MDPI

*Article*

# Representation learning for crop type classification of multispectral multitemporal remote sensing data

**Andrea González-Ramírez [1,*], Clement Atzberger [2], Deni Torres-Roman [1] and Josué López[2]**

[1] Center for Research and Advanced Studies of the National Polytechnic Institute, Telecommunications Group, Av del Bosque 1145, Zapopan 45017, Mexico
[2] Mantle Labs, Grünentorgasse 19, Vienna 1090, Austria
* Correspondence: andrea.gonzalez@cinvestav.mx

**Abstract:** Remote sensing (RS) spectral time series are a substantial source of information for earth monitoring tasks. Supervised deep learning algorithms with large number of training samples are usually used to develop accurate solutions for these applications. However, such approaches often face the lack of reliable labeled datasets. In addition, RS images acquired by optical sensors are frequently degraded by clouds/shadows, producing missing observations of an area of interest, creating irregular observation pattern. To address these issues, efforts have been made to implement frameworks that generate meaningful representations from the available data and alleviate the deficiencies of the data sources and supervised algorithms. Here, we propose a conceptually and computationally simple representation learning (RL) approach based on autoencoders (AEs) to generate informative and discriminative features for crop type classification using only a tiny set of reference samples. Our AEs architecture has very few parameters compared to other models proposed in the state-of-the-art, leading to a scalable model able to process very large areas in low computational time. The proposed methodology ensembles a set of single layer AEs with very limited number of neurons, each one trained with mono temporal spectral features of a set of samples per class. The reconstruction difference vector between input samples and its corresponding estimation are averaged over all cloud/shadows free temporal observations of a pixel location. This averaged reconstruction difference vector is the base for the representations. Experimental results show that the proposed extremely light-weight architecture indeed generates separable features for competitive performances in crop type classification. A simple classification fully connected network (FCN) was trained and tested with representations generated for the Sentinel-2 multispectral multitemporal dataset named BreizCrops. Our method reaches 74.44% overall accuracy which is 1% higher than a convolutional neural network OmniscCNN, and 5.56% lower than a transformer model, while our method is composed by $6.8k$ parameters, $\sim 400x$ less than the OmnicsCNN and $\sim 27x$ less than Transformers. These results prove that our method is competitive in classification performance compared with state-of-the-art methods while requiring much lower computational load.

**Keywords:** crop types; multispectral time series; autoencoder; representation learning; reconstruction

**Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:



## 1. Introduction

Monitoring and analysis of the Earth's surface using remote sensors has been increasingly used for crop surface studies, given the quantity and availability of spectral-temporal images. The multi-spectral time series from sensors such as Landsat or Sentinel-2 have provided very cost-effective technical support to achieve the reliable identification and monitoring of large cropping areas [1–4]. While a wide number of data sources and supervised classification algorithms have been used for crop mapping [3–11], limited efforts have been made in feature selection as well as the use of un- and self-supervised learning algorithms to alleviate scarcity of labels and missing data produced by clouds. Notable overviews and examples are provided in [12–16].

The use of multitemporal observations of multispectral data has a strong impact in crop type classification since the spectral difference in the crop growth over time are highlighted [1,13]. Each crop type has a distinct seasonal spectral behavior depending on local weather conditions [3]. Therefore, many researchers center their works on making use of multi-temporal information instead of using single acquisition [1,5,9,17].

The most common methods for crop type classification are based on supervised learning algorithms [18–22]. The aim of these algorithms is to train a discriminative model using labeled data. However, it is complicated to find tagged datasets, since it requires human accurate intervention, and normally datasets contain a huge amount of samples, which results in an expensive and time consuming task. Examples of supervised machine learning (ML) models include decision trees (DT) [23], random forest (RF) [24], support vector machine (SVM) [25] and artificial neural networks (ANN) [26]. The mentioned algorithms provide usually similar classification performance, but require extensive prepocessing steps such as compositing and gap-filling when incomplete (e.g., cloud-corrupted) time series are analyzed.

Unsupervised algorithms, such as autoencoders (AEs), mitigate the reliance on labeled datasets, although in principle these algorithms are not designed for the same purpose as supervised ones. An AE has as objective to compress data into a lower dimensional space, known as code, and then reconstruct the input [27]. The code is regarded to be a set of features, also called representations, which condense the necessary information to recover the original data [28].

AEs have been widely used as change detection methods by generating representations from the reconstruction difference of samples that belong to a particular probability distribution [29–31]. Moreover, representation learning (RL) is a broad subfield in machine learning, which is a set of techniques focused on automatically learning and identifying meaningful features from the input data. The field is closely related to the learning of low-dimensional manifolds within high-dimensional feature spaces [32–38].

In this work, we propose to train a light-weight deep learning model with individual time-tagged spectral signatures, while bypassing gap-filling and compositing methods. In our framework, we use an ensemble of AEs to generate new informative and discriminative features for crop type classification. Here, we arbitrarily chose one simple AE per class, any other number of AEs would also be possible. Instead of using the AE codes as representations, we calculate the respective reconstruction difference vector between the input and the output and use this concatenated reconstruction difference vector from the different AEs as representations. We evaluate the perfomance of the derived representations with a simple fully connected network (FCN) using the Breizhcrops dataset [11]. We compare the outcomes against a number of competing approaches using the same dataset.

### *1.1. Related work*

Russwurm et al. [11], propose a satellite image time series dataset for crop type mapping named BreizhCrops. They extract time series from Sentinel-2 at both pre-processing levels: top- (TOA) and bottom-of-atmosphere (BOA). They then use this dataset to benchmark a series of seven classification algorithms including RF and six deep learning methods based either on convolution, recurrence, or attention models for building a state-of-the-art benchmark on methods for crop type mapping.

Paris et al. [10] address the challenges in crop type classification, such as the presence of clouds that corrupts the multi-temporal spectral signature of remote sensing images and the necessity of labeled samples, using Sentinel-2 time series and a Long-short term memory (LSTM) model.

In [14] a method to merge spectral, textural and environmental features is proposed, with the purpose of designing a more accurate and efficient crop type classification method.

Swope et al. [32] propose a new self-supervised training, named contrastive sensor fusion, wich is a technique for learning unsupervised representations through a "Siamese network" training scheme. They used shared information from multiple sensors and

spectral bands by training a single model to produce a representation that remains similar when any subset of its input channels is used. Yuan et al. [39] propose a method called SITS-Former, that is pre-trained with unlabeled Sentinel-2 time series data to learn spatio-spectral-temporal features via a missing-data imputation proxy task based on self-supervised learning. Lisaius et al. [40] use representations derived from a spectral-temporal Barlow Twin (STBT) for crop type classification and RL.

The research in [31] is particularly interesting in the use of AEs. They propose an approach that uses the reconstruction losses of joint AEs to detect non-trivial changes (permanent changes and seasonal changes that do not follow common tendency) between two co-registered images in a satellite image time series.

An approach using AEs for unsupervised feature-learning in hyperspectral data was proposed in [41]. The method permits to evaluate the separability of the feature spaces for clustering tasks.

To address the problem of missing data, other techniques in the state, such as the ones based on combination of optical and SAR data [19,42,43], fusion of multiple sensors [7,44,45] or data interpolation [11,46] have been developed. It is worth mentioning that sensor combination/fusion implies the enormous challenge of handling huge amounts of data, which leads to higher computational load and increases in processing time.

**Table 1.** Summary of relevant works related to crop types classification and representation learning.

| **References**(year)[cites] | **Satellite** | **Time range** | **Method** | **Number of classes** | **Feature selection** |
|---|---|---|---|---|---|
| Kalinicheva, E., et al, (2019)[19] [31]. | SPOT-5 | 2002 2008 | AEs | Not specified | N/A |
| Windrim, E., et al, (2019)[42] [41]. | AVIRIS and others | Not specified | AEs | Not specified | N/A |
| Paris, Claudia, et al (2020)[13] [10]. | S2 | 09/2017 08/2018 | LSTM | 12 | N/A |
| Russwurm, Marc, et al. (2020)[65]. [11] | S2 | 01/01/2017 31/12/2017 | RF and others | 9 | N/A |
| Zhiwei Yi, et al. (2020)[56]. [13] | S2 | 23/04/2019 20/09/2019 | RF | 8 | Spectro temporal |
| Shan He, et al. (2022)[3]. [14] | MODIS | 01/01/2009 31/12/2009 | KS | 4 | Spectral textural environmental |
| Leikun Yin, et al. (2020)[35]. [12] | S2 | 01/04/2018 31/10/2018 | RF | 3 | Spectro temporal |
| Lisaius, et al. (2024)[-]. [40] | S2 | 01/01/2017 31/12/2018 | STBT | 8 | Spectro temporal |
| **Proposal in this work** | S2 | 01/01/2017 31/12/2017 | AEs | 9 | N/A |

### *1.2. Contributions*

The main contributions of this work are the following:

1. To tackle cloud-corrupted time series analysis, the proposed framework processes individual time-tagged spectral signatures for feature extraction and thereby completely avoids the use of gap-filling and compositing methods.
2. The proposed methodology uses neural networks with a reduced number of neurons to keep the computational load low, thereby facilitating the processing of large geographic areas.
3. The proposed pixel-wise framework provides a robust solution with respect to the number of available cloud-free observations, while achieving competitive results even under high levels of cloudiness.

