BENCHMARKING OF CONVOLUTIONAL NEURAL NETWORK APPROACHES FOR VEGETATION LAND COVER MAPPING

Carpentier Benjamin, Masse Antoine, Lavergne Emeric, Sannier Christophe

CLS, Parc Technologique du Canal, 11 Rue Hermès, 31520 Ramonville-Saint-Agne, France

KEY WORDS: Satellite image time series, Deep learning, Convolutions, Classification, Land cover map, Sparse annotations

ABSTRACT:

Satellite Image Time Series (SITS) are becoming available at high spatial, spectral and temporal resolutions across the globe by the latest remote sensing sensors. These series of images can be highly valuable when exploited by classification systems to produce frequently updated and accurate land cover maps. The richness of spectral, spatial and temporal features in SITS is a promising source of data for developing better classification algorithms. However, machine learning methods such as Random Forests (RF), despite their fruitful application to SITS to produce land cover maps, are structurally unable to properly handle intertwined spatial, spectral and temporal dynamics without breaking the structure of the data. Therefore, the present work proposes a comparative study of various deep learning algorithms from the Convolutional Neural Network (CNN) family and evaluate their performance on SITS classification. They are compared to the processing chain coined *iota*^{2}, developed by the CESBIO and based on a RF model. Experiments are carried out in an operational context using with sparse annotations from 290 labeled polygons. Less than 80 000 pixel time series belonging to 8 land cover classes from a year of Sentinel-2 monthly syntheses are used. Results show on a test set of 131 polygons that CNNs using 3D convolutions in space and time are more accurate than 1D temporal, stacked 2D and RF approaches. Best-performing models are CNNs using spatio-temporal features, namely *3D-CNN*, *2D-CNN* and *SpatioTempCNN*, a two-stream model using both 1D and 3D convolutions.

1. INTRODUCTION

1.1 Context

Land cover maps provide spatial information on the variety of different types, or classes, covering the Earth's surface. Such maps were originally produced by using only spectral features available in satellite images sensed by Earth observation systems. However, some land cover classes, despite their characteristic spectral signatures, remain difficult to classify with a lack of spatial and temporal information. In order to make these maps available on time, accurate, robust, and reliable, automatic methods need to better handle multidimensional data such as spectral, spatial and temporal domains.

The present paper will benchmark various CNN models to produce land cover maps from SITS, defined as a sequence of satellite images of the same scene taken at subsequent times. As the Earth's surface is rapidly changing due to natural and socioeconomic factors, land cover maps are an essential tool for mapping and monitoring its biophysical cover. They are highly valuable in many applications such as urbanization, natural resources management, and during extreme events accentuated by climate change such as drought, flooding, wildfires or biomass changes. Indeed, SITS can provide detailed information on the status and evolution dynamics of different land cover classes and hence make possible leveraging class-specific spectro-temporal profiles to improve the classification. However, the majority of land cover maps are still only relying on spectral information or as in recent studies, in spectral and spatial information. Consequently, the use of temporal dependencies has been poorly investigated as explained in (Gómez et al., 2016) and (Gbodjo et al., 2020).

1.2 Related work

Due to the recent availability of SITS and their increasing spatial and temporal resolutions, an array of new methods to better handle multidimensional data takes form in the literature. Among them, CNNs are the most widely used and frequently beat state-of-the-art approaches from machine learning such as RF.

The present work will restrict its investigation to the benchmark of CNNs on SITS land cover classification even though other deep learning architectures have proven to be successful recently, such as recurrent-based models in (Rußwurm and Körner, 2018) with convolutional LSTM cells, or attention-based models as in (Rußwurm and Körner, 2019). We focused on CNNs for their (i) relative ease of training compared to recurrent models, (ii) ease of deployment in an operational production framework, and (iii) ability to efficiently blend spatial and temporal information in convolution kernels. Moreover, recent experiments in (Garnot et al., 2019) on the respective performance of recurrent, convolution and hybrid models show that best-performing approaches are reached when up to 90 % of the models' parameters are allocated to modelling the temporal dimension of the data, suggesting that simple convolutional architectures are well-suited and probably sufficient for SITS classification.

One important aspect of this work is that the reference data is only sparsely labeled, meaning that only a a small subset of pixels is labeled in each training instance, leading to multiple issues with patch-based approaches such as spatial CNNs. Since only a fraction of pixels of the training images is labeled, much of the geometric information is simply not present in the data at first. Class borders pixels and spatial arrangements between classes are rarely annotated and, in our case, where labeled pixels are localised using small polygons within larger class

objects, they are totally absent of the data. Geometric degradation such as smoothing of corners and erosion or dilation of small elements in the classified map is a well-known drawback induced by CNNs. Unfortunately, this phenomenon is accentuated when using sparse annotations. Indeed, since the training data drive both the feature extraction and the classification steps, the learning of rich patterns is impossible if the data is not rich enough.

Most conventional methods that try to incorporate the temporal dynamics of the data heavily rely on either on a simple temporal stacking in the channel dimension or on handcrafted feature descriptors. While a straightforward stacking of time acquisitions is oblivious to the temporal structure and causality present in the first place, feature engineering is based on domain knowledge and may fail to capture the relevant part of the raw data. In the meantime, there is a strong need to leverage simultaneously spatial and temporal features to perform land cover classification, preferably jointly learnt to take the most out of the feature interplay that guides the dynamics of SITS.

