finally, pixel descriptions are analyzed for each class separately using redescription mining (Galbrun and Miettinen, 2017) to automatically extract correspondence rules between the interpretable features and the different activation level configurations. The redescription mining technique adopted in this paper is presented in Section 3. The proposed approach is assessed using a Sentinel-2 time series and a spatiotemporal deep neural network implementing a channel-separated processing as well as a channel-based attention mechanism. This experimental setting is described in Section 4. Obtained results are made available in Section 5. They show that, while preserving good performances, the attention-based explanations of the decisions are meaningful and can be enriched with class level explanations relying on intepretable features identified by redescription mining. Section 6 concludes this paper by summarizing the main contribution, discussing its limits and pointing to possible future work directions.

* Corresponding author

2. GROUPED FREQUENT SEQUENTIAL PATTERNS

The first step of the proposed approach consists in characterizing each pixel for which a ground truth is available with interpretable features. It is assumed that some of these features are captured by deep SITS land cover classifiers and could help understand how the latter distinguish classes. With the object of exploiting all SITS dimensions, spatiotemporal features extracted from each available channel are focused on. They are obtained by mining the so-called *Grouped Frequent Sequentiel patterns* originally proposed in (Julea et al., 2011). These patterns are spatiotemporal data mining patterns designed to describe and summarize SITS at the pixel level in an unsupervised way (Julea et al., 2011). We propose to characterize each pixel with such patterns by extracting them from each channel separately. Let us consider a given channel. Its pixel values are first quantized with standard techniques such as equal frequency bucketing or clustering (Julea et al., 2011). Quantized pixel values are then associated with symbols denoting the quantization intervals they belong to. As a result, each pixel is described by a symbolic series containing as many symbols as the number of images involved in the SITS. The set of symbolic series describing pixels is then mined to extract all possible *sequential* patterns such as $2 \rightarrow 3 \rightarrow 2$. This pattern, if observed for a symbolic series, indicates that, some time in the series, the value of the pixel it describes is denoted by quantization interval '2', then, some times later, by interval '3', and finally, some times later by interval '2' once again. No timing constraint is imposed, and there might be other symbols occurring in-between the occurrences of the pattern symbols. Moreover, if this pattern is observed for several other symbolic series, i.e., it affects other pixels, no synchronisation constraint between pattern occurrences is considered. In other words, such a pattern is allowed to occur anywhere in space and in time.

In order to reject spurious patterns and safely prune the search space, two spatial constraints must be fulfilled. A pattern is retained if: 1) it affects a sufficient number of pixels, i.e., it covers a minimum surface and 2) affected pixels are sufficiently connected to each other on average, i.e., they form homogeneous regions in space, whatever their shapes. The first constraint is simply set using a *minimum surface threshold* termed *minimum frequency threshold* and denoted σ. The second constraint is evaluated by considering the 3×3 neighbourhood of each affected pixel, counting the number of immediate neighbours that are also affected by the same pattern, averaging all of these counts and checking whether this mean exceeds or not a *minimum grouping threshold* denoted κ. If unclassified pixels are located in the neighborhood of affected pixels, it is assumed that they are not affected by any pattern. Applying these spatial constraints allow to target sequential patterns whose occurrences are *frequent* and *grouped*, hence the name of Grouped Frequent Sequential Patterns or *GFS-patterns*. In addition, only *maximal* GFS-patterns are selected to focus on the most specific ones. A pattern is maximal if it is not contained in any other pattern of the output collection. The reader is referred to (Julea et al., 2011) for a more formal definition of GFS-patterns and details regarding the corresponding extraction algorithm. Finally, maximal GFS-patterns are ranked using a dedicated and efficient randomization approach designed to guide end-users towards the most promising GFS-patterns, i.e., the patterns that are the less or the more likely to occur in randomized versions of the symbolic datasets. More details about this ranking method can be found in (Méger et al., 2019). Back to this paper, it is proposed to describe each pixel for which a ground truth is available by indicating, for each channel, which are the most promising maximal GFS-patterns occurring in its symbolic series, if any. In the following, the most promising maximal GFS-patterns are simply referred to as *patterns* when clear from the context.

3. REDESCRIPTION MINING

Redescription mining is *'a data analysis task that aims at finding distinct common characterizations of the same objects'* (Galbrun and Miettinen, 2017). For instance, if geographical areas are described on one side by the presence of mammal species, and, on the the other side by their temperatures, it is possible to find the following *redescription*: *'The areas inhabited by either the Eurasian lynx or the Canada lynx are approximately the same areas as those where the maximum March temperature ranges from -24.4 °C to 3.4 °C.'* (Galbrun and Miettinen, 2017). More precisely, let us suppose that each studied object is described by as many tables as they are distinct characterisations. Within such tables, each object, i.e. a row, is described by attributes, i.e. columns, whose type can be numerical, categorical or Boolean. Each table can be separately used to produce *descriptions* of the objects, i.e. expressions built using the attributes of the table. Each expression allows to assign a Boolean value to each object by checking whether it characterizes it, even partially, or not. Back to the first redescription example, it can be alternately expressed with descriptions $p = Eurasian\ lynx \wedge Canadian\ lynx$ and $q = [-24.5 \leq March\_maximum\_temperature \leq 3.4]$ originating from the table denoting the presence of mammal species and the table reporting the temperatures respectively. The set of objects for which a description is true is termed *support*. A *redescription* is a pair of descriptions such as (p, q), also denoted $p \sim q$, each description being produced from a different table. In order to rank redescriptions according to their accuracy, the Jaccard index of each description $p \sim q$ is computed as $\frac{|supp(p) \cap supp(q)|}{|supp(p) \cup supp(q)|}$. It allows to evaluate to which extent the objects for which a description is valid can be also characterized by the other description, i.e., to which extent descriptions p and q are similar. The set of objects characterized by both p and q is the *support* of redescription $p \sim q$. Finally, redescription mining is about finding all redescriptions by taking into account potential additional constraints such as limiting the support of redescriptions, limiting the length of descriptions or selecting redescriptions that are statistically significant. The reader is referred to (Galbrun and Miettinen, 2017) for a more formal and general overview of redescription mining.