The remainder of this work is organized as follows. Section 2 presents the concept of RL and the respective mathematical definitions, as well as a brief description of AEs.

In Section 3 the problem statement of this work and the mathematical formulation of the proposed framework are introduced. Section 4 describes quantitative and qualitative experimental results. Sections 5 and 6 present the discussion and conclusions, respectively, of the results obtained in our experiments.

## 2. Materials and Methods

### 2.1. Representation Learning

Representation learning, also called feature learning, is a subfield of machine learning that aims to automatically learn and identify meaningful features, or representations, from the input data. Representations are expected to be more informative for downstream tasks such as clustering, regression, or classification [47].

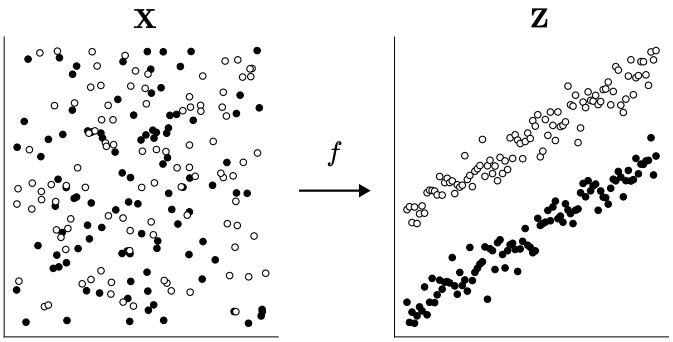



**Figure 1.** Illustration of RL as a function $f$, mapping vectors from a dimensional space to a representation space.

Mathematically, RL is defined as a function $f : \mathbf{X} \to \mathbf{Z}$, that transforms the input data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_S\}$, into features $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_S\}$, where each vector $\mathbf{x}_s \in \mathbb{R}^n$ and its image $\mathbf{z}_s \in \mathbb{R}^p$, and $\mathbb{R}^p$ denotes the representation space. The objective function $f$ leads the model to learn meaningful representations of the input data, preserving information, reducing redundancy and generally reducing dimensionality.

In recent years, many RL methods have been proposed from different perspectives / families [48], e.g., contrastive learning methods (InfoNCE [49–51]), deep metric learning (SimCLR [52,53], NNCLR [54], etc.), non-contrastive methods (VICReg [55], BarlowTwins [40,56], etc), among others. Such approaches are particularly useful in cases where the observed data is generated by a limited set of variables [57]. However, RL is not limited to these families of methods, and conventional neural network models, such as autoenconders can form a representation learning method.

### 2.2. Autoencoders

AEs are a specific type of ANN used for unsupervised learning (Figure 2) [58]. They have applications in various research fields, such as anomaly detection, data compression, and feature learning. Their aim is to encode the input into a compressed representation, and then reconstruct the input from this representation, so that the reconstruction is as similar as possible to the input [47,59]. Both under- and overcomplete versions exist, as well as variational AE [60].

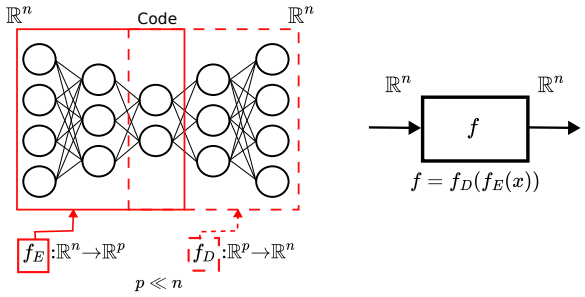



**Figure 2.** Example of an AE architecture with mathematical definition as a function. In the present work, not the codes are used as representations, but the recostruction difference between input and output.

The aim of the AEs, formally defined in [61], is to learn the functions $f_E : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $f_D : \mathbb{R}^p \rightarrow \mathbb{R}^n$, where $f_E$ denotes the encoder function, $f_D$ is the decoder function, $n$ is the input and output dimension, and $p$ denotes the dimension of the code. Generally $p \ll n$, leading to learn compressed features of the data.

## 3. Problem statement and proposed method

Consider a multi-spectral multi-temporal dataset acquired by an optical sensor, where each sample has been acquired at different times. From the entire set of observations, only a subset will usualy be useful as climate conditions such as clouds, cirrus, cloud shadows, snow, among others, occasionally obstruct the land surface. Missing data produced by these conditions commonly leads to poor perfomance on particular tasks, such as land use / land cover classification or change detection. Therefore, it is of utmost importance to extract and use only the land related information, either by filtering the data, or generating new features (often in the form of composites).

### *3.1. Mathematical formulation*

Let $\mathscr{X} \in \mathbb{R}^{P \times B \times T}$ be a multispectral time series dataset represented as a third-order array, where $P$ represents the number of geographic points on the earth surface, $B$ is the number of spectral bands, and $T$ denotes the number of temporal observations, and each geographic point is denoted as a vector $\mathbf{x} \in \mathbb{R}^{B \cdot T}$ and $\mathbf{x} \in \mathscr{X}$. The aim is to transform each vector $\mathbf{x}$ into a representation vector $\mathbf{z} \in \mathbb{R}^R$, where $R$ is the number of new features named representations, to address label scarcity and missing data produced by clouds, and to alleviate tasks such as crop type classification (See Figure 3).

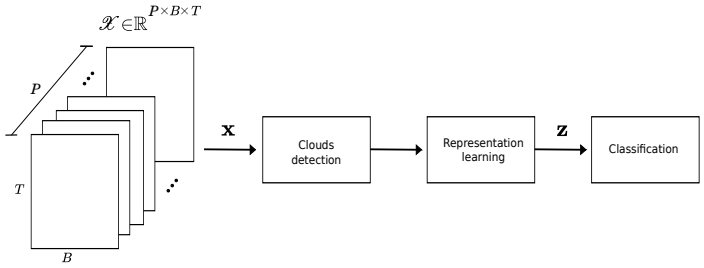



**Figure 3.** First level proposed workflow. Scene classification product provided by the European Space Agency (ESA) is used to mask out cloudy samples from a geographic point (pixel) shaped as a $T \times B$ array. The pixel cloud free observations are the inputs to generate a representations vector, which is a concatenation of the reconstruction differences per AE, and are then the input data to a classification model.

### 3.2. Methodology

The methodolgy of this work consists of four processes: data downloading/prepocessing, model training, inference (representations formation) and, as downstream task, classification. The proposed framework is shown in more detail in Figure 4.

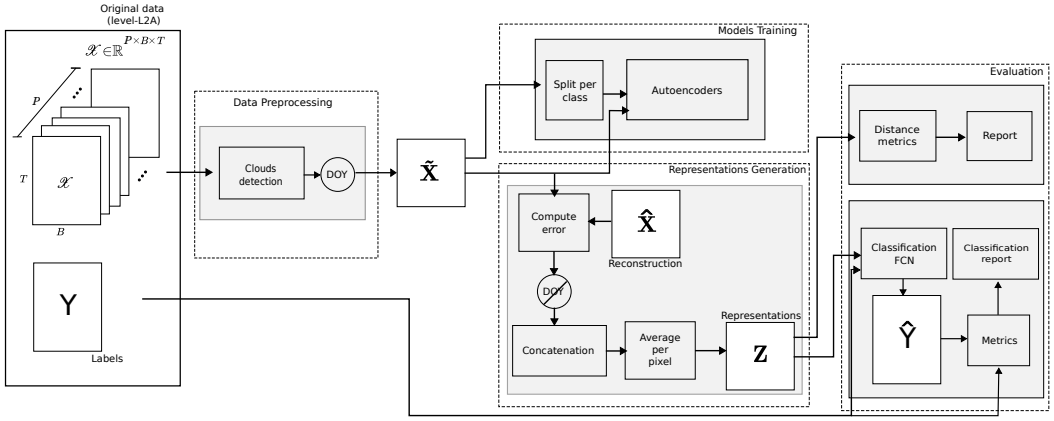



**Figure 4.** Proposed framework block diagram. The full methodology is composed by four main blocks: data preprocessing, model training, representation generation and evaluation.

#### 3.2.1. Data downloading/preprocessing

Reference, crop type labels were extracted from a public benchmark dataset named BreizCrops [11] (field level). Google earth engine (GEE) was used to download full multitemporal multispectral data from a region of interest (ROI) (see Figure 5).

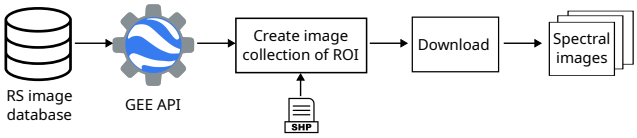



**Figure 5.** Process of downloading datasets using the GEE API.

Given the multispectral multitemporal remote sensing dataset, a scene classification product (e.g., sen2cor) is used to get a cloud/non-cloud mask for each sample $\mathbf{x}$, i.e., for each temporal observation at pixel level. Cloudy samples are excluded from the dataset producing a set of pixels with variable number of cloud free temporal observations.