Time series could help handle intra-class variability across time, which is one of the major aspect of land cover mapping that plummets its performance. Obviously, deep learning made much more advances in recent years than traditional methods. Especially, CNNs are promising candidates to address the task of SITS land cover classification.

1.3 Goal

The present paper proposes a benchmark of different CNN-based approaches and a RF classifier on SITS land cover classification from sparsely annotated data. A special focus is put on the handling of spectral, spatial and temporal information. The benefit of the proposed methods will be assessed by comparing evaluation metrics on a separate test set. This has been divided into two sub-goals:

1. Exploiting spatial and temporal dependencies

   Most land cover classification systems rely on spectral features and lack spatial or temporal information. Indeed, while a temporal stack contains temporal statistics, it does not model the sequential nature of the data. Indeed, shuffling the temporal order has no consequence on the model and results. We aim to fill this gap, especially for land cover classes that vary over space and time and which are hence prone to misclassification. We will study the ability for different CNNs to extract relevant temporal and spatial features in SITS to better classify them. Indeed, such features may help discriminate between certain land cover classes which may have similar spectral signatures at some point in time and being radically different at a later time, especially vegetation.

2. Operational solution for real-world applications

   Since data is scarce and costly in operational works, we aim to propose a scalable solution for real world applications defined by large areas and a little amount of sparsely annotated ground truth data. Such data is indeed expensive since it is provided by photo-interpreters that manually label it. Additionally, the solution is expected to be computationally light and feasible.

2. METHODS

2.1 Random Forest (RF)

State-of-the-art methods extensively use machine learning to perform land cover classification. Methods such as RF classifiers are commonly found in the literature and is used herein as a baseline model. Particularly, we mention the solution *iota*^{2} developed by (Inglada et al., 2017) at the CESBIO. Briefly, a RF is an ensemble of decision trees, acyclic graphs that can be used to make decisions. In each node of the graph, a given feature in the input feature vector is submitted to a binary question. In this way, one can construct trees of various depths that are used to classify a given input feature vector.

To account for the temporal domain, temporal stacking in the channel dimension is performed. Therefore, inputs to the model are time series vectors of length $c \cdot T$ where c the number of spectral bands and T the number of time acquisitions. As for any pixel-based approach and without the addition of spatial features beforehand, this algorithm sees each pixel irrespective of its spatial neighborhood. Moreover, since the temporal ordering has no effect on the classification process, this type of classifier cannot properly leverage temporal dynamics and causality present in the data.

Two versions of this model are proposed: a basic one denoted by *iota*^{2}*_base* which only relies on spectral input features, and a spatially-aware one denoted by *iota*^{2}*_ctx*, developed by (Derksen et al., 2020) at the CESBIO, which is enhanced by a prior object spatial segmentation on which spatial features are computed and stacked to the spectral ones. This additional feature engineering step that we decided to avoid in our proposed deep learning methods introduces more complexity to the model but provides it with valuable spatial information. Both RFs are implemented on CPUs using the same set of hyperparameters: minimum number of samples in each node of 20, maximum depth of 50, maximum number of trees of 100. Other parameters are set to default values as described in Orfeotoolbox documentation (Grizonnet et al., 2017).

2.2 Deep learning

Deep learning models are increasingly used to perform land cover classification. Especially, given the filtering nature of convolution kernels, CNNs can be applied to extract relevant spectral, spatial and temporal features from data. This ability to handle multidimensional data makes them promising candidates to produce more accurate land cover maps from SITS.

2.2.1 1D CNN 1D CNNs have been used on the spectral (Hu et al., 2015) and temporal (Pelletier et al., 2019a) dimension. Given the filtering properties of convolution kernels, they are very appropriate to handle temporal information. For instance, 1D convolutions are typically applied to sequential data such as sentences in natural language processing, audio signals, and more generally to any structured one-dimensional signal like time series. Figure 1 shows an example of a 1D convolution filter sliding through a pixel time series of a particular spectral band.

CNNs using 1D convolutions have been used for land cover classification of SITS as in (Pelletier et al., 2019b) with a model coined *TempCNN*. This model performs convolutions through pixel time series. Hence, no spatial information is taken into account but it shows competitive results due to its ability to handle

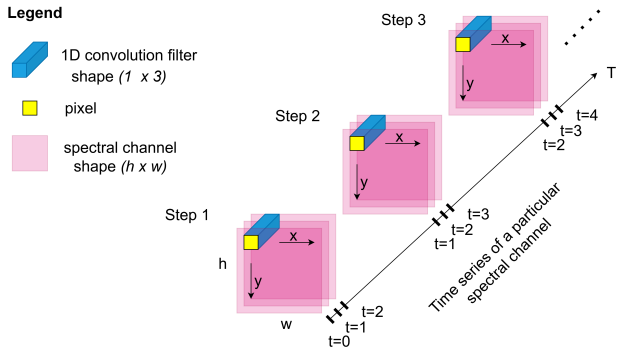

Figure 1. Representation of a 1D convolution on a spectral channel. Only the three first sliding steps are depicted.

temporal dynamics. Inputs to the model are tensors of shape $c \times T$. Figure 2 from the original paper shows an overview of the model.