In this paper, objects are pixels described by 1) the different activation levels of the neurons of the layer that is assumed to capture GFS-patterns, and 2) GFS-patterns themselves. Let a_i denote the activation level of neuron i. Since we aim to automatically match the different activation level configurations with the presence of patterns, redescriptions such as $[0.2 \leq a_1 \leq 0.3] \wedge [0.7 \leq a_{17} \leq 0.9] \sim 2 \rightarrow 3 \rightarrow 2$ are expected. The *ReReMi* algorithm proposed in (Galbrun and Miettinen, 2012a) is therefore considered. It can indeed handle Boolean data and automatically determine the optimal numerical intervals that should be considered when establishing a description from numerical attributes. In addition, a wide range of descriptions are explored. Disjonctions and conjonctions can be employed, and variables can be negated. Some restrictions are nevertheless considered to make extractions tractable: the descriptions are evaluated from left to right, irrelevant of the operator precedence, and every variable can be used only once.

Basically, *ReReMi* can extract redescriptions whose Jaccard index is above a user-defined threshold and whose descriptions are statistically dependent. This dependence is checked using a p-value expressing the probability that the supports of descriptions overlap as much as observed. Such a test tends to favor redescriptions with low support and can be counter-balanced by rejecting those whose support is below a user-defined threshold. Even if such additional constraints are fulfilled, the search space remains exponential and an heuristic pruning is performed: the algorithm starts from the best redescriptions whose descriptions contain only one variable and greedily expand their descriptions as long as they form the best redescriptions. More details about the algorithm can be found in (Galbrun and Miettinen, 2012a).

4. EXPERIMENTAL SETTING

The experimental setting comprises a SITS for which a ground truth and accurate classifiers are available in the literature as well as a dedicated deep spatiotemporal land cover classifier whose design is inspired by state-of-the-art classifiers.

4.1 The Sentinel-2 Time Series

A Sentinel-2 SITS covering the Réunion island, for which the ground truth is available in (Dupuy et al., 2020) and land cover classifiers have been proposed in (Ienco et al., 2020) and (Courteille et al., 2021), is used. It consists of 21 images with size 6667 pixels $\times$ 5916 pixels. They were acquired between January and December 2017 and cover a 67 km x 59 km scene with a 10 m spatial resolution. Clouds are filtered using a multilinear interpolation (Ienco et al., 2020). The following spectral bands are available: B2 (blue), B3 (green), B4 (red) and B8 (near-infrared). The Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI) (McFeeters, 1996) are also supplied. These standard indexes are defined by $NDVI = f(B8, B4)$ and $NDWI = f(B3, B8)$ with $f(x, y) = \frac{x-y}{x+y}$, an homogeneous function from $\mathbb{R}_+^* \times \mathbb{R}_+^*$ to $[-1; 1]$. Regarding the ground truth, 2% (880,828 pixels) of the pixels are annotated according to 11 unbalanced land cover classes listed in Table 3 along with their class ratios.

4.2 The Deep Spatiotemporal Land Cover Classifier

The *Deep Spatiotemporal Land Cover Classifier (DSLCC)* used in this paper is designed to be as interpretable as possible. Neural networks based on recurrent cell models such as *Long Short Term Memory (LSTM)* ones (Ienco et al., 2017) are thus discarded in favor of *Convolutionnal Neural Networks (CNN)* such as (Pelletier et al., 2019) or (Ienco et al., 2020). Moreover, the CNN *field of view (fov)*, i.e., the extent of the input neighborhood influence, can be controlled by design. This allows us to specify the temporal fov, which is crucial when dealing SITS as they generally contain few acquisitions. Following the work of (Courteille et al., 2021), all the convolutions are performed for each channel separately, and a auxiliary attention branch weighting the importance of each channel in the decision of the classifier is added. Such an attention operator is chosen because it delivers meaningful explanations that can be easily merged for each class using box plots (Courteille et al., 2021). Its architecture is detailled in Table 1. Inspired by the work of (Pelletier et al., 2019), temporal convolutions (layers ③ and ④) are applied at the pixel level. They are carried out after having computed spatiotemporal ones whose spatial footprint is 3×3 to match the spatial extent used to extract GFS-patterns (layers ②). As a result, a description based on 96 neurons is obtained for each channel independently (layers ④). These neurons are assumed to summarize a large part of the processing performed by the network and capture high level concepts similar to GFS-patterns. Their number is set so as to limit the number of activation level configurations and ease interpretability. Finally, channel-wise features are gathered and then summarized with dimension reduction layers (⑤, ⑦ and ⑧). For each channel, an auxiliary attention branch (layer ⑥) is added to weight the importance of each channel in the decision of the classifier. To fit the Réunion SITS with $T = 21$ timestamps, the size of convolutional kernels is adjusted to set a temporal field of view such that $k_1 = k_2 = k_3 = 7$. Finally, the number of convolutional filters progressively decreases with $N_1 = 256$, $N_2 = 64$, and $N_3 = 32$.

Layer	Operation	Specifications	Tensor output shape
①	split	6 input channels	$(T, 3, 3, 1)(\times 6)$
② (x 6)	conv3d	kernel=$(k_1, 3, 3)$, no padding, $n_{filters} = N_1$	$(T_1, 1, 1, N_1)$
③ (x 6)	conv1d	kernel=(k_2), with or no padding , $n_{filters} = N_2$	(T_2, N_2)
④ (x 6)	conv1d	kernel=(k_3), with or no padding, $n_{feat} = N_3$	(T_3, N_3)
⑤	stack	6 channels from 4	$(6, T_3, N_3)$
⑥	attention	6 weights α_i	
⑦ from ⑤	hidden dense	+ relu	256 neurons
⑧	final dense	+ softmax	11 neurons

Table 1. DSLCC architecture. Features are extracted by 3 different convolutional layers computed for each one of the 6 channels (6 times layers ②, ③, ④). Attention layer ⑥ is outing 6 channel attention weights.