To leverage the temporal information for the particular task of crop type classification, we add temporal embeddings to each sample, which aim to extend the vector space to one where similar spectral curves of different crop types at different growth stages are separable. Hence, each sample $\tilde{\mathbf{x}} \in \mathbb{R}^F$ has $F$ features, i.e., $B$ spectral bands plus two values denoting the sensing Day Of Year (DOY), and are computed, given the cyclical character of the nature seasons, using sin and cosine functions as follows

$$d_{\sin} = \left(\sin\left(\frac{2\pi d}{365}\right) + 1\right)/2 \tag{1}$$

and

$$d_{\cos} = \left(\cos\left(\frac{2\pi d}{365}\right) + 1\right)/2, \tag{2}$$

where $d$ denotes the DOY as a numeric value from 1 to 365, and $d_{\sin}$ and $d_{\cos}$ are in the range 0 to 1.

### 3.2.2. Model training

The principle of this work is to train a set of $C$ independent autoenconders with vectors $\tilde{\mathbf{x}}_c$ that belong to a particular class/cluster $c$ for $c = 1, \ldots, C$, resulting in $C$ semi-supervised trained models able to reconstruct samples from the same class/cluster, approaching the reconstruction difference to zero, while using the ensemble of reconstruction differences to derive the representations (see Figure 6).

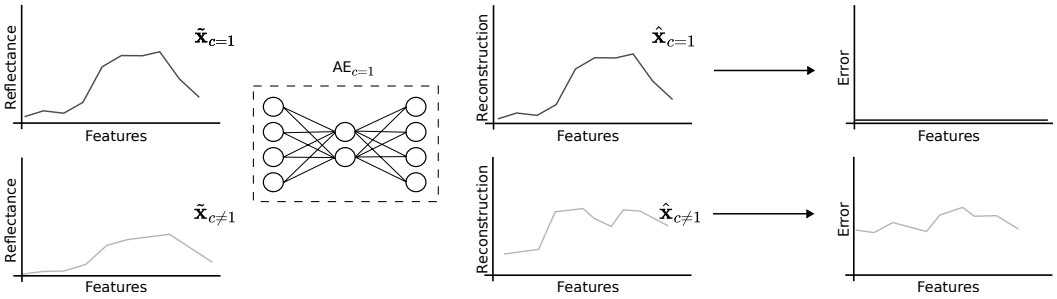



**Figure 6.** Example of the expected output for positive and negative samples. The "errors" from the ensemble of AEs constitute the representations for the downstream task.

The training process can be semi-supervised, given a labeled dataset, as in this work, or unsupervised, with no ground truth data (e.g., by training a set of random AEs not associated to specific crop types or classes). The scope of this work addresses the semi-supervised approach, with a crop type labeled dataset as pairs $(x, y)$, where $y$ is an integer value which indicates the class that $x$ belongs to. Samples are split in as many subdatasets as classes and each subdataset is used to train a different AE. This approach is graphically represented in Figure 7.

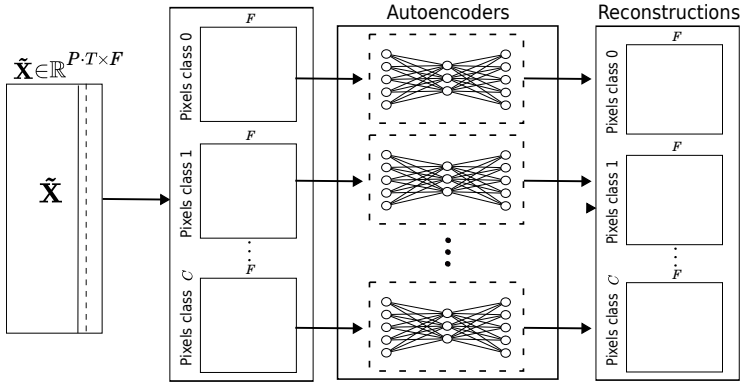



**Figure 7.** AEs training. Each AE is trained with a set of individual spectral curves belonging to one of the crop types. The reconstructions from the $C$ classes are used to calculate the difference vector across the ensemble, that is the final set of representations.

As mentioned in Section 2.2, an AE is formally defined as a composed function $f(x) = f_D(f_E(x))$. For our input vectors $\tilde{\mathbf{x}}_c$, each AE can be seen as a function

$$f(\tilde{\mathbf{x}}_c) = f_D(f_E(\tilde{\mathbf{x}}_c))) \tag{3}$$

where, $f_E(\tilde{\mathbf{x}}_c)$ denotes the encoder function, which maps the input vectors that belong to class $c$, from the input space $\mathbb{R}^F$ to an embedding space $\mathbb{R}^P$, known as code, and, at the same time, it is a composed function

$$f_E(\tilde{\mathbf{x}}_c) = f_{E_I}(f_{E_{I-1}}(\ldots(f_{E_1}(\tilde{\mathbf{x}}_c)))) \tag{4}$$

where $f_{E_i}(\cdot)$, for $i = 1, \ldots, I$ are regression functions

$$f_{E_i}(\mathbf{y}_{i-1}) = \alpha_i(\mathbf{W}_i \mathbf{y}_{i-1} + \mathbf{b}_i) \tag{5}$$

where $i$ denotes the layer number, $\mathbf{y}_{i-1}$ stands for the output of layer $i-1$, $\mathbf{W}_i$ and $\mathbf{b}_i$ denotes the weights and bias at layer $i$ respectively, and $\alpha_i$ the activation function.

The decoder function $f_D(\cdot)$, also a composed function as $f_E(\cdot)$ in eq. 4, maps the code to a estimated reconstruction

$$\hat{\mathbf{x}} = f_D(f_E(\tilde{\mathbf{x}}_c)) \tag{6}$$

which is approximated to the input, updating the weights $\mathbf{W}_i$ and bias $\mathbf{b}_i$ by backpropagating the error $\mathbf{d}$ computed by a loss function $L$ as

$$\mathbf{d} = L(\tilde{\mathbf{x}}_c, f_D(f_E(\tilde{\mathbf{x}}_c))) = L(\tilde{\mathbf{x}}_c, \hat{\mathbf{x}}) \tag{7}$$

where $L$ computes the distance between $\tilde{\mathbf{x}}_c$ and $\hat{\mathbf{x}}$, and $\mathbf{W}_i$ and $\mathbf{b}_i$ are updated approaching $\mathbf{d} \to \mathbf{0}$, by an optimization algorithm such as stochastic gradient descent, Adam, Adagrad.

#### 3.2.3. Inference (representations formation)

Given the $C$ trained AEs, denoted as $f_c$, the set of cloud free observations of individual geographic points (pixels) form an array $\tilde{\mathbf{X}} \in \mathbb{R}^{t \times F}$, where $t$ denotes the number of cloud free samples for a given pixel. $\tilde{\mathbf{X}}$ is the input to all the AEs, and the reconstruction array is obtained from each AE as

$$\hat{\mathbf{X}}_c = f_c(\tilde{\mathbf{X}}) \tag{8}$$

where $\hat{\mathbf{X}}_c$ represents the reconstruction estimated by AE $c$. Then, the difference vector is computed by the same loss function used in training phase (eq. 7) as

$$\mathbf{D}_c = L(\tilde{\mathbf{X}}, \hat{\mathbf{X}}_c) \tag{9}$$

and the pixel mean reconstruction difference $\bar{\mathbf{d}}_c$ is computed by

$$\bar{\mathbf{d}}_c = \frac{1}{t} \sum_{s=0}^{t} \mathbf{d}_{sc} \tag{10}$$

where $\mathbf{d}_{sc}$ denotes the $s$-th row of $\mathbf{D}_c$ and the representations are formed by concatenating the mean pixel reconstruction erros as

$$\mathbf{z} = \bar{\mathbf{d}}_1 \oplus \bar{\mathbf{d}}_2 \oplus \cdots \oplus \bar{\mathbf{d}}_C \tag{11}$$

where $\oplus$ denotes the vector concatenation and $\bar{\mathbf{d}}_\mathbf{c}$ the mean pixel reconstruction difference vector from AE $c$. The inference phase of our proposed framework is presented in Figure 8

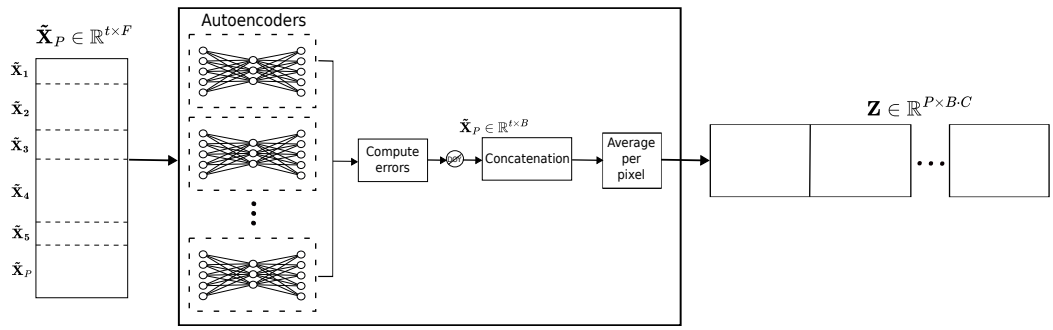



**Figure 8.** Inference workflow of the proposed framework. For each temporal set of cloud-free reflectance spectra, the average reconstruction errors are calculated for each of the $C$ AEs and concatenated to define the representations of this pixel.