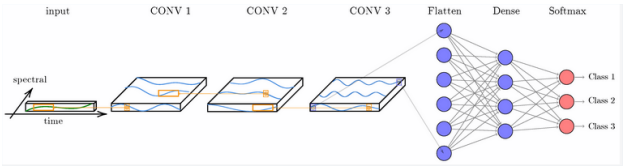

Figure 2. *TempCNN* model (Pelletier et al., 2019b)

2.2.2 2D CNN In 2D CNNs, spatial convolutions are applied in both x and y directions to extract relevant spatial features from images. More sophisticated architectures try to leverage multi-scale spatial information by downsizing input feature maps at subsequent stages in the network as a Fully Convolutional Network in (Maggiori et al., 2017) and the well-known U-net architecture in (Stoian et al., 2019). Yet, since our input training images have a size of 32×32, this approach is not conceivable as most of the information would be lost in the early layers. To account for the temporal dimension, the use of standard 2D convolutions is not straightforward and workarounds such as temporal stacking in the channel dimension are needed (Kussul et al., 2017). Consequently, the input data becomes of shape $c \cdot T \times h \times w$ where c denotes the spectral channels and h and w denote image height and width. Figure 3 shows an example of such 2D convolution with temporal stacking. This solution does not fully model the temporal dynamics that exist in the data since retraining the model with a different temporal order would statistically provide similar results.

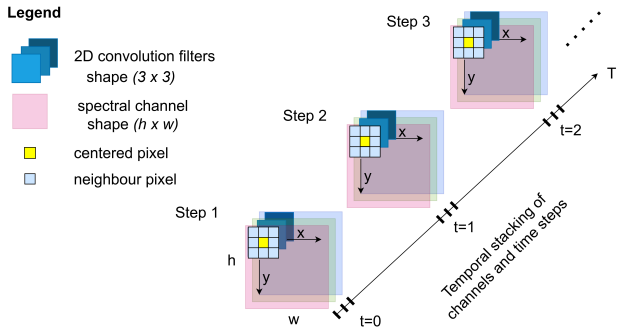

Figure 3. Representation of a 2D convolution using temporal stacking. Only the three first sliding steps are depicted.

The architecture of the model is similar to that of Figure 5 except that 2D convolution kernels are used and inputs at subsequent time steps are stacked along the channel dimension. The forward pass comprises a series of convolutional layers followed by batch normalization (BatchNorm), Rectified Linear Unit (ReLu) activations and Dropout layers. The last two layers are fully convolutional layers instead of fully connected ones since they permit filters to remain spatially invariant and allow to keep the input image shape. The model outputs a tensor of shape $K \times h \times w$ where K is the number of classes. Each of the K channel is the probability for a given pixel in $h \times w$ to belong to class K.

2.2.3 3D CNN Since 3D convolutions allow to convolve in more dimensions, 3D CNNs have been tried across spatial and spectral domains (Ben Hamida et al., 2018) and across spatial and temporal ones for crop classification (Ji et al., 2018). Figure 4 shows an example of a 3D convolution filter applied in both spatial and temporal dimensions.

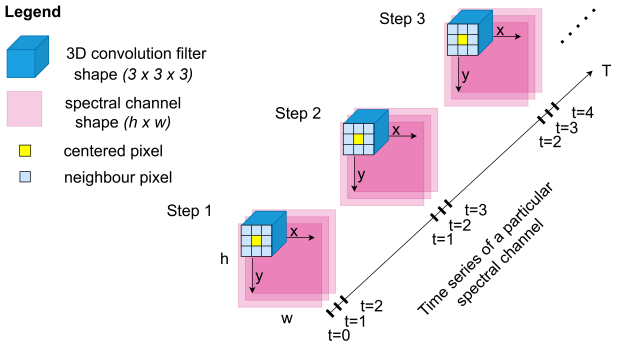

Figure 4. Representation of a 3D convolution on a spectral channel. Only the three first sliding steps are depicted.

CNNs using 3D convolutions are rare in the literature. By using cubes of convolutions of shape $i \times j \times k$, they are able to operate on temporal and spatial dimensions simultaneously. A 3-dimensional convolution filter uses all input channels and slides along three dimensions. Inputs to the model are 4D tensors of shape $c \times T \times h \times w$. We chose convolution filters of shape $3 \times 3 \times 3$, which means a temporal extent of 3 and a spatial window of 3×3 pixels. Figure 5 details the architecture of the model.

2.2.4 Two-stream models Two-stream models are often used to extract two different types of features (e.g. spatial and temporal ones) by using two models in parallel and combining their respective feature vectors in a single one as in (Benedetti et al., 2018), (Interdonato et al., 2019). In this work, we propose a two stream model to benefit from both patch-based and pixel-based approaches.

Patch-based methods such as spatial CNNs inevitably loose geometric precision on the classified map since any pixel prediction takes into account its direct neighborhood, for example in a 3×3 window. Therefore, small spatial features can be totally erased and large ones dilated in the classification output. To tackle this issue and inspired from ensembling methods, a hybrid model is proposed. Often seen in the literature in order to merge temporal and spatial features, we propose to combine our two best performing models into a single one to balance the respective disadvantages of pixel-based and patch-based models. Figure 6 depicts this model ensembling. The output prediction is controlled by a weighting trade-off parameter set manually.