5. EXPERIMENTS

5.1 Pattern extraction

All pixels for which a ground truth is available are characterized using patterns (see Section 2). The whole extraction process, from pixel value quantization to pattern ranking is run using the free prototype *DFTS-P2miner* (Nguyen et al., 2019). Original pixel values are quantized with three intervals using equal frequency bucketing. As a result, for each channel, pixel values are denoted by symbols '1', '2', and '3', that respectively report 'low values', 'medium values' and 'high values'. Such a preprocessing is commonly adopted to mine patterns (Julea et al., 2011). The minimum surface of a pattern is set so as to be able to find patterns for each class. The class that is less represented, namely 'greenhouse crops', only contains $1,931$ pixels. Assuming that about half of these pixels share the same kind of evolution, the minimum surface was set to 881 pixels, which is a very weak constraint since it represents 0.1% of annotated pixels. The average connectivity constraint imposed on the pixels affected by a pattern is set to 5, which is a standard setting when processing optical SITS (Julea et al., 2011). The 20 patterns that are the more likely to occur in randomized versions of the data are retained along with the 20 patterns that are the less likely to do so, which is once again a standard setting (Méger et al., 2019). For a given channel, each pixel is thus described by indicating the absence or presence of 40 patterns.

5.2 Training of the DSLCC

The model optimisation is performed using TensorFlow2. Inputs are normalised channel-wise between 0 and 1 using a min-max feature scaling to facilitate the convergence of the training

process. A stratified sampling preserving class ratios is employed to split annotated pixels into a training dataset (60%), a validation and test dataset (20% each). Pixels belonging to a same object all belong to the same dataset. The network is learnt by considering the standard unweighted categorical cross-entropy CE and defining the loss as $L_{global} = CE(Y, Y_{main}) + 0.5 \times CE(Y, Y_{aux})$ where Y_{main} and Y_{aux} are the model main and auxiliary outputs. Gradients are back-propagated using an Adam optimizer and an $\mathbb{L}^2$-regularization with a weight decay of 1.10^{-6} to avoid overfitting on all layers.

The performances of the DSLCC is assessed against a classical random forest (500 trees, 200 splits) and two deep land cover classifiers, namely *TempCNN* (Pelletier et al., 2019) and *Sdeep-B-Multi-ii* (Courteille et al., 2021). Both work at the pixel level, ignore the spatial dimension, exploit all bands and indexes, and rely on temporal convolutions. In addition, Sdeep-B-Multi-ii is equipped with an attention operator weighting the importance of each channel in the final classification decision. The random forest, *TempCNN*, *Sdeep-B-Multi-ii* and the DSLCC respectively reach 90.4%, 91.3%, 92.2% and 84.4% of accuracy. Though less accurate, the performance of DSCLC is still decent. It is detailled in Table 2.

Class	Precision	Recall	% of annotated pixels
Sugar cane	88	89	10.1
Pasture	88	84	7.7
Market gardening	61	65	2.0
Greenhouse crops	17	35	0.2
Orchards	62	67	3.8
Wooded areas	83	86	23.3
Moor	84	79	17.6
Rocks	92	92	17.5
Relief shadows	78	91	6.2
Water	98	84	9.2
Urban area	81	80	2.2
Mean	**75.8**	**77.4**	-

Table 2. Precision and recall by class for the DSLCC on the test set (176,166 pixels).

5.3 Explaining the DSLCC decisions

In order to produce explanations as general as possible, all annotated pixels are supplied to the DSLCC to infer their classes. As expected, the overall accuracy is still high with a score of 87%. The precision and the recall obtained for each one the classes are listed in Table 3.

Class	Precision	Recall	% of annotated pixels
Sugar cane	96.7	96.6	12.4
Pasture	92.8	94.0	7.3
Market gardening	75.1	74.3	2.3
Greenhouse crops	52.9	52.3	0.2
Orchards	80.1	83.7	3.9
Wooded areas	87.3	94.4	23.5
Moor	92.4	77.9	16.0
Rocks	97.5	97.7	21.4
Relief shadows	94.3	98.7	5.1
Water	99.9	99.4	6.1
Urban area	84.6	91.0	1.8
Mean	**86.7**	**87.3**	-

Table 3. Precision and recall by class for the DSLCC for all annotated pixels (880,828 pixels).

5.3.1 Attention: As shown in Figure 1 for some classes of interest and all annotated pixels, the attention weights differ from one class to another, which illustrates the interest of the attention operator. It can be noted that all bands are exploited to take decisions. These weights are extremely similar to the ones obtained at the scale of the validation and test sets. One exception is class 'Water'. This can be explained by the presence of different evolution sub-classes. Finally, let us remark that merging attention weights is simple when considering channels or the temporal dimension. Contrarily, doing so for the spatial dimension or a set of dimensions is far from being straightforward.

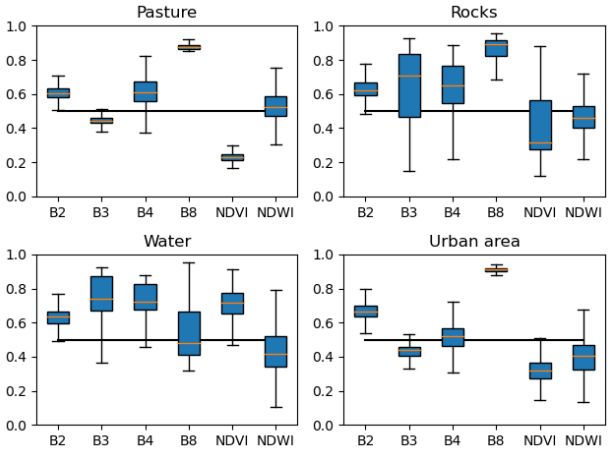

Figure 1. All annotated pixels: box plots of channel weight attention for 4 classes, normalized by class sums. Horizontal lines depict activation thresholds (0.5 for a sigmoid).

