A Sentinel-2 Multiyear, Multicountry Benchmark Dataset for Crop Classification and Segmentation With Deep Learning

Dimitrios Sykas, Maria Sdraka , Dimitrios Zografakis , and Ioannis Papoutsis , *Member, IEEE*

Abstract—In this work, we introduce Sen4AgriNet, a Sentinel-2-based time series multicountry benchmark dataset, tailored for agricultural monitoring applications with machine and deep learning. Sen4AgriNet dataset is annotated from farmer declarations collected via the land parcel identification system (LPIS) for harmonizing country-wide labels. These declarations have only recently been made available as open data, allowing for the first time the labeling of satellite imagery from ground truth data. We proceed to propose and standardize a new crop type taxonomy across Europe that addresses common agriculture policy (CAP) needs, based on the Food and Agriculture Organization (FAO) indicative crop classification scheme. Sen4AgriNet is the only multicountry, multiyear dataset that includes all spectral information. It is constructed to cover the period 2016–2020 for Catalonia and France, while it can be extended to include additional countries. Currently, it contains 42.5 million parcels, which makes it significantly larger than other available archives. We extract two subdatasets to highlight its value for diverse deep learning applications—the object aggregated dataset (OAD) and the patches assembled dataset (PAD). OAD capitalizes zonal statistics of each parcel, thus creating a powerful label-to-features instance for classification algorithms. On the other hand, PAD structure generalizes the classification problem to parcel extraction and semantic segmentation and labeling. The PAD and OAD are examined under three different scenarios to showcase and model the effects of spatial and temporal variability across different years and different countries.[1]

Index Terms—Benchmark satellite dataset, crop harmonization taxonomy, crop type classification, deep learning.

I. INTRODUCTION

OVER the past years, Copernicus Sentinels (1 & 2) and NASA's Landsat satellites have been consistently collecting images harmonized in the[1] spectral, temporal, and spatial dimensions. The free and open distribution of such an imagery archive has enabled, among others, the consistent and robust monitoring of agricultural activities. Furthermore, the recent developments in Artificial Intelligence (AI) algorithms and models has propelled the adoption and implementation of novel machine learning (ML) techniques, paving the way for a more efficient modeling of the complex agricultural ecosystems and the generation of expertise to support smart farming, common agriculture policy (CAP) implementation, and agricultural insurance.

One of the major problems faced by researchers in this field is the absence of country-wide labeled data that are harmonized along space and time. Specifically in the EU, the CAP has placed a stepping stone to overcome this issue by legally establishing paying agencies in each EU country which are responsible for distributing subsidies to farmers. In order to fulfill their objectives, paying agencies systematically collect the cultivated crop type and parcel geometries for every farmer and record it via the land parcel identification system (LPIS) [1] in a standardized way for each country. Unfortunately, public access to these farmer declarations has been restricted for several years, thus making it almost impossible to get country-wide ground truth data. However, since 2019 and for the first time, these datasets are gradually becoming open (e.g., France, Catalonia, Estonia, Croatia, Slovenia, Slovakia, and Luxemburg). This change offers a significant opportunity for the earth observation (EO) community to explore novel and innovative data-driven agricultural applications, by exploiting this abundance of new LPIS information.

In principle, this fusion of the LPIS data sources has tremendous potential but there are still some barriers to overcome. First of all, the LPIS system of each country is customly configured to utilize the local language of the crop types and the specific taxonomy structure of the crops that matches the local subsidies policy implementation. This nonstandardization of the labels prohibits the spatial generalization of deep learning (DL) models, and thus, needs to be carefully handled to achieve a common representation consistent among countries. On top of these contextual/semantic barriers, parcels are mapped in the corresponding national cartographic projection which in all cases is different from the cartographic projection of the satellite images and pose an additional challenge on the preparation of a consistent, proper and at scale DL-ready dataset.