## 4. Experiments

### *4.1. Dataset*

For this work the Breizhcrops dataset [11] was used for experiments and evaluation. The provided multi-temporal multi-spectra data is from the Brittany region in the northwest of France and is composed of labeled Sentinel-2 images from January 1st to December 31st, 2017. Labels are assigned to the "average of reflectance values over the bounds of the field geometry retrieved from the dataset" [11].

This dataset is organized in four regions (see Table 2), and each region contains nine crop categories: barley, wheat, rapeseed, corn, sunflower, orchards, nuts, permanent meadows and temporary meadows. To allow for a direct comparison to the work published in [11], we use the regions FRH01 and FRH02 for training, FRH03 for validation, and FRH04 for evaluation. The data split is described in Table 4, which outlines the features employed in this experiment. These include DOY (sin and cosine), 10 spectral bands (10 and 20 meters resampled to 10m) and five well-known spectral indices (NDWI, NDVI, NDTI, NDSVI and EVI).

Table 3 describes the number of samples per class used for training, validation and test respectively. It is worth noting that the dataset is imbalanced, i.e., each class has different number of samples. This makes the classification model more sensitive to overfitting and also makes an accuracy evaluation more difficult [62].

**Table 2.** Regions of Britany (France) with number of field parcels and spectral data for the atmospherically corrected surface reflectances at the bottom-of-atmosphere (L2A) [11]. The regions FRH01 and FRH02 were used for training, FRH03 for validation, and FRH04 for evaluation.

| Regions | NUTS-3 | L2A |
|---|---|---|
| Côtes-d'Armor | FRH01 | 178,632 |
| Finistère | FRH02 | 140,782 |
| Ille-et-Vilaine | FRH03 | 166,367 |
| Morbihan | FRH04 | 122,708 |
| Total | | 608,489 |

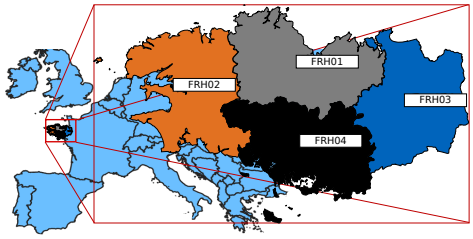



**Table 3.** Number of samples per class used for training, validation and test.

| Class | Training | Validation | Test | Total |
|---|---|---|---|---|
| Barley | 23,787 | 7,154 | 5,981 | 36,922 |
| Wheat | 45,406 | 27,202 | 17,009 | 89,617 |
| Rapeesed | 7,945 | 3,557 | 3,244 | 14,746 |
| Corn | 80,623 | 42,011 | 31,361 | 153,995 |
| Sunflower | 7 | 10 | 2 | 19 |
| Orchards | 1,285 | 1,217 | 552 | 3,054 |
| Nuts | 28 | 10 | 11 | 49 |
| Perm. Meadows | 69,177 | 32,524 | 25,134 | 126,835 |
| Temp. Meadows | 91,156 | 52,682 | 38,414 | 182,252 |

It is worth mentioning that this datasets provides only spectral signatures in tabular format for the center pixel in a field and not Sentinel 2 images. Nevertheless, to enable our qualitative analysis and produce output classification maps, we downloaded Sentinel 2 images using GEE. However, not all Sentinel 2 products in 2017 are available in GEE database and less images than in Breizhcrops dataset were used for this experiment.

### *4.2. AEs training.*

With the aim of developing an algorithm capable of being scalable to relatively large geographic areas, the AE is composed by a single layer FCN as encoder, and its counterpart for the decoder. This keeps computational load and processing times lower than other

models such as convolutional or recurrent networks. Batch size, learning rate, number of units in hidden layer and the loss function were set in accordance with the results acquired through the hyperparameter random search presented in Appendix A. Table 5 presents the AEs configuration.

**Table 4.** Dataset split.

| Parameter | Value |
|---|---|
| Training size | 319,414 |
| Validation size | 166,367 |
| Testing size | 122,708 |
| Features | 10 bands,<br>2 DOY,<br>5 spectral indices |
| Classes | 9 |

**Table 5.** Setting up of the AEs hyperparameters.

| Hyperparameters | |
|---|---|
| Epochs | 1000 |
| Early stop | True |
| Patience | 10 |
| Min. delta | 1e-5 |
| Batch size rate | 0.05* |
| Units in hidden layers | 5 |
| Learning rate | 1e-4 |
| Optimizer | Adam |
| Loss | MAE |

* proportion of samples for each class

Figure 9 shows, for each AE, the training and validation loss function computed as the mean absolute error (MAE)

$$MAE = \frac{1}{M}\sum_{m=0}^{M}|x_m - \hat{x}_m| \tag{12}$$

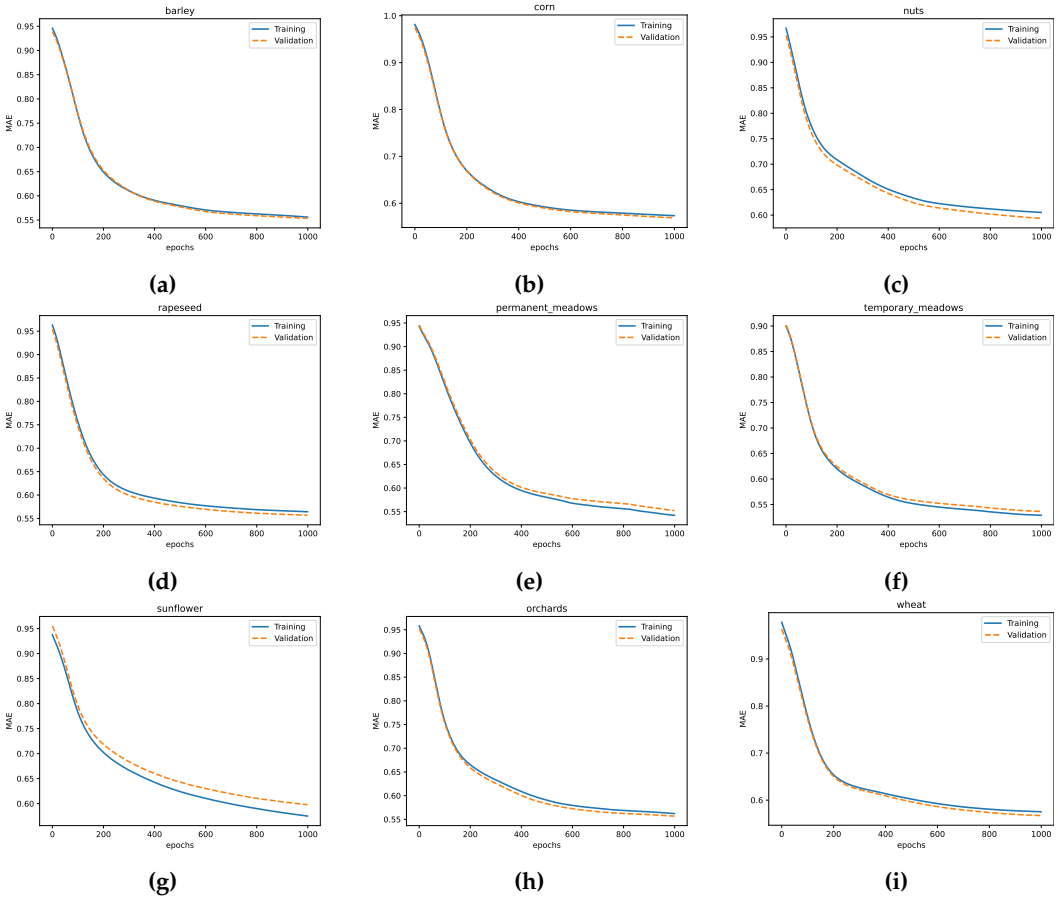
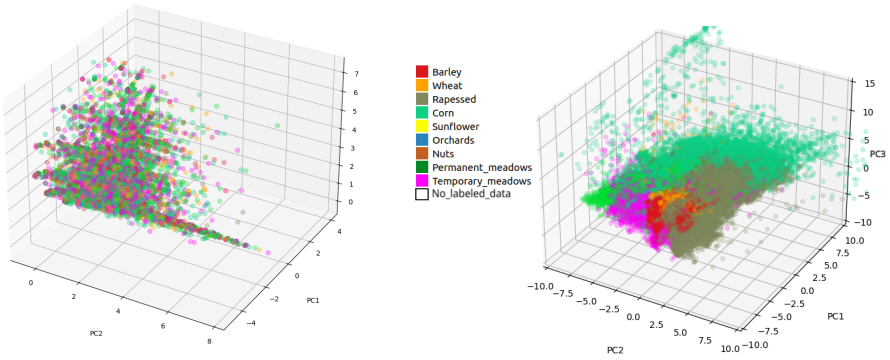



**Figure 9.** Loss functions of AEs trained with (a) barley, (b) corn, (c) nuts, (d) rapeseed, (e) permanent meadows, (f) temporary meadows, (g) sunflower (h) orchards and (i) wheat samples.

### *4.3. Separabilty assessment and distance metrics*

For qualitative assessment of the inter-class separability in the generated representations space, 3D scatterplots of the test spectral-temporal RS data and their corresponding representations produced by our method, reduced to a three-dimensional space by principal component analysis (PCA), are shown in Figures 10a and 10b respectively. The scatter plot of representations, compared with that of the initial data, shows that the density of points belonging to each of the crop types is much better clustered.