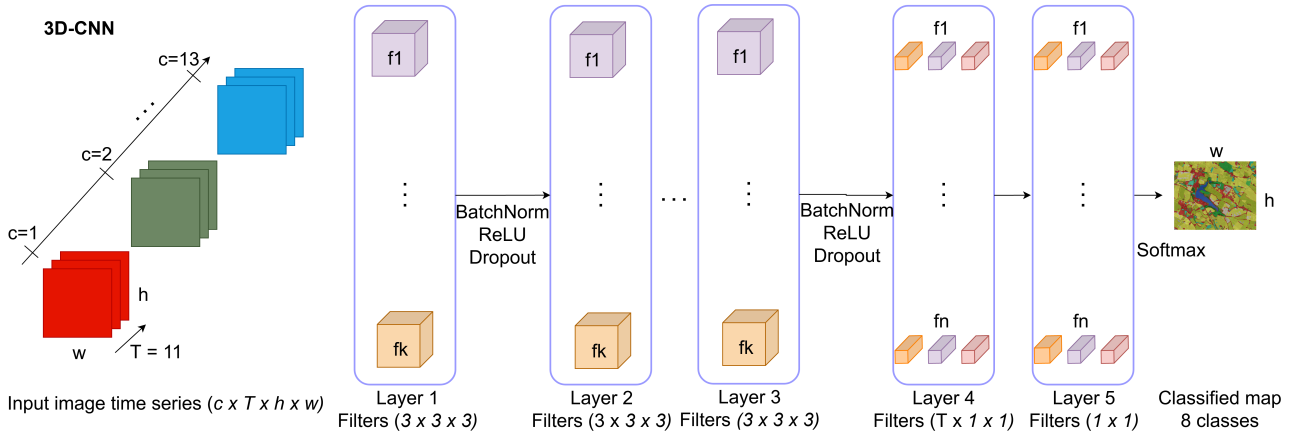

Figure 5. *3D-CNN* model.

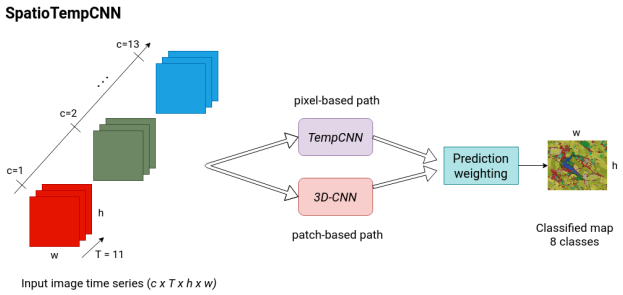

Figure 6. SpatioTempCNN: a two-stream model.

3. BENCHMARK

This section first describes the data used in this paper and describes the aforementioned benchmark.

3.1 Training data

Our dataset is composed of 11 dates of Sentinel-2 images of the tile ID 31TCJ processed at level L3A, which are monthly syntheses produced from their L2A counterparts acquired every 5 days. The time series span monthly from February to December 2018 as depicted in Table 1. The tile 31TCJ covers an area of $110\,\mathrm{km} \times 110\,\mathrm{km}$ and is located near Toulouse, France.

Table 1. Span of the L3A tile time series.

Timestep	Date
1	2018-02-15
2	2018-03-15
3	2018-04-15
4	2018-05-15
5	2018-06-15
6	2018-07-08
7	2018-08-15
8	2018-09-15
9	2018-10-15
10	2018-11-15
11	2018-12-15

Each training instance consists of a pair of an image time series of shape $(c \times T \times h \times w)$ and a binary label mask indicating the presence or absence of labeled pixels. The label masks are created by extracting patches around each labeled polygons. Information about ground truth data collection is provided in the next section.

3.2 Ground truth data

The labeled polygons have been collected by trained photo-interpreters on satellite imagery of the tile 31TCJ and reviewed by experts. The polygon sampling strategy, crucial for accuracy assessment, is let to the expert's knowledge of the area. Since all classes are not uniformly distributed over large regions, the strategy must account for the specificity of the terrain and the distribution of classes in order to avoid accentuating class imbalance. Table 2 describes the class nomenclature which comprises 8 land cover classes.

Table 2. Ground truth dataset repartition using CLC+ nomenclature.

Code	Class name	# labeled pixels
1	Built up	9528
2	Woodland needle-leaved trees	4532
3	Woodland broad-leaved trees	11512
5	Shrubland	672
6	Permanent herbaceous land	4132
7	Periodically herbaceous land	11875
10	Non-vegetated land	1012
11	Water	9444

The best way to unsure a fair and correlation-free strategy consists in splitting the training and validation sets at the polygon level rather than the pixel level. Indeed, pixels extracted from the same polygons and found in both sets are more likely to be similar and have higher auto-correlation than pixels from separate polygons. Therefore, the dataset is split between a training and a validation set while ensuring a balanced class distribution. The training data accounts for 70 % of the polygons and the remaining 30 % are kept separate for the validation during which the performance metrics are computed.