5.3.2 Redescription mining: Channel attention-based explanations given by Figure 1 are enriched with redescriptions extracted from annotated pixel descriptions, i.e., patterns and the activation levels of layer ④ observed at inference time. This is performed for each predicted class of interest and each channel using the free prototype *SIREN* (Galbrun and Miettinen, 2012b). In order to extract expressive but yet simple redescriptions, disjunctions and negated variables are not considered, the maximum number of variables is set to 4 for the neuron activation level ones and 1 for pattern ones. The minimum Jaccard index is set to the SIREN default value, i.e., 0.01, to extract as much redescriptions as possible. The minimum number of pixels supporting a redescription is arbitrarily set to the SIREN default value, i.e. 5%. The maximum p-value is set to the standard value 0.05. All other SIREN parameters resort to default values. Finally, once redescriptions are extracted, redondant ones are filtered out by removing those supported by pixels supporting other redescriptions whose Jaccard index is better.

The following table gives, for each predicted class of interest and each band, the number of extracted redescriptions (r), the fraction of pixels supporting all of them (S_{all}), the minimum and maximum cardinality of their supports (S_{min}, S_{max}) as well as their minimum and maximum Jaccard indices (J_{min} and J_{max}). Though redescriptions are statistically significant since their p-values are lower or equal to 0.05, the match between the activation level configurations and the patterns is not complete. According to J_{min} and J_{max}, redescription accuracy indeed varies between 0.08 and 0.66. Moreover, since S_{all} reaches 0.7 as a maximum, redescriptions extracted from different bands should be therefore considered jointly to explain as much decisions as possible, which is no suprise since the classifier itself adopts this strategy to infer classes (see Section 5.3.1).
The cover of decisions by redescriptions C_r, i.e. the percentage of decisions that can be explained with one or more redescriptions is given for each class in Table 5. This cover is provided along with the minimum, maximum and mode numbers of redescriptions per decision, R_m, R_M, R_d, and the minimum, maximum and mode numbers of bands from which these decisions

Class / Band	r	S_{all}	S_{min}	S_{max}	J_{min}	J_{max}
Pasture / B2	2	0.28	0.05	0.23	0.18	0.45
Pasture / B3	3	0.33	0.05	0.18	0.19	0.47
Pasture / B4	5	0.70	0.06	0.20	0.23	0.40
Pasture / B8	2	0.48	0.08	0.40	0.19	0.42
Pasture / NDVI	1	0.05	0.05	0.05	0.15	0.15
Pasture / NDWI	2	0.41	0.14	0.27	0.27	0.45
Rocks / B2	2	0.11	0.05	0.06	0.08	0.33
Rocks / B3	3	0.20	0.05	0.10	0.11	0.41
Rocks / B4	5	0.33	0.05	0.09	0.15	0.40
Rocks / B8	3	0.15	0.05	0.05	0.12	0.24
Rocks / NDVI	1	0.05	0.05	0.05	0.17	0.17
Rocks / NDWI	4	0.38	0.08	0.12	0.22	0.39
Water / B2	2	0.31	0.14	0.17	0.48	0.51
Water / B3	4	0.36	0.05	0.16	0.17	0.44
Water / B4	1	0.06	0.06	0.06	0.39	0.39
Water / B8	0	-	-	-	-	-
Water / NDVI	0	-	-	-	-	-
Water / NDWI	1	0.60	0.60	0.60	0.66	0.66
Urban area / B2	1	0.15	0.15	0.15	0.48	0.48
Urban area / B3	1	0.05	0.05	0.05	0.08	0.08
Urban area / B4	2	0.10	0.05	0.05	0.13	0.21
Urban area / B8	4	0.26	0.06	0.08	0.23	0.39
Urban area / NDVI	2	0.11	0.05	0.06	0.19	0.23
Urban area / NDWI	2	0.27	0.09	0.18	0.20	0.27

Table 4. Support and accuracy of extracted redescriptions.

originate, B_m, B_M, B_d. As it can be observed, a majority of decisions are covered by redescriptions (between 65% and 91%) using up to 8 redescriptions and all 6 bands. Reported modes also show that decisions are most frequently covered by one ore two redescriptions extracted from one or two bands. Considering several patterns coming from different bands to explain decisions also makes sense when considering input data characteristics. For example, 97% of the pixels labelled as 'Pasture' by the classifier are indeed affected by up to 10 redescription patterns (i.e., the patterns occurring in redescriptions, not redescriptions themselves) while no more than 5 redescriptions patterns can be supplied by a single band for class 'Pasture'. This cover of decisions by redescription patterns, C_{rp}, is given Table 5 for each class along with the minimum, the maximum and mode numbers of redescription patterns covering decisions, R_m, R_M, and B_d. Finally, since not all decisions can be associated with redescriptions, other extraction parameter settings could be tried, in particular for the support of redescriptions and the number of patterns retained to describe each band.

Class	C_r	C_{rp}	R_m	R_M	R_d	B_m	B_M	B_d	RP_m	RP_M	RP_d
Pasture	0.91	0.97	1	8	2	1	6	2	1	10	2
Rocks	0.72	0.83	1	6	1	1	5	1	1	8	1
Water	0.73	0.78	1	7	1	1	4	1	1	7	1
Urban area	0.65	0.80	1	5	1	1	4	1	1	8	2

Table 5. Decision covers.