**(a)** **(b)**

**Figure 10.** 3D-scatterplot of (a) S2 BOA fixed-length time series (45 observations) and (b) representation, over three principal components obtained by PCA.

Distance metrics were used to assess the distance between classes. Table 6 presents a comparison of the inter-class separability on the input spectral-temporal data and the generated representations, measured by Silhouette score (SS), Calinski-Harabasz index (CH) and Davies-Bouldin Index (DBI) (see Appendix B for a description). In the same way as in the qualitative analysis, distance scores demonstrate much higher separability on the representation space than on the initial data.

**Table 6.** Class distance assessment of the input S2 optical dataset and the representations produced by the proposed method. $SS$ ranges from -1 for incorrect clustering and +1 for highly dense clustering. $CH$ larger scores indicates better separability. $DBI$ ranges from 0 to $\infty$ and the closer to zero the better partition.

| Distance metric | BOA | Our approach |
|---|---|---|
| $SS$ | -0.76 | **0.18** |
| $CH$ | 1.4 | **48678.5** |
| $DBI$ | 72.44 | **9.73** |

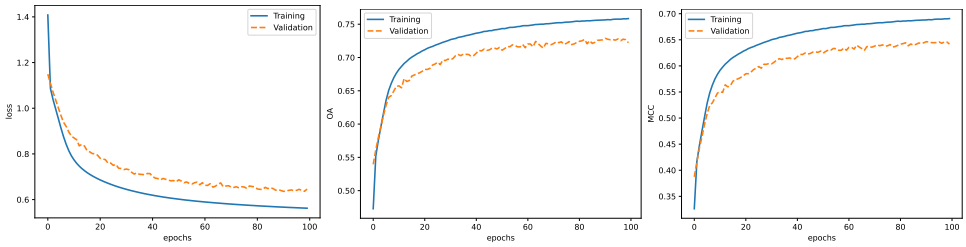
### *4.4. Evaluating representations in the classification of crop types*

After the representations have been learned, a 3-layer FCN was used as classification model, where the inputs are the generated representations and their corresponding labels. The parameters of the classifier are detailed in Table 7. The representation values derived by DOY embeddings are removed from our representation vectors, since averaging these values induces redundancy to the classification model. Hence, the input dimensionality is defined by $B$ spectral bands plus 5 spectral indices.

**Table 7.** Classification model hyperparameters. $B$ denotes the number of bands, and $C$ the number of AEs. Note that only 5 additional features (the spectral vegetation indices) are added to the spectral bands, since DOY features are excluded from the representations in the training of the classification model.

| Hyperparameters | |
|---|---|
| Input size | $(B+5) \times C = 135$ |
| Epochs | 100 |
| Batch size | 100 |
| Units in hidden layers | 128, 64, 32 |
| Learning rate | 1e-5 |
| Optimizer | Adam |
| Loss | Categorical crossentropy |

In addition to monitoring training and validation loss functions, the overall accuracy (OA) is monitored as indicator of classification model performance. Due to the clear imbalance among classes, BreizhCrops is an extremely challenging dataset. Therefore, the Matthews correlation coefficient (MCC), which is robust to class imbalance, is also monitored for a less biased evaluation. Figure 11 presents the progress of these metrics during model fitting. In Figure 11a MAE on validation dataset follows the same descending trend as in the training dataset, indicating that no overfitting occurs during training. From Figures 11b and 11c, it can be seen that OA and MCC are slightly better on the training dataset.




**(a)** **(b)** **(c)**

**Figure 11.** FCN training and validation (a) MAE (b) OA and (c) MCC functions over epochs.

The confusion matrix shown in Figure 12 presents the model performance on the testing dataset. While the model performs relatively accurate for wheat, rapeseed and corn samples, permanent and temporary meadows are not separable. This type of miss-classification is due to the similarity in nature of both crop types. Additionally, the small and challenging classes in this dataset, i.e., sunflower, orchards and nuts, are also totally misclassified mainly due to the very limited number of samples.

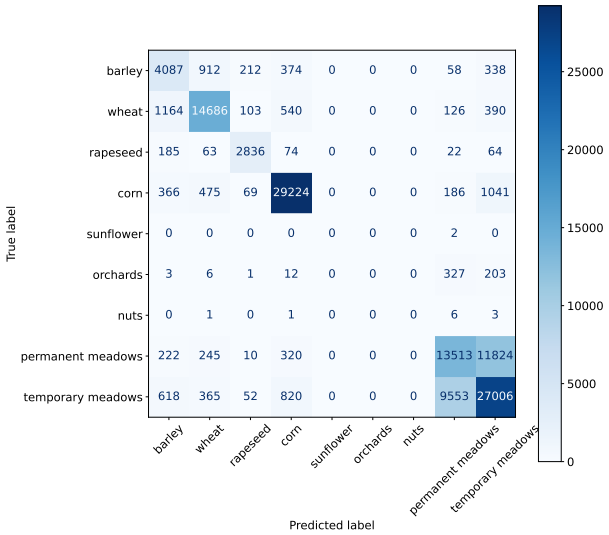



**Figure 12.** Confusion matrix of the FCN prediction for the testing data.

Table 8 presents a performance comparison of our method with convolutional, recurrence and attention methods. TempCNN, OmniscCNN, LSTM, StarRNN, Transformer, and our proposed method with the generated representations as input to a FCN are evaluated by overall accuracy (OA), average accuracy (AA), F1 score, and Cohen's kappa coefficient ($\kappa$). Additionally, details on the processor used, the number of parameters, and runtime are presented. We present the results reported in [11], since the code used for those experiments is not publicly available and it is not straightforwardly reproducible. Notwithstanding, the same test data points were used in our experiments, hence, results are directly comparable.

The OA achieved with our method is 0.74, while a Transformer model reaches 0.80. However, Transfomer is composed by 188,429 trainable parameters, while ours only needs 6,825, i.e., 28 times less trainable parametes, which is directly related to the computational load and processing time. Moreover, other complex convolutional models, such as TempCNN and OmniscCNN, achieve 0.79 and 0.73 OA, i.e., only 0.05 higher and 0.01 lower than our method respectively, but more 400 times more parameters. Likewise, Transformer's AA has the highest score 0.58, while our method reached 0.53, slightly below transformers but still within a competitive range.

The F1 score, which balances precision and recall, and therefore provides a balanced assessment of the model, was also the highest in the LSTM and Transformer models with 0.80, while our representations-based FCN achieved 0.74, reflecting its ability to maintain a reasonable balance despite the imbalanced class distribution. Similarly, LSTM and Transformer models reach the highest Kappa's coefficient (0.75), while our method's is strongly penalized due mostly to the proportion of mismatches on barley, and permanent and temporary meadows. Although our method reaches 0.10 lower score than the competing models, it is still relatively good, especially considering the shallowness of the model.

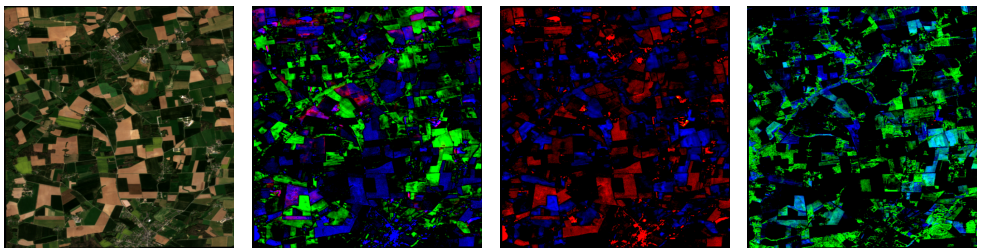
**Table 8.** Classification performance evaluation of benchmarked models. All models were evaluated over the same testing dataset.

| | TempCNN | OmniscCNN | LSTM | StartRNN | Transformer | **Representations-FCN** |
|---|---|---|---|---|---|---|
| OA | 0.79 | 0.73 | **0.80** | 0.79 | **0.80** | **0.74** |
| AA | 0.55 | 0.52 | 0.57 | 0.56 | 0.58 | 0.53 |
| F1 | 0.79 | 0.72 | 0.80 | 0.79 | 0.80 | 0.74 |
| k | 0.73 | 0.65 | 0.74 | 0.73 | 0.75 | 0.66 |
| Processor | 8X NVIDI Tesla P100 16GB/GPU 28,672 Total NVIDIA CUDA Cores | | | | | Intel® Core™ i5-1035G1 CPU @ 1.00GHz ×8 Mesa Intel® UHD Graphics (ICL GT1) |
| N° param | 3,199,501 | 2,739,737 | 1,339,431 | 72,103 | 188,429 | **6,825** |
| Runtime in $[it/s]$ | 1.25 | 1.02 | 1.16 | 1.02 | 1.20 | 0.75 |

### *4.5. Qualitative results*

As BreizhCrops data are sparse geographic points and not images, a qualitative assessment based on classification maps cannot be directly performed on this dataset. Nevertheless, to enable a qualitative analysis, 67 Sentinel 2 multispectral images from 2017 of a subregion in FRH04 (test region) were downloaded and preprocessed. A representative area was defined drawing a polygon where most of the classes (barley, wheat, corn, rapeseed, temporary meadows and permanent meadows) are present (Figure 13a).