In terms of pixel counts inside the polygons, the data is unbalanced with two minority classes: *Shrubland* and *Non_vegetated*. A straightforward oversampling strategy during training is chosen to alleviate this class imbalance issue.

3.3 Spectral features

Pixel time series are composed of 13 spectral values (10 bands and 3 spectral indices). We used the Sentinel-2 $10\,\mathrm{m}$ bands (B2, B3, B4, B8) and $20\,\mathrm{m}$ bands (B5, B6, B7, B8A, B11, B12) resampled to $10\,\mathrm{m}$. The three following spectral indices are

added. The NDVI is designed for vegetation detection and is defined by

$$NDVI = \frac{B8 - B4}{B8 + B4} \quad (1)$$

Likewise, the NDWI as defined by McFeeters in 1996 in (McFeeters, 1996) is used to perceive changes in water bodies and is defined by

$$NDWI = \frac{B3 - B8}{B3 + B8} \quad (2)$$

Lastly, the BI is sensitive to the brightness of soils where high soil brightness is linked with soil humidity and presence of salts. It is defined by

$$BI = \sqrt{\frac{B4^2 + B3^2}{2}} \quad (3)$$

While an increasing amount of research experiments shows that adding handcrafted spectral indices may be useless when training deep learning models as in (Pelletier et al., 2019b), we do not assess their usefulness in the present work.

3.4 Spatial context and temporal dynamics

Our hypothesis states that the extraction of combined spectral, temporal and spatial features is a key factor when analyzing SITS. While state-of-the-art approaches focus on either one or two of these dimensions, only too few have investigated to do it all at once in an end-to-end fashion. Indeed, spectral information only is sometimes insufficient to identify certain land cover classes that are similar at a particular time.

3.4.1 Spatial context Pixel-based approaches suffer from this lack of information since they consider each pixel irrespective of their spatial context.

3.4.2 Temporal dynamics Figure 7 shows examples of time series for the three spectral indices. One can quickly notice general trends that correspond to what we should expect. For instance, the NDWI curve, which is used to detect water bodies, clearly separates it from other classes (Figure 7, middle). On the contrary, BI shows important variations for some classes depending the month of the year (Figure 7, right).

Additionally, some classes can look similar in terms of spectral signatures at a given moment in time while being totally different at a later time. Figure 8 shows that particular spectral features patterns are very characteristic for certain classes at different times.

As expected, vegetation classes such as *Woodland needle-leaved trees*, *Permanent herbaceous land* and *Shrubland* show a high normalized NDVI (green starry curve). Especially, infrared bands (B6, B7, B8, B8A) clearly show the expected variations along the year for the *Periodically herbaceous land* class: a steady increase during winter and mid-spring followed by a rapid decrease during summer when the harvest season comes. This observation is in accordance with the vegetation phenology. We can notice that most of the *Built-up* spectral features

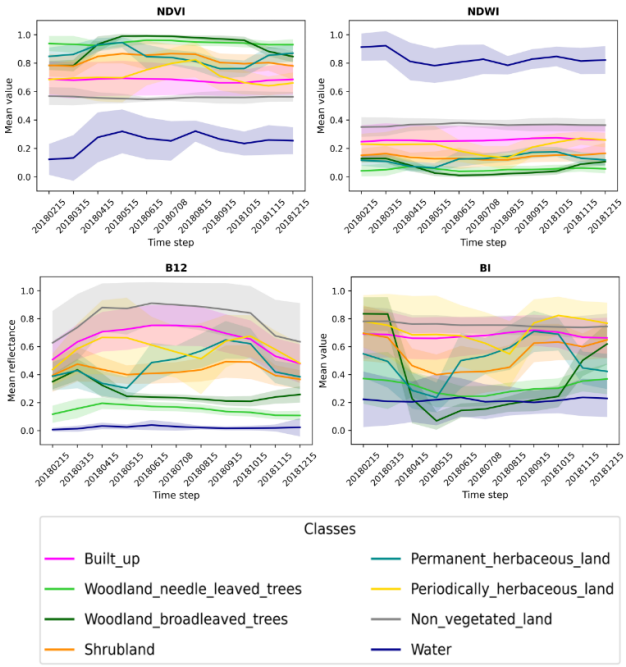

Figure 7. Class separability over time shown for three spectral indices and band B12. Intra-class variability computed using class standard deviations is represented in colored areas around each curve.

are close and follow similar trends, except for NDWI. This particular pattern, which could make more difficult to discriminate this class, is related its inherent heterogeneity. Indeed, this class often contains different features such as buildings, trees, grassland and roads.

In order to enrich classification systems, there is a strong need to incorporate both spatial and temporal information that may help discriminate between classes. Consequently, our approaches will focus in adding these valuable insights to the classification process. We believe that intertwined feature modeling can have a high potential for the leverage of relevant information to improve land cover classification systems.

3.5 Methodology

A comparative research methodology using a benchmark is adopted in this paper. It is facilitated by a generic training and validation of deep learning models that use sparsely annotated data from labeled polygons. This framework and all proposed models are coded in Python with the deep learning library PyTorch (Paszke et al., 2017).