In this work, we introduce, present, and use Sen4AgriNet, a unique benchmark EO dataset for agricultural monitoring with the following key characteristics. 1) It is pixel based to capture spatial parcel variability. 2) It is multitemporal to capture the crop phenology phases. 3) It is multiannual to model the seasonal variability. 4) It is multicountry to model the geographic spatial variability. 5) It is object-aggregated to

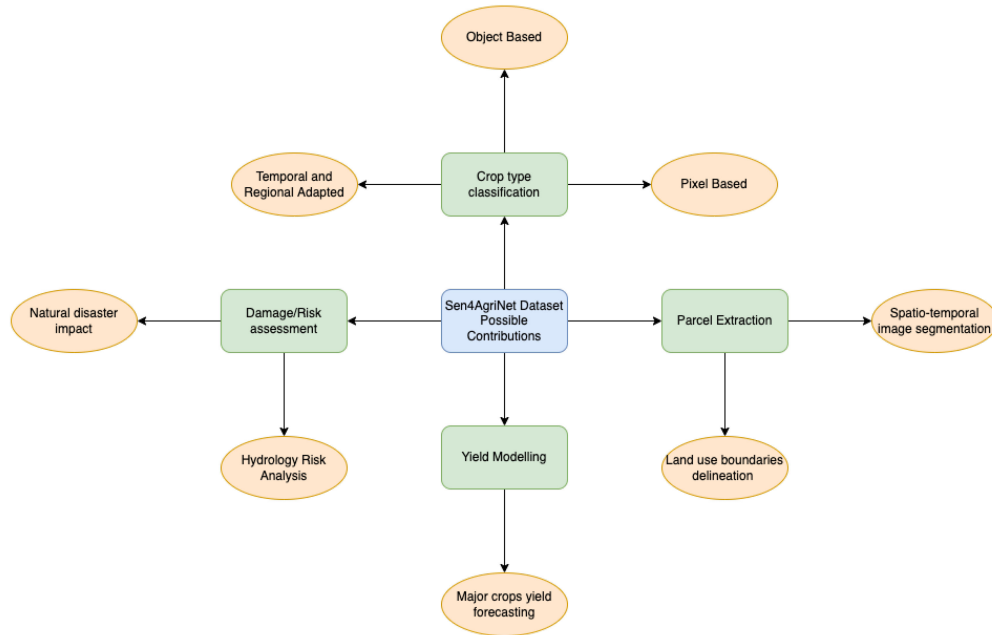[1][Online]. Available: https://github.com/Orion-AI-Lab/S4A

Fig. 1. Taxonomy of Sen4AgriNet potential applications.

further incorporate ground truth data (parcel geometries) in the process. 6) It is modular since it can be enlarged with parcels from more EU countries or expanded in a straightforward way to include additional sensor and non-EO data (e.g., meteorological data). A preliminary version of Sen4AgriNet was introduced and discussed in [2]. This article provides an extended version for Sen4AgriNet, provides additional technical details on its construction, and includes a series of new experiments on top of the dataset. In Fig. 1, we provide a diagram presenting potential applications for Sen4AgriNet, highlighting the major challenges for which the dataset can be used to address. Starting from crop type classification and parcel extraction with their variants, the dataset can also be used for yield modeling applications, where strategic plans for efficiency and sustainability can be investigated, as well as damage/risk assessment tasks for successful planning of relief and regeneration actions.

The Sen4AgriNet contains approximately 225 000 five-year multitemporal Sentinel-2 patches coregistered with open LPIS data for regions in Spain and France with a total size of 10 TB. The dataset is splitted into two distinct subdatasets: the patch aggregated dataset (PAD) and the object aggregated dataset (OAD). PAD contains the original patches of Sentinel-2, i.e., the raster reflectance bands as time series. OAD is built on top of PAD, by aggregating raster values at parcel level, thus producing one sample per parcel which contains all the averaged spectral values with statistics for all bands and available time stamps.