Representations for this study area are produced by passing individual pixels from the imagery dataset through the inference workflow outlined in Figure 8. The false color images generated by combining three random representations bands presented in Figures 13b, 13c and 13d, as well as individual representations shown in Figure 14, contrast the crop fields in the new representations space.



**(a)** **(b)** **(c)** **(d)**

**Figure 13.** (a) True color image of the study area in 2017 and false color images combining three random representations per map (b) 59-84-81, (c) 30-11-141, and (d) 45-66-57.

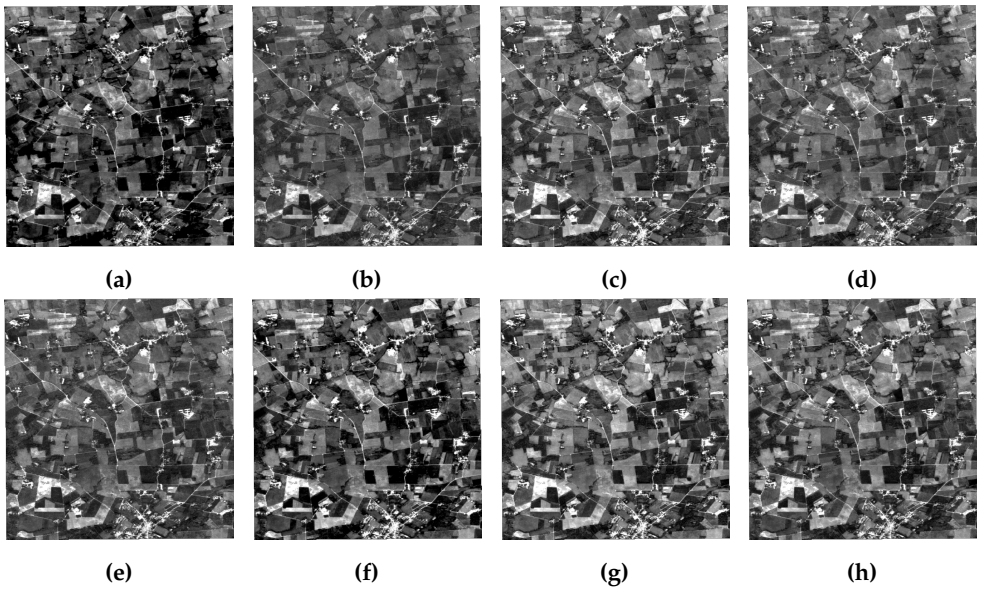



**Figure 14.** Visualization of a subset of the representation bands produced by our RL AE-based framework. We show here only 8 out of $9 \times 15$ available representations.

The classification map produced by the trained classification FCN, with the representations as input data, is presented in Figure 15 together with the ground truth. While the spatial distribution of some classes seems well captured, a number of classes are only poorly mapped.

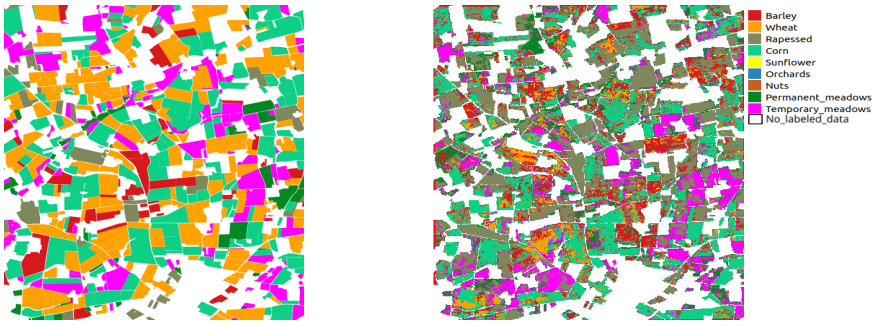



**Figure 15.** (a) Study area ground truth at field level (polygons) and (b) representations-based FCN pixel-wise classification (raster).

## 5. Discussion

Our AEs-based methodology for representation learning (RL) addresses the problem of cloud-corrupted optical data by mapping RS spectral-temporal features into informative, and gap free representations. Our spectral and temporal-based approach produces pixel level comprehensive representations, avoiding the need to employ complex spatial-based classifiers.

The method proposed in this paper has as main advantage its ability to produce pixel-wise representations independently of the number of cloud free samples. Therefore, complex interpolation/gap filling methods, used in other approaches, are not needed. Other solutions, such as obtaining fixed-length time series matching with the input size of a neural network are not needed. However, as other methods, our approach is still negatively affected when limited cloud free temporal observations are available.

Despite the shallowness of the neural networks used for both, representations learning as well as classification, our method performs satisfactory extracting meaningful information for downstream crop classification. While other deep classification networks are

capable of achieving higher OA scores, our light-weight model performs with similar quality.

Models such as TempCNN, OmniscCNN, LSTM, StarRNN, and Transformer, need to be executed on powerful equipment well-suited for handling complex models. In contrast, our full framework was easily launched on a significantly less powerful CPU. This showcases our method's efficiency and adaptability to lower-end hardware and/or scalability to large geographic areas. In terms of number of trainable parameters, convolutional and recurrent models require millions of parameters, which indicates their high computational demands. A shallow three layer FCN, such as the one tested in this paper, is significantly less computationally complex.

The dataset used in the experiments of this work is particularly challenging, as sunflower, orchards, and nuts were not separable, mainly due to the limited number of labeled samples, which restricts our model from learning enough informative and significative representations before classification. However, there are no computational or data limitations to apply our approach on different areas and with datasets from other optical sensors, such as Landsat, and even from radar sensors. Although, this research presents specifically crop type classification task-guided representations, the extrapolation to other classification task is straightforward.

## 6. Conclusions

Based on the results reported in this paper, we derive the following conclusions:

1. Qualitative evaluation based on distance metrics demonstrate that the representations produced by our method accomplished the objective of mapping RS spectral-temporal data to a feature space where inter-class separability is higher than in the initial Sentinel 2 BOA time series.
2. Classification scores obtained by our method, in combination with the comparison of number of trainable parameters and execution time, demonstrate that our method, althought it is not the best classification, reaches competitve accuracy with much less computational load. Therefore, scalability to larger areas is feasible and not excesively time consuming.
3. The confusion matrix, as well as the classification map provide valuable insights that our method in fact performs correctly on the majority of classes evaluated in this work, especially for those with sufficient training samples. However, the challenge of distinguishing between crops with similar spectral characteristics remains. Addressing this issue will be essential for improving the robustness and precision of crop classification models in future studies and will most probably involve the concurrent use of additional sensor modalities.

## 7. Open Issues

Outside the scope of this work, there are still some points to consider in future research:

- Implementation of a fully unsupervised methodology for training autoencoders without relying on labeled data.
- Evaluation of the proposed methodolgy on other optical sensors, radar sensors or combination of both data sources.
- Fine-tuning the classification model to find a better balance between performance metrics and number of trainable parameters.

**Author Contributions:** Conceptualization, A.G. and C.A.; methodology, C.A., A.G. and J.L.; software, A.G. and J.L.; writing—original draft preparation, A.G. and J.L.; writing—review and editing, C.A. and D.T.; visualization, A.G.; supervision, C.A. and D.T.; project administration, A.G.; funding acquisition, D.T.

**Funding:** This research was funded by CONAHCYT grant number 1001207.

**Conflicts of Interest:** The authors declare no conflict of interest

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AA | Average accuracy |
| AEs | Autoencoders |
| ANN | Artificial neural networks |
| BOA | Bottom of atmosphere |
| CH | Calinski Harabasz |
| DBI | Davies Bouldin Index |
| DOY | Day of the year |
| DT | Decission trees |
| FCN | Fully connected network |
| GEE | Google earth engine |
| $\kappa$ | Cohen's kappa |
| LSTM | Long-short term memory |
| MAE | Mean absolut error |
| MCC | Matthews correlation coefficient |
| ML | Machine learning |
| OA | Overall accuracy |
| PCA | Principal component analysis |
| RF | Random forest |
| RL | Representation learning |
| ROI | Region of interest |
| RS | Remote sensing |
| SS | Silhouette score |
| STBT | Spectral-temporal Barlow twins |
| SVM | Support vector machine |
| TOA | Top of atmosphere |
| UA | User's accuracy |

## Appendix A Hyperparameters random search

AEs hyperparameters were defined after an extensive random search. One hundred configurations with four variable hyperparameters were launched and evaluated with three classification and three distance metrics. The search spaces for each hyperparameter are:

- Units: $U\{1, 16\}$
- Batch size rate: $U[0.1, 0.3]$
- Learning rate: $U\left[1 \times 10^{-3}, 9 \times 10^{-6}\right]$
- Loss: $\{0, 1\}$

where $U\{\cdot\}$ and $U[\cdot]$ denote uniform discrete and continuous distribution respectively. Final configuration reported in Table 5 was defined according to the pairwise correlation between hyperparameters and metrics presented in Figure A1.