3.5.1 Training As with all deep learning algorithms, training occurs over multiple repetitions, or epochs, of some optimization procedure. The training session stops either when the number of epochs is filled or when the optimization has converged according to some stop criterion. The latter option is chosen in this project using an early stopping regularization mechanism monitoring the F1-score macro averaged over all classes on the validation set using a patience parameter of 5 epochs.

3.5.2 Validation After each training epoch, the model's performance is measured on the validation set using an array of evaluation metrics such as OA, F1-scores for each class, F1-score macro averaged over all classes, confusion matrices, or

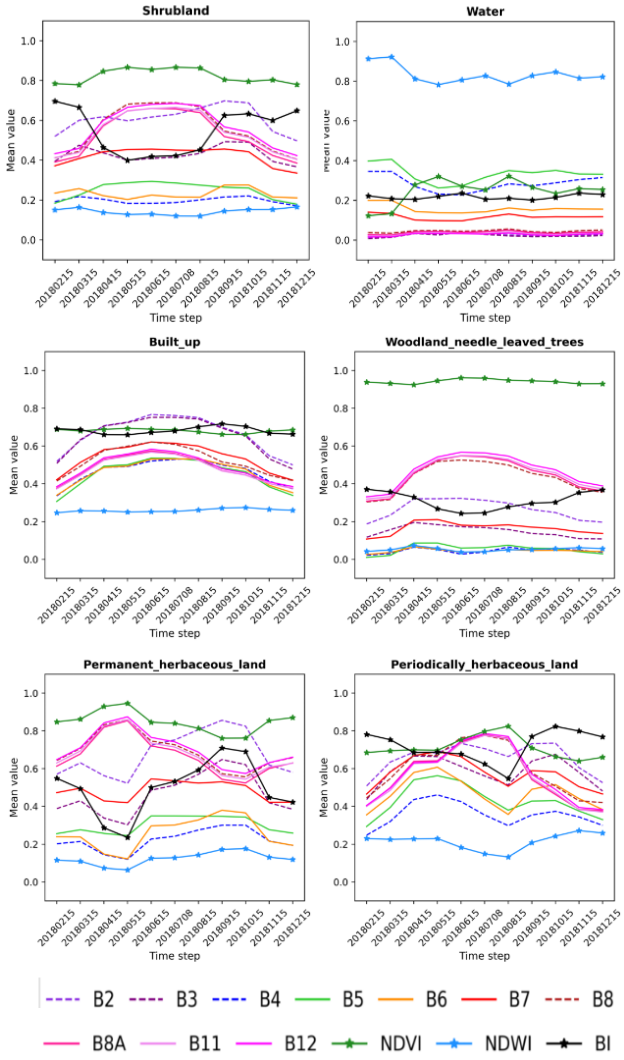

Figure 8. Examples of normalized spectral features time series per class.

the model loss. We used the macro-averaged F1-score across all classes as criterion of early stopping. The macro-average takes into account class imbalance by assigning the same weight to each class, irrespective of their population size. It is widely used to assess multi-class classification results and is defined as the geometric average of precision and recall. Since we perform 5 training sessions for each model to ensure statistical reliability, the best model among these trials is again chosen according to the evaluation metrics.

3.5.3 Testing All proposed models are eventually assessed using a separate test set of 131 polygons whose repartition is shown in Table 3. The test set is labeled by an external photo-interpreter to minimize any bias in the labelling procedure between training and test sets.

3.6 Results and Analysis

Class F1-scores on the validation and test sets for the benchmark models are shown in Table 4. Test results show that *3D-CNN* is the best performing model with a mean F1-score of 0.804. It is followed by *2D-CNN*, *Spatio-TempCNN*, *iota*^{2}*_ctx*, *TempCNN* and *iota*^{2}*_base* with F1-scores of 0.799, 0.798, 0.753, 0.750 and 0.723 respectively.

Table 3. Test set repartition using CLC+ nomenclature.

Code	Class name	# labeled pixels
1	Built up	154556
2	Woodland needle-leaved trees	13768
3	Woodland broad-leaved trees	35200
5	Shrubland	3887
6	Permanent herbaceous land	4504
7	Periodically herbaceous land	32546
10	Non-vegetated land	2130
11	Water	19311

Statistical reliability and efficiency results are also provided in Table 4. Training and inference time are measured using two NVIDIA Tesla V100 GPUs. This experiment proves that the performance of the models are reliable and robust as shown by the standard deviation values. Yet, the small sized dataset and class imbalance in both training and validation sets may limit this reliability assessment.

Besides evaluation metrics, a visual inspection of classification maps reveals interesting properties of each model. Pixel-based models such as *iota*^{2}*_base* and 1D temporal *TempCNN* tend to produce more speckle-like noise in classification maps as expected since they are oblivious to spatial context. On the contrary, patch-based models like *3D-CNN* produce more spatially coherent maps with homogeneously classified areas. Figure 9 shows an example of classified area by different models. This example shows one of the recurring error we observed: RF-based models (top row) tend to wrongly classify *Periodically herbaceous land* (crop fields) areas as *Built up* in red or as *Non-vegetated* in grey. This is mostly due to the fact that crop fields vary much over time and can be spectrally similar to other classes over the year. We also observe that CNN models produce less speckle-like labelling. Roads, paths and class borders are often labeled as *Built-up*. As mentioned earlier, small features are often erased in the patch-based *3D-CNN* as individual pixel features are smoothed out in the kernel window. On the other hand, certain features can be dilated. To help combat these issues, one would need to enrich the dataset using more dense polygons where mixed pixels in class borders are available during training.