For class 'Pasture', a very high attention is reported for B8, which is expected as it is a vegetation class. Redescriptions obtained for B8 are given hereafter along with their accuracy, J, and the number of pixels supporting them, S:

- $r1$: $[3.347221 < a_{10}] \sim 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3$ ($J = 0.42,\ S = 0.40$)
- $r2$: $[-35.42756 < a_{33} < -5.023269] \wedge -[7.611569 < a_{62} < -1.214998] \wedge [0.322748 < a_{70} < 1.388656] \sim 3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2$ ($J = 0.19,\ S = 0.08$)

These descriptions illustrates the fact that a pattern can be captured using a single or more neurons. Redescription $r1$ indicates that 40% of the decisions are associated with a pattern exhibiting a continuous presence of vegetation at a quite high level: symbol '3' occurs in 20 images (the series contains 21 images). Though accounting for 8% of the decisions, redescription $r2$ is interesting as it shows a loss of biomass with a series of 4 symbols '3' followed by five symbols '2'. Both patterns can be localized within the ground truth areas actually covered by pastures using synthetic colors to denote both their spatial extent and the date at which they ends in the SITS. The temporal color palette used in this paper is given by Figure 2. The white color is also used to point out the pixels of the ground truth belonging to class 'Pasture' that are not affected by patterns. Remaining pixels, i.e. black ones, are simply not covered by pastures according to the ground truth. These SpatioTemporal Localizations Maps (STL-maps) (Méger et al., 2019) have been computed and cropped to focus on a 12 km $\times$ 12 km area of interest located inland, in the region of *la Plaine des Cafres*. The resulting maps are presented with figures 3 and 4 for the $r1$ and $r2$ patterns respectively. No significant temporal behavior is observed for the $r1$ pattern, which is expected as it covers almost all the images of the SITS, i.e., 20 images out of 21. Contrarily, the $r2$ pattern shows more dispersed settings since it is shorter and can end at different dates. This dispersion is particularly visible in the northern part of *La Plaine des Cafres* as shown in Figure 5. As also evidenced by the STL-maps, and since no pixel can be covered both by the $r1$ pattern and the $r2$ pattern[1], these patterns are complementary spatially. These maps thus show that using these patterns to predict class 'Pasture' makes sense even if, as underlined by white pixels, they do not characterize all of its pixels. Knowing that the classifier is very accurate for that class, this is far from being surprising since about half of the decisions are explained by $r1$ and $r2$ (S=0.48). STL-maps computed with respects to decisions 'Pasture' (and not the actual class 'Pasture') are extremely similar for the same reason. They are presented in figures 6 and 7. The same color palette is used to denote the presence of the pattern as well as its ending dates. The white color is used to show decisions for which the activation level conditions are not fulfilled and the pattern is absent. The brown color is associated with decisions such that the activation levels conditions are satisfied and the pattern is absent. If the pattern is present and the activation level conditions are not meet, then the gray color is assigned to pixels.

Regarding NDVI, even if designed to monitor vegetation, the DSLCC hardly exploits it according attention weights (see Figure 1). Let us have a look at the only one and quite weak redescription extracted for this index. It is supported by 5% of the pixels with an accuracy of 0.15%. It refers to pattern $1 \to 1 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$ and thus expresses a gain and then a loss of biomass. Since behaviors similar to B8 ones could also be expected, this confirms that mobilizing NDVI to detect pastures is indeed not of primary interest for this SITS.

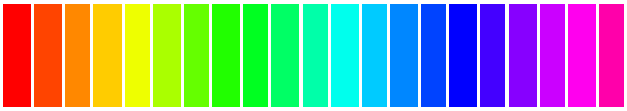

Figure 2. Temporal color palette: 21 acquisitions, from January 2017 (red) to December 2017 (magenta).

Class 'Water' is detected by exploiting all channels at quite high levels according attention weights (see Figure 1). Though being the less exploited according to attention weights, NDWI is generally expected to detect water. This is confirmed by a single but strong redescription:

[1] There is only one symbol left undescribed for pixels affected by the $r1$ pattern, and this pattern contains no symbol '2' while the $r2$ pattern contains 5 symbol '2'.

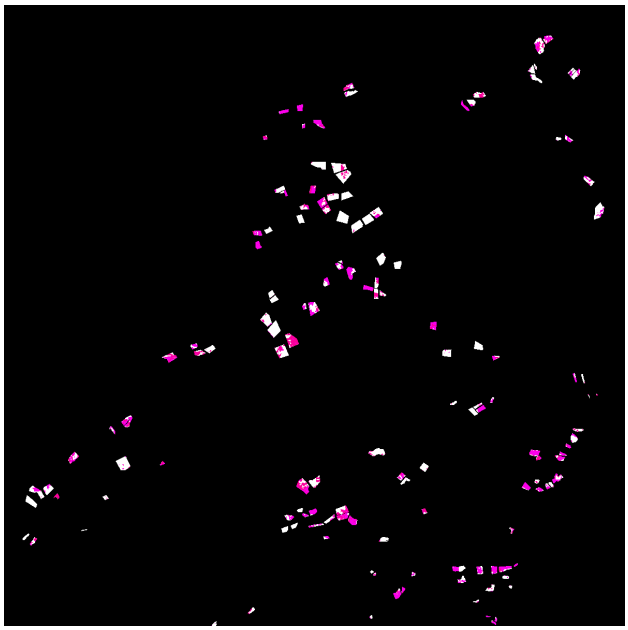

Figure 3. STL-map of the $r1$ pattern, for ground truth areas covered by pastures, in the region of *la Plaine des Cafres*.