The motivation for creating this dataset is based on the existing issues and challenges that the agriculture remote sensing domain has faced over the last years, despite the access to open satellite data archives. Namely, we aim to provide an outlet for standardized label nomenclature, connection of crop types to operational usage of the results (e.g., EU CAP) and enhancement of the generalization ability of the trained classifiers. Our main contributions can be summarized as follows.

1) We develop and deliver Sen4AgriNet, a benchmark dataset tailored for ML/deep learning (DL) that can be used for a variety of EO-based applications, such as crop type classification, parcel extraction, parcel counting, and semantic segmentation.
2) We introduce a unified crop taxonomy based on the Food and Agriculture Organization (FAO) indicative crop classification (ICC) scheme, to directly address CAP needs.
3) We construct and deliver two reduced versions of Sen4AgriNet, one tailored for pixel-based and one for object-based applications.
4) We conduct a series of baseline experiments to assess for the first time the generalization capabilities of state-of-the-art DL methods across space (different countries) and time (different cultivation years).
5) We open up for reuse Sen4AgriNet, our trained models and our code for generating DL analysis ready datasets.

The rest of this article is organized as follows: initially we review similar existing datasets in EO designed for ML/DL applications. Then the methodology for developing Sen4AgriNet is presented, including the proposed taxonomy and its main classes structure. Third, we present the experimental designs and DL model architectures used to test different generalization scenarios in the context of the pixel and object-based datasets. Finally, the results of the experiments are presented and their implications are briefly discussed.

II. REMOTE SENSING DATASETS FOR ML APPLICATIONS

The concept of creating reference datasets in remote sensing problems targeted at AI algorithms to solve scientific questions has recently gained traction in the community. Most of the existing datasets focus on problems related to land use/land cover classification exploiting preexisting open data sources.

One of the first attempts in this domain is BigEarthNet [3], a large-scale benchmark archive of 125 Sentinel-2 tiles corresponding to acquisitions over ten European countries from June 2017 to May 2018. These tiles were atmospherically corrected with sen2cor [4] and splitted into 590 326 nonoverlapping image patches to better address computer vision problems. For each created patch the multiple corresponding land cover classes were subsequently exported via the CORINE Land Cover (CLC) database [5] and added as label annotations, establishing BigEarthNet a good fit for multiclass, multilabel image classification applications.

A similar approach to BigEarthNet is presented in [6] (Eurosat), where a multiclass annotation dataset with all 13 spectral bands of Sentinel-2 is proposed. Eurosat consists of ten land cover classes with 27 000 labeled and geo-referenced image patches. A more specialized dataset named so2Sat is presented in [7]. It focuses on urban areas and classes across the planet combining both Sentinel-1 and Sentinel-2 image patches. Key asset of the dataset is the manual labeling process which was designed and performed by 15 domain experts.

More relevant from a domain scope to our proposed dataset is the "CV4A Kenya Crop Type Competition" dataset [8]. It combines the temporal aspect of satellite image acquisition with crop types. The dataset uses both the multitemporal coverage of satellite images and the spatial distribution of farm holdings. Key shortcoming of this dataset are the small number of agricultural parcels, the small number of different crop type classes, and the fact that the included satellite and label data cover a single year and no multiannual data are recorded.

Another relevant to the agricultural domain dataset is reported in [9]. The BreizhCrop dataset is used for the supervised classification of field crops from satellite time series and consists of combined labeled data and Sentinel-2 top-of-atmosphere as well as bottom-of-atmosphere time series in the region of Brittany, north-east France, spanning throughout 2017. Key characteristic of the dataset is the object-based approach, i.e., each parcel is represented as one observation (spatial aggregation) with several features (Sentinel-2 reflectances over different timesteps).

In [10], a large-scale land-cover dataset with Gaofen-2 (GF-2) satellite images was created. The Gaofen Image Dataset (GID) focuses on land cover classes, with special remarks on the coverage size and spatial distribution resolution. GID consists of two