4. CONCLUSIONS AND PERSPECTIVES

This paper introduced a benchmark of CNNs and compared their classification performance on multi-class land cover classification of SITS using sparse annotations. The proposed models are compared to a RF-based approach with and without prior spatial segmentation. A sparsely annotated ground truth dataset constituted by 290 polygons (less than 80 000 labeled time series) belonging to 8 land cover classes is used and sampled from the Sentinel-2 tile 31TCJ near Toulouse. All models are eventually evaluated on a separate test set of 131 polygons.

Results show that *3D-CNN*, a spatio-temporal CNN using 3D convolutions is the best performing model of the benchmark with a mean F1 score of 0.804 on a test set. On this set, all CNNs exploiting spatio-temporal features outperform RF-based approaches. Therefore, a proper exploitation of the spatial context and temporal dynamics in satellite images appears as a powerful lever arm to improve land cover maps, especially for classes usually prone to misclassification such as vegetation ones. Moreover, given the honorable performance of the 1D

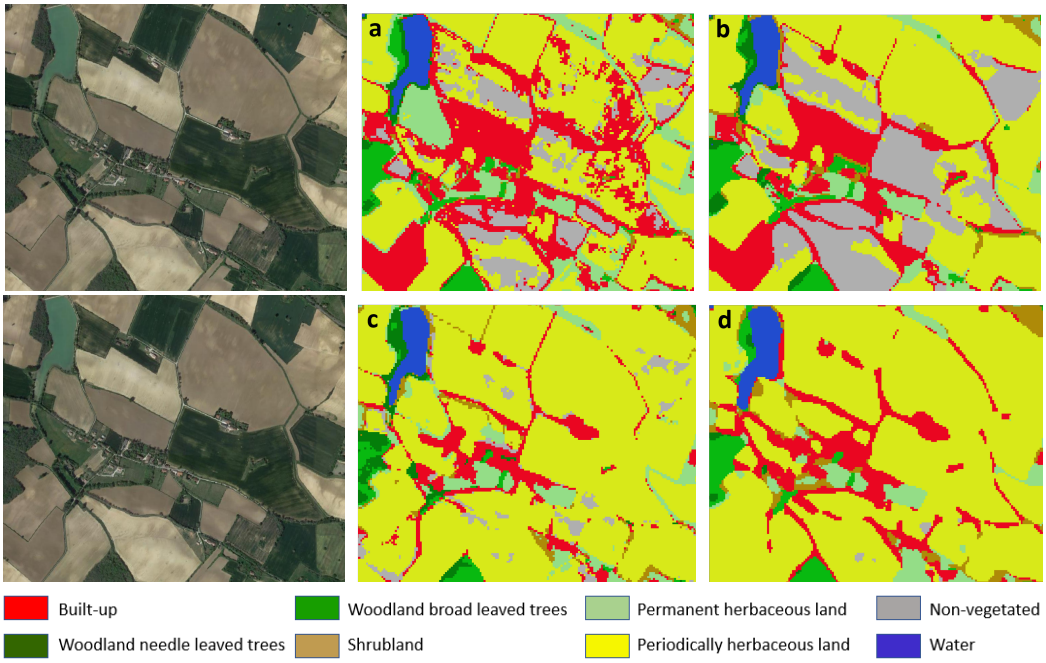

Figure 9. Example of a classified area with presence of crop fields: *iota*^{2}*_base* (a), *iota*^{2}*_ctx* (b), *TempCNN* (c), *3D-CNN* (d).

Table 4. Benchmark class F1-scores on validation and test sets. Model reliability is measured by mean F1-scores and standard deviations over 5 trainings. Model efficiency is measured by number of parameters, training and inference on the tile 31TCJ.

Class name	**TempCNN**	**3D-CNN**	**2D-CNN**	**Spatio-TempCNN**	*iota*^{2}*_ctx*	*iota*^{2}*_base*
	F1 scores on validation set					
Built-up	**0.973**	0.965	0.947	0.940	0.924	0.924
Woodland needle-leaved trees	0.967	0.997	**0.998**	0.987	0.990	0.842
Woodland broad-leaved trees	0.980	**0.996**	0.986	0.983	0.990	0.913
Shrubland	0.615	0.695	0.380	0.557	**0.719**	0.080
Permanent herbaceous land	**0.933**	0.900	0.936	**0.943**	0.932	0.799
Periodically herbaceous land	0.980	**0.985**	0.982	0.980	0.974	0.923
Non-vegetated land	0.807	0.653	0.674	0.600	0.576	**0.961**
Water	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Mean	**0.907**	0.899	0.863	0.874	0.888	0.680
	F1 scores on test set					
Built-up	0.984	**0.987**	0.981	0.984	0.968	0.962
Woodland needle-leaved trees	**0.492**	0.461	0.451	0.467	0.459	0.405
Woodland broad-leaved trees	0.863	0.490	0.867	**0.870**	0.861	0.854
Shrubland	0.682	**0.778**	0.619	0.694	0.627	0.586
Permanent herbaceous land	0.731	0.620	**0.753**	0.733	0.580	0.573
Periodically herbaceous land	0.962	0.939	0.966	**0.968**	0.947	0.948
Non-vegetated land	0.282	**0.774**	0.757	0.670	0.362	0.458
Water	**1.000**	**1.000**	**1.000**	**1.000**	0.999	0.998
Mean	0.750	**0.804**	0.799	0.798	0.753	0.723
	Efficiency and Reliability					
# parameters	493 k	**379 k**	463 k	867 k	NA	NA
Training (min)	1.0	**0.7**	1.0	3.0	+60	+40
Inference (min)	40	40	40	40	40	40
F1-score $\pm$ std on 5 trainings	0.89 ± 0.01	0.86 ± 0.03	0.82 ± 0.03	0.86 ± 0.02	NA	NA