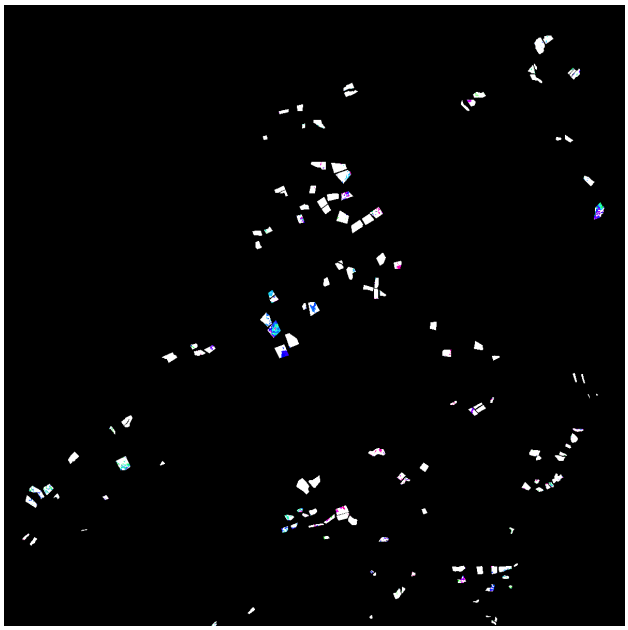

Figure 4. STL-map of the $r2$ pattern, for ground truth areas covered by pastures, in the region of *la Plaine des Cafres*.

- $r3$: $[-31.58766 < a_{92} < -2.46482] \sim 1 \nrightarrow 1 \nrightarrow 2 \nrightarrow 1 \nrightarrow 1 \nrightarrow 1$ $(J = 0.66,\ S = 0.60)$

Redescription $r3$ allows to describe 60% of the decisions leading to class 'Water', which is the highest support reported for the redescriptions extracted for the classes of interest. Its 22 km $\times$ 17 km STL-maps, for areas located along the northwest coast between the towns of Saint-Paul and Saint-Denis, are given by Figure 8 and Figure 9. Though the $r3$ pattern contains few symbols, there is few temporal dispersion with ending dates located at the end of the year. According to the ground truth, it mainly characterizes maritime waters, i.e., the eight large blocks picked up by the experts in coastal waters. Contrarily, a set of inland water zones, grouped at the lower left corner of the image is not

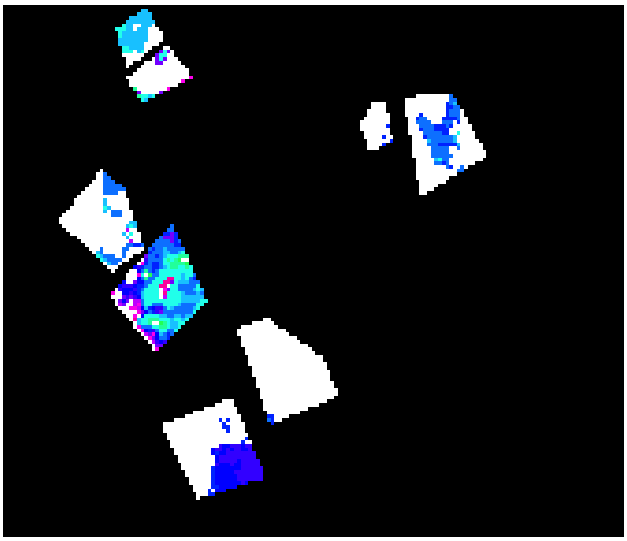

Figure 5. STL-map of the $r2$ pattern for ground truth areas covered by pastures, in the northern part of *la Plaine des Cafres*.

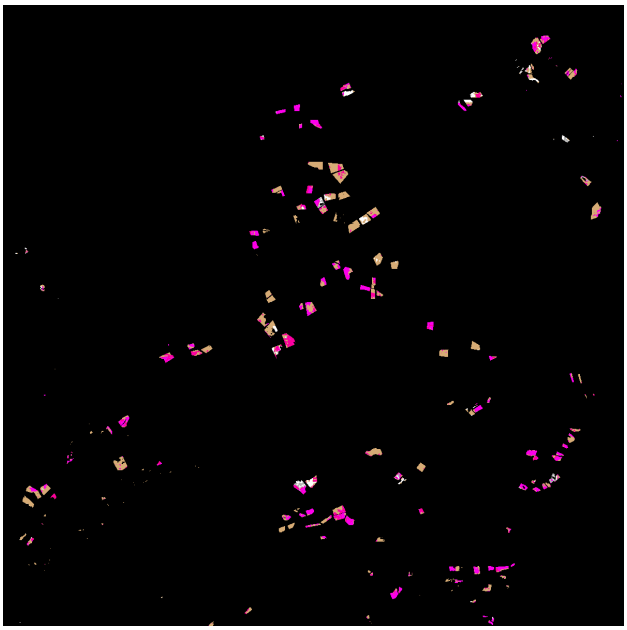

Figure 6. STL-map of the $r1$ pattern, for areas identified as pastures by the DSLCC, in the region of *la Plaine des Cafres*.

captured by the pattern. They belong to the *Réserve naturelle nationale de l'étang de Saint-Paul* and can not be detected using redescription patterns extracted from B2, B3 or B4. The latter indeed focus on maritime waters as well. In more details, B2 is associated to redescription patterns expressing gradual increases as well series of changes, all reflectance levels being mobilized. Band B3 supplies redescription patterns expressing the stability of either medium or high reflectance values. Band B4 is exploited through a redescription expressing an abrupt increase from symbol '1' to symbol '3'.

Finally, B8 and NDVI do not seem to be captured by the network in the way patterns do since no redescription is to be reported for these bands. Knowing that the DSLCC identifies maritime waters as well as inland waters such as those of the *Réserve naturelle nationale de l'étang de Saint-Paul*, one could simply argue that patterns do not characterize such areas.