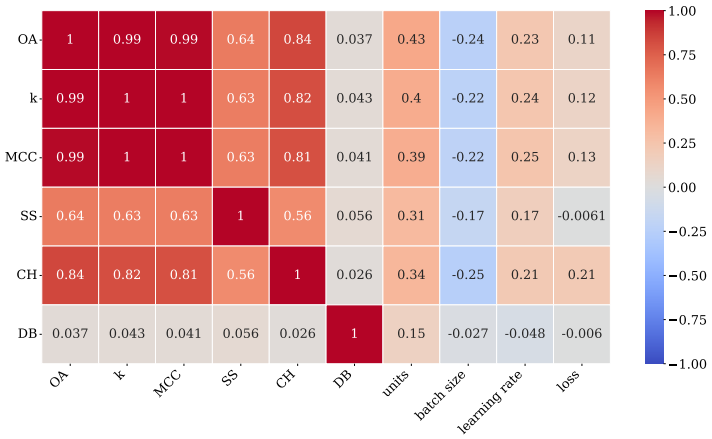



**Figure A1.** Hyperparameters and quality indicators correlation matrix.

## Appendix B Separability metrics

These metrics quantify how separable a set of classes/clusters are from each other.

**Silhouette score**:

$$SS = \frac{b - a}{max(a, b)} \tag{A1}$$

where $a$ is the mean distance between a sample and all other points in the same class, $b$ is the mean distance between a sample and all other points in the next nearest cluster. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering.

**Calinski-Harabasz Index**

$$CH = \frac{\left[\frac{\sum_{k=1}^{K} n_k \|c_k - c\|^2}{K-1}\right]}{\left[\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \|d_i - c_k\|^2}{N-K}\right]} \tag{A2}$$

where $d_i$ is the feature vector of data point $i$, $n_k$ is the size of the $k^{th}$ cluster, $c_k$ is the feature vector of the centroid of the $k^{th}$ cluster, $c$ is the feature vector of the global centroid of the entire dataset, $N$ is the total number of data points. The higher the score is the better separation.

**Davies-Bouldin Index**

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{A3}$$

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} R_{ij} \tag{A4}$$

where $s_i$ is the average distance between each point of cluser $i$ and the centroid of that cluser, $d_{ij}$ is the distance between cluser centroids $i$ and $j$. The score is between 0 and $\infty$, and the values closer to zero indicate a better partition.

## Appendix C Classification metrics

For evaluating the predictions obtained to the FCN, we consider to compute the same metrics that the authors in [11] used for comparative purposes of this work. We compute through of the confussion matrix the equations shown follow:

Given a confusion matrix $\mathbf{M} \in \mathbb{R}^{C \times C}$ where $C$ is the number of classes, the $OA$ is computed with the equation A5.

$$OA = \frac{\sum_{i=1}^{C} \mathbf{M}_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} \mathbf{M}_{ij}} \tag{A5}$$

From $\mathbf{M}$ to a class-wise confusion matrix following the approach one versus all, the producers accuracy also known as precision is computed by

$$PA_c = \frac{TP_c}{TP_c + FP_c} \tag{A6}$$

where $TP_c$ is the true positive and $FP_c$ is the false positive of the class $c$.
Then, the AA is computed as follows

$$AA = \frac{\sum_{c=1}^{C} PA_c}{C} \tag{A7}$$

The user's accuracy ($UA_c$) also known as recall is compute as follows

$$UA_c = \frac{TP_c}{TP_c + FN_c} \tag{A8}$$

where $TP_c$ is the true positive and $FN_c$ is the false negative of the class $c$.
With the equation A6 and A8 we can compute the Weighted F1-score per class ($F1_c$) as follows:

$$F1_c = 2\frac{PA_c \times UA_c}{PA_c + UA_c} \tag{A9}$$

The formula for Cohen's kappa ($k$) is the probability of agreement minus the probability of random agreement, divided by one minus the probability of random agreement.

$$k = \frac{p_o - p_e}{1 - p_e} \tag{A10}$$

where $p_o$ is is the relative observed agreement among raters, and $p_e$ is the hypothetical probability of chance agreement.

## References


1. Vuolo, F.; Neuwirth, M.; Immitzer, M.; Atzberger, C.; Ng, W.T. How much does multi-temporal Sentinel-2 data improve crop type classification? *Int. J. Appl. Earth Obs.* **2018**, *72*, 122–130. https://doi.org/10.1016/j.jag.2018.06.007.
2. Pelletier, C.; Webb, G.; Petitjean, F. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing* **2019**, *11*, 523. https://doi.org/10.3390/rs11050523.
3. Foerster, S.; Kaden, K.; Foerster, M.; Itzerott, S. Crop type mapping using spectral–temporal profiles and phenological information. *Comput. Electron. Agr.* **2012**, *89*, 30–40. https://doi.org/10.1016/j.compag.2012.07.015.
4. Chen, B.; Zheng, H.; Wang, L.; Hellwich, O.; Chen, C.; Yang, L.; Liu, T.; Luo, G.; Bao, A.; Chen, X. A joint learning Im-BiLSTM model for incomplete time-series Sentinel-2A data imputation and crop classification. *Int. J. Appl. Earth Obs.* **2022**, *108*, 102762. https://doi.org/10.1016/j.jag.2022.102762.
5. Hu, Q.; Wu, W.; Song, Q.; Yu, Q.; Lu, M.; Yang, P.; Tang, H.; Long, Y. Extending the Pairwise Separability Index for Multicrop Identification Using Time-Series MODIS Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6349–6361. https://doi.org/10.1109/tgrs.2016.2581210.
6. Palchowdhuri, Y.; Valcarce-Diñeiro, R.; King, P.; Sanabria-Soto, M. Classification of multi-temporal spectral indices for crop type mapping: a case study in Coalville, UK. *The Journal of Agricultural Science* **2018**, *156*, 24–36. https://doi.org/10.1017/s0021859617000879.
7. Heupel, K.; Spengler, D.; Itzerott, S. A Progressive Crop-Type Classification Using Multitemporal Remote Sensing Data and Phenological Information. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science* **2018**, *86*, 53–69. https://doi.org/10.1007/s41064-018-0050-7.
8. Li, Q.; Tian, J.; Tian, Q. Deep Learning Application for Crop Classification via Multi-Temporal Remote Sensing Images. *Agriculture-london.* **2023**, *13*, 906. https://doi.org/10.3390/agriculture13040906.

9. Dong, H.; Ma, W.; Wu, Y.; Zhang, J.; Jiao, L. Self-Supervised Representation Learning for Remote Sensing Image Change Detection Based on Temporal Prediction. *Remote Sensing* **2020**, *12*, 1868. https://doi.org/10.3390/rs12111868.
10. Paris, C.; Weikmann, G.; Bruzzone, L. Monitoring of agricultural areas by using Sentinel 2 image time series and deep learning techniques. In Proceedings of the Image and Signal Processing for Remote Sensing XXVI; Notarnicola, C.; Bovenga, F.; Bruzzone, L.; Bovolo, F.; Benediktsson, J.A.; Santi, E.; Pierdicca, N., Eds. SPIE, 2020. https://doi.org/10.1117/12.2574745.
11. Rußwurm, M.; Pelletier, C.; Zollner, M.; Lefèvre, S.; Körner, M. BREIZHCROPS: A TIME SERIES DATASET FOR CROP TYPE MAPPING. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2020**, *XLIII-B2-2020*, 1545–1551. https://doi.org/10.5194/isprs-archives-xliii-b2-2020-1545-2020.
12. Yin, L.; You, N.; Zhang, G.; Huang, J.; Dong, J. Optimizing Feature Selection of Individual Crop Types for Improved Crop Mapping. *Remote Sensing* **2020**, *12*, 162. https://doi.org/10.3390/rs12010162.
13. Yi, Z.; Jia, L.; Chen, Q. Crop Classification Using Multi-Temporal Sentinel-2 Data in the Shiyang River Basin of China. *Remote Sensing* **2020**, *12*, 4052. https://doi.org/10.3390/rs12244052.
14. He, S.; Peng, P.; Chen, Y.; Wang, X. Multi-Crop Classification Using Feature Selection-Coupled Machine Learning Classifiers Based on Spectral, Textural and Environmental Features. *Remote Sensing* **2022**, *14*, 3153. https://doi.org/10.3390/rs14133153.
15. Dumeur, I.; Valero, S.; Inglada, J. Self-Supervised Spatio-Temporal Representation Learning of Satellite Image Time Series. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2024**, *17*, 4350–4367. https://doi.org/10.1109/jstars.2024.3358066.
16. Wang, S.; Azzari, G.; Lobell, D.B. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sens. Environ.* **2019**, *222*, 303–317. https://doi.org/10.1016/j.rse.2018.12.026.
17. Roy, D.; Yan, L. Robust Landsat-based crop time series modelling. *Remote Sens. Environ.* **2020**, *238*, 110810. https://doi.org/10.1016/j.rse.2018.06.038.
18. Feng, S.; Zhao, J.; Liu, T.; Zhang, H.; Zhang, Z.; Guo, X. Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2019**, *12*, 3295–3306. https://doi.org/10.1109/jstars.2019.2922469.
19. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. https://doi.org/10.1109/lgrs.2017.2681128.
20. Cai, Y.; Guan, K.; Peng, J.; Wang, S.; Seifert, C.; Wardlow, B.; Li, Z. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sens. Environ.* **2018**, *210*, 35–47. https://doi.org/10.1016/j.rse.2018.02.045.
21. Manish Lad, A.; Mani Bharathi, K.; Akash Saravanan, B.; Karthik, R. Factors affecting agriculture and estimation of crop yield using supervised learning algorithms. *Mater. Today:. Proc.* **2022**, *62*, 4629–4634. https://doi.org/10.1016/j.matpr.2022.03.080.
22. Inglada, J.; Vincent, A.; Arias, M.; Tardy, B.; Morin, D.; Rodes, I. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sensing* **2017**, *9*, 95. https://doi.org/10.3390/rs9010095.
23. Rokach, L.; Maimon, O., Decision Trees. In *Data Mining and Knowledge Discovery Handbook*; Springer-Verlag; pp. 165–192. https://doi.org/10.1007/0-387-25465-x_9.
24. Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/a:1010933404324.
25. Cortes, C. Support-Vector Networks. *Mach. Learn.* **1995**.
26. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. https://doi.org/10.1037/h0042519.
27. *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*; Springer International Publishing, 2023. https://doi.org/10.1007/978-3-031-24628-9.
28. Lopez Pinaya, W.H.; Vieira, S.; Garcia-Dias, R.; Mechelli, A., Autoencoders. In *Machine Learning*; Elsevier, 2020; pp. 193–208. https://doi.org/10.1016/b978-0-12-815739-8.00011-0.
29. Lopez-Fandino, J.; Garea, A.S.; Heras, D.B.; Arguello, F. Stacked Autoencoders for Multiclass Change Detection in Hyperspectral Images. In Proceedings of the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2018. https://doi.org/10.1109/igarss.2018.8518338.
30. Luppino, L.T.; Hansen, M.A.; Kampffmeyer, M.; Bianchi, F.M.; Moser, G.; Jenssen, R.; Anfinsen, S.N. Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 60–72. https://doi.org/10.1109/tnnls.2022.3172183.
31. Kalinicheva, E.; Sublime, J.; Trocan, M., Change Detection in Satellite Images Using Reconstruction Errors of Joint Autoencoders. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing*; Springer International Publishing, 2019; pp. 637–648. https://doi.org/10.1007/978-3-030-30508-6_50.
32. Swope, A.M.; Rudelis, X.H.; Story, K.T. Representation Learning for Remote Sensing: An Unsupervised Sensor Fusion Approach, 2021. https://doi.org/10.48550/ARXIV.2108.05094.
33. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. https://doi.org/10.1109/tpami.2013.50.
34. Neumann, M.; Pinto, A.S.; Zhai, X.; Houlsby, N. In-domain representation learning for remote sensing, 2019. https://doi.org/10.48550/ARXIV.1911.06721.
35. Li, W.; Chen, K.; Chen, H.; Shi, Z. Geographical Knowledge-Driven Representation Learning for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. https://doi.org/10.1109/tgrs.2021.3115569.