temporal *TempCNN* which has no spatial information, it seems that temporal dynamics only can be determining.

Our initial hypothesis stating that combining spectral, spatial and temporal features contained in SITS would improve land cover classification systems is verified by the conducted experiments on the dataset at hand. In addition, the proposed models are light and meet the efficiency requirements needed in operational contexts for real-world applications. Besides, a generic training and validation framework for deep learning models has been developed for further research and development.

However, the little amount of labeled data is a serious impediment in deep learning models. In this regard, the proposed benchmark is currently being trained and evaluated using a richer and dense dataset of ground truth data on another area and first results clearly show the superiority of CNN-based approaches when massive amounts of data are available. Given the overall good but heterogeneous performance of CNNs, hybrid approaches such as the two-stream *SpatioTempCNN* will be studied using learning-based ensembling methods instead of rule-based ones.

REFERENCES

Ben Hamida, A., Benoit, A., Lambert, P., Ben Amar, C., 2018. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8), 4420–4434. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

Benedetti, P., Ienco, D., Gaetano, R., Osé, K., Pensa, R., Dupuy, S., 2018. M3Fusion: A Deep Learning Architecture for Multi-{Scale/Modal/Temporal} satellite data fusion. *arXiv:1803.01945 [cs]*. http://arxiv.org/abs/1803.01945. arXiv: 1803.01945.

Derksen, D., Inglada, J., Michel, J., 2020. Geometry Aware Evaluation of Handcrafted Superpixel-Based Features and Convolutional Neural Networks for Land Cover Mapping Using Satellite Imagery. *Remote Sensing*, 12(3). https://www.mdpi.com/2072-4292/12/3/513.

Garnot, V. S. F., Landrieu, L., Giordano, S., Chehata, N., 2019. Time-space tradeoff in deep learning models for crop classification on satellite multi-spectral image time series.

Gbodjo, Y. J. E., Ienco, D., Leroux, L., Interdonato, R., Gaetano, R., Ndao, B., 2020. Object-Based Multi-Temporal and Multi-Source Land Cover Mapping Leveraging Hierarchical Class Relationships. *Remote Sensing*, 12(17), 2814. https://www.mdpi.com/2072-4292/12/17/2814. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

Grizonnet, M., Michel, J., Poughon, V., Inglada, J., Savinaud, M., Cresson, R., 2017. Orfeo ToolBox: Open source processing of remote sensing images. *Open Geospatial Data, Software and Standards*, 2(1), 15.

Gómez, C., White, J. C., Wulder, M. A., 2016. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, 55–72.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sensing*, 7(11), 14680–14707. https://www.mdpi.com/2072-4292/7/11/14680.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sensing*, 9(1), 95. https://www.mdpi.com/2072-4292/9/1/95. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Interdonato, R., Ienco, D., Gaetano, R., Ose, K., 2019. DuPLO: A DUal view Point deep Learning architecture for time series classificatiOn. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149, 91–104.

Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y., 2018. 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images. *Remote Sensing*, 10(1), 75. https://www.mdpi.com/2072-4292/10/1/75. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778–782. Conference Name: IEEE Geoscience and Remote Sensing Letters.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional Neural Networks for Large-Scale Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55, 645-657. https://hal.inria.fr/hal-01369906.

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. https://doi.org/10.1080/01431169608948714.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. *NIPS-W*.

Pelletier, C., Webb, G. I., Petitjean, F., 2019a. Deep Learning for the Classification of Sentinel-2 Image Time Series. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 461–464. ISSN: 2153-7003.

Pelletier, C., Webb, G. I., Petitjean, F., 2019b. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5), 523. https://www.mdpi.com/2072-4292/11/5/523. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Rußwurm, M., Körner, M., 2018. Convolutional LSTMs for Cloud-Robust Segmentation of Remote Sensing Imagery. *CoRR*, abs/1811.02471. http://arxiv.org/abs/1811.02471.

Rußwurm, M., Körner, M., 2019. Self-Attention for Raw Optical Satellite Time Series Classification. *CoRR*, abs/1910.10536. http://arxiv.org/abs/1910.10536.

Stoian, A., Poulain, V., Inglada, J., Poughon, V., Derksen, D., 2019. Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems. *Remote Sensing*, 11(17), 1986. https://www.mdpi.com/2072-4292/11/17/1986. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.