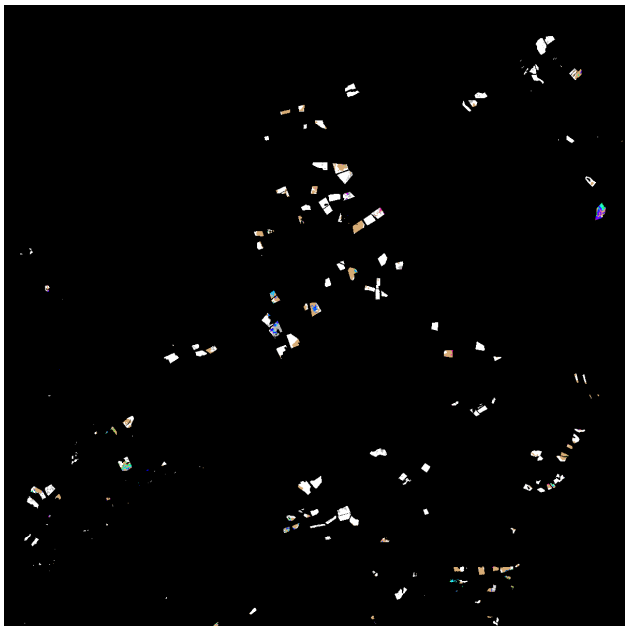

Figure 7. STL-map of the $r2$ pattern, for areas identified as pastures by the DSLCC, in the region of *la Plaine des Cafres*.

Among the 40 patterns automatically extracted from B8, 7 can be manually identified as being able to detect such areas. This is reasonable since vegetation is present in this pond region. Pattern $3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2$ is a good example. Its STL-map for the whole 22 km $\times$ 17 km is presented with Figure 10. Please note that it is not restricted to actual water areas or areas identified as such by the DSLCC. As evidenced, the inland water of the *Réserve naturelle nationale de l'étang de Saint-Paul* as well as other vegetated areas are well exhibited while maritime waters are not detected. The pattern extraction parameters as well as the redescription extraction parameters can thus be questioned. For example, more patterns could be considered to characterize each band and/or a lower redescription support could be envisaged. If the latter option is chosen, and if the minimum number of pixels supporting a decision is set to 1% instead of 5%, then 5 redescriptions are extracted. Two of them rely on the inland water patterns identified manually, and pattern $3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$ is one of them. It is thus possible to relate inland water decisions with patterns, as long as the support is correctly set. Interestingly, no such patterns are reported for NDVI while its attention weights are higher than those of B8.

According to attention weights, class 'Rocks' mainly mobilizes visible and near-infrared channels. Rocks are indeed less likely to be detected by vegetation nor water indices. They are also not likely to change that much, which is confirmed by available redescription patterns. They indeed indicate that reflectance levels are quite stable. Though symbol '1' can occur in a row, these levels are mainly medium and high ones for B2, B3, and B4 (symbols '2' and '3') while they are lower ones for B8, NDVI, and NDWI. (symbols '1' and '2'). Regarding the redescriptions provided for class 'Urban area', quite stable evolutions based on medium or high reflectance values are extracted for B2, B3, B4. Contrarily, an increase from symbol to '1' to '2' is reported for NDVI and NDWI. NDWI also provides a redescription pattern showing stability at low levels. Band B8 unveils more complex evolutions, i.e., stable ones with series of symbol '2' as well as decreases, either from symbol '2' to symbol '1' or from symbol '3' to symbol '2'. This diversity is

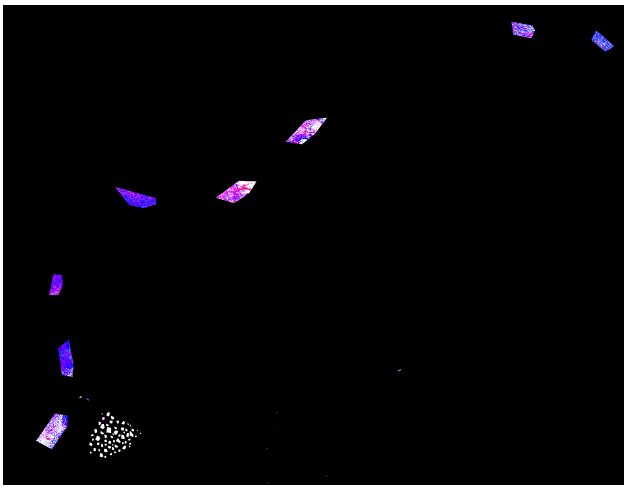

Figure 8. STL-map of the $r3$ pattern for ground truth areas covered by maritime waters, between the towns of Saint-Paul and Saint-Denis.

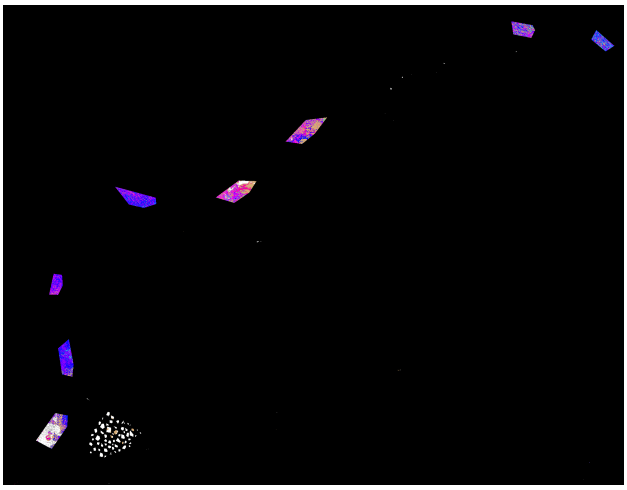

Figure 9. STL-map of the $r3$ pattern for areas identified as waters by the DSLCC, between the towns of Saint-Paul and Saint-Denis.

exploited by the DSLCC since it mainly relies on B8 according to attention weights (see Figure 1).