36. Bengio, Y. Deep Learning of Representations for Unsupervised and Transfer Learning. In Proceedings of the Proceedings of ICML Workshop on Unsupervised and Transfer Learning; Guyon, I.; Dror, G.; Lemaire, V.; Taylor, G.; Silver, D., Eds., Bellevue, Washington, USA, 02 Jul 2012; Vol. 27, *Proceedings of Machine Learning Research*, pp. 17–36.
37. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2019**, *109*, 373–440. https://doi.org/10.1007/s10994-019-05855-6.
38. Ericsson, L.; Gouk, H.; Loy, C.C.; Hospedales, T.M. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Process. Mag.* **2022**, *39*, 42–62. https://doi.org/10.1109/msp.2021.3134634.
39. Yuan, Y.; Lin, L.; Liu, Q.; Hang, R.; Zhou, Z.G. SITS-Former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification. *Int. J. Appl. Earth Obs.* **2022**, *106*, 102651. https://doi.org/10.1016/j.jag.2021.102651.
40. Lisaius, M.C.; Blake, A.; Keshav, S.; Atzberger, C. Using Barlow Twins to Create Representations From Cloud-Corrupted Remote Sensing Time Series. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2024**, *17*, 13162–13168. https://doi.org/10.1109/jstars.2024.3426044.
41. Windrim, L.; Ramakrishnan, R.; Melkumyan, A.; Murphy, R.J.; Chlingaryan, A. Unsupervised Feature-Learning for Hyperspectral Data with Autoencoders. *Remote Sensing* **2019**, *11*, 864. https://doi.org/10.3390/rs11070864.
42. Bégué, A.; Arvor, D.; Bellon, B.; Betbeder, J.; de Abelleyra, D.; P. D. Ferraz, R.; Lebourgeois, V.; Lelong, C.; Simões, M.; R. Verón, S. Remote Sensing and Cropping Practices: A Review. *Remote Sensing* **2018**, *10*, 99. https://doi.org/10.3390/rs10010099.
43. Orynbaikyzy, A.; Gessner, U.; Conrad, C. Crop type classification using a combination of optical and radar remote sensing data: a review. *Int. J. Remote Sens.* **2019**, *40*, 6553–6595. https://doi.org/10.1080/01431161.2019.1569791.
44. Pierre Pott, L.; Jorge Carneiro Amado, T.; Augusto Schwalbert, R.; Mateus Corassa, G.; Antonio Ciampitti, I. Crop type classification in Southern Brazil: Integrating remote sensing, crop modeling and machine learning. *Comput. Electron. Agr.* **2022**, *201*, 107320. https://doi.org/10.1016/j.compag.2022.107320.
45. Moreno-Martínez, A.; Izquierdo-Verdiguier, E.; Maneta, M.P.; Camps-Valls, G.; Robinson, N.; Muñoz-Marí, J.; Sedano, F.; Clinton, N.; Running, S.W. Multispectral high resolution sensor fusion for smoothing and gap-filling in the cloud. *Remote Sens. Environ.* **2020**, *247*, 111901. https://doi.org/10.1016/j.rse.2020.111901.
46. Kandasamy, S.; Baret, F.; Verger, A.; Neveux, P.; Weiss, M. A comparison of methods for smoothing and gap filling time series of remote sensing observations – application to MODIS LAI products. *Biogeosciences* **2013**, *10*, 4055–4071. https://doi.org/10.5194/bg-10-4055-2013.
47. Tzelepi, M.; Nousi, P.; Passalis, N.; Tefas, A., Representation learning and retrieval. In *Deep Learning for Robot Perception and Cognition*; Elsevier, 2022; pp. 221–241. https://doi.org/10.1016/b978-0-32-385787-1.00015-4.
48. Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; et al. A Cookbook of Self-Supervised Learning, 2023. https://doi.org/10.48550/ARXIV.2304.12210.
49. Tschannen, M.; Djolonga, J.; Rubenstein, P.K.; Gelly, S.; Lucic, M. On Mutual Information Maximization for Representation Learning **2019**. [arXiv:cs.LG/1907.13625]. https://doi.org/10.48550/ARXIV.1907.13625.
50. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive Representation Learning: A Framework and Review. *IEEE Access* **2020**, *8*, 193907–193934. https://doi.org/10.1109/access.2020.3031549.
51. Aitchison, L.; Ganev, S. InfoNCE is variational inference in a recognition parameterised model, 2021. https://doi.org/10.48550/ARXIV.2107.02495.
52. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the Proceedings of the 37th International Conference on Machine Learning; III, H.D.; Singh, A., Eds. PMLR, 13–18 Jul 2020, Vol. 119, *Proceedings of Machine Learning Research*, pp. 1597–1607.
53. Bachman, P.; Hjelm, R.D.; Buchwalter, W. Learning Representations by Maximizing Mutual Information Across Views, 2019. https://doi.org/10.48550/ARXIV.1906.00910.
54. Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; Zisserman, A. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations, 2021. https://doi.org/10.48550/ARXIV.2104.14548.
55. Bardes, A.; Ponce, J.; LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, 2021. https://doi.org/10.48550/ARXIV.2105.04906.
56. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction, 2021. https://doi.org/10.48550/ARXIV.2103.03230.
57. Coifman, R.R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. A.* **2006**, *21*, 5–30. https://doi.org/10.1016/j.acha.2006.04.006.
58. Bank, D.; Koenigstein, N.; Giryes, R. Autoencoders, 2020. https://doi.org/10.48550/ARXIV.2003.05991.
59. Zhang, L.; Qi, G.J.; Wang, L.; Luo, J. AET vs. AED: Unsupervised Representation Learning by Auto-Encoding Transformations rather than Data, 2019. https://doi.org/10.48550/ARXIV.1901.04596.
60. Valero, S.; Agullo, F.; Inglada, J. Unsupervised Learning of Low Dimensional Satellite Image Representations via Variational Autoencoders. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, 2021. https://doi.org/10.1109/igarss47720.2021.9554661.
61. Baldi, P. Autoencoders, Unsupervised Learning, and Deep Architectures. In Proceedings of the Proceedings of ICML Workshop on Unsupervised and Transfer Learning; Guyon, I.; Dror, G.; Lemaire, V.; Taylor, G.; Silver, D., Eds., Bellevue, Washington, USA, 02 Jul 2012; Vol. 27, *Proceedings of Machine Learning Research*, pp. 37–49.

62. Foody, G.M. Status of land cover classification accuracy assessment. *Remote Sens. Environ.* **2002**, *80*, 185–201. https://doi.org/10.1016/s0034-4257(01)00295-4.