6. CONCLUSION

An original approach for explaining the behavior of a deep spatiotemporal land cover classifier is presented in this paper. It unveils, for each channel and each class, the intepretable spatiotemporal features captured at the pixel level by the last convolutional layer. These features are extracted using dedicated data mining patterns and matched with the different neuronal activation level configurations using redescription mining. In order to demonstrate the feasibility of the proposed approach, a simple architecture applying spatiotemporal and and temporal convolutions for each channel separately is considered for the deep classifier. As shown by experiments on a Sentinel-2 time series, such explanations are meaningful and can enrich channel attention-based explanations. Nevertheless, though indeed captured by the network, it is yet unclear to which extent they do account for the final decisions. Masking methods are for example envisaged to assess this crucial aspect. Moreover, since some expected redescription-based explanations are missing and

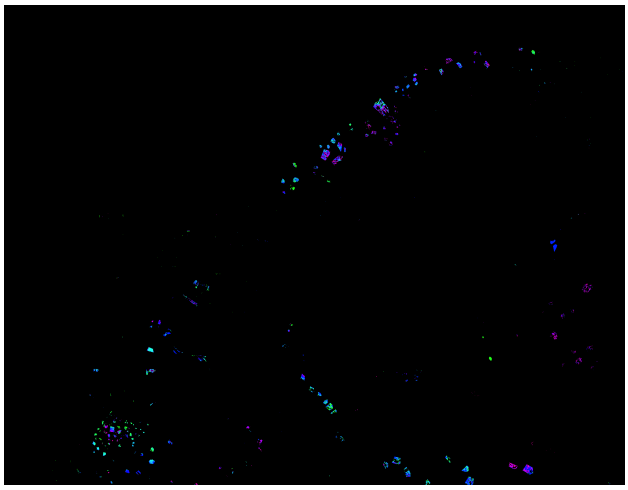

Figure 10. STL-map of pattern $3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$, between the towns of Saint-Paul and Saint-Denis.

decisions are not always covered by redescriptions, the selection of the patterns and the redescriptions can also be questioned. In particular, the number of most promising GFS-patterns describing the pixels could be higher and the support of the redescriptions could be lower. A method for setting them automatically would be of interest. More complex redescriptions, e.g., with more activation level variables and disjunctions could also be extracted. In addition, it would be of interest to check whether the presence/absence of patterns could help in understanding classifier errors or not. Finally, other layers such as the final dense one should be studied, and other interpretable features could be considered, in particular the WECS (Wavelet Energies Correlation Screening), a wavelet-based change measure recently proposed by (Fonseca et al., 2021).

REFERENCES

Campos-Taberner, M., Haro, F., Martinez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., Gilabert, M., 2020. Understanding deep learning in land use classification based on Sentinel-2 time series. *Scientific Reports*, 10.

Censi, A. M., Ienco, D., Gbodjo, Y. J. E., Pensa, R. G., Interdonato, R., Gaetano, R., 2021. Attentive Spatial Temporal Graph CNN for Land Cover Mapping From Multi Temporal Remote Sensing Data. *IEEE Access*, 9, 23070-23082.

Courteille, H., Benoît, A., Méger, N., Atto, A. M., Ienco, D., 2021. Channel-based attention for land cover classification using sentinel-2 time series. *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2021, Brussels, Belgium, July 11-16, 2021*, IEEE, 1077–1080.

Dupuy, S., Gaetano, R., Le Mézo, L., 2020. Mapping land cover on Reunion Island in 2017 using satellite imagery and geospatial ground data. *Data in Brief*, 28, 104934.

Fonseca, R., Pinheiro, A., Atto, A., 2021. Wavelet spatio-temporal change detection on multi-temporal polsar images.

Galbrun, E., Miettinen, P., 2012a. From black and white to full color: extending redescription mining outside the Boolean world. *Statistical Analysis and Data Mining*, 5(4), 284-303.

Galbrun, E., Miettinen, P., 2012b. Siren: An Interactive Tool for Mining and Visualizing Geospatial Redescriptions . *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12*, Beijing, China.

Galbrun, E., Miettinen, P., 2017. *Redescription Mining*. SpringerBriefs in Computer Science, Springer, Cham.

Garnot, V. S. F., Landrieu, L., Giordano, S., Chehata, N., 2020. Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention. *CVPR 2020*, CVPR, Seattle, United States.

Ienco, D., Gaetano, R., Dupaquier, C., Maurel, P., 2017. Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10), 1685-1689.

Ienco, D., Gbodjo, Y. J. E., Gaetano, R., Interdonato, R., 2020. Weakly Supervised Learning for Land Cover Mapping of Satellite Image Time Series via Attention-Based CNN. *IEEE Access*, 8, 179547-179560.

Julea, A., Méger, N., Bolon, P., Rigotti, C., Doin, M.-P., Lasserre, C., Trouvé, E., Lazarescu, V. N., 2011. Unsupervised Spatiotemporal Mining of Satellite Image Time Series Using Grouped Frequent Sequential Patterns. *IEEE Transactions on Geoscience and Remote Sensing*, 49(4), pp.1417-1430.

Lundberg, S. M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 4765–4774.

McFeeters, S. K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425-1432.

Méger, N., Rigotti, C., Pothier, C., Nguyen, T., Lodge, F., Gueguen, L., Andréoli, R., Doin, M.-P., Datcu, M., 2019. Ranking evolution maps for Satellite Image Time Series exploration: application to crustal deformation and environmental monitoring. *Data Mining and Knowledge Discovery*, 33(1), 131-167.

Nguyen, T., Méger, N., Rigotti, C., Pothier, C., Gourmelen, N., Trouvé, E., 2019. A pattern-based mining system for exploring Displacement Field Time Series. *19th IEEE International Conference on Data Mining (ICDM) Demo*, IEEE, Beijing, China, 1110–1113.

Pelletier, C., Webb, G., Petitjean, F., 2019. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5), 523. http://dx.doi.org/10.3390/rs11050523.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical Satellite Time Series Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169, 421 - 435.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618–626.

Vali, A., Comai, S., Matteucci, M., 2020. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sensing*, 12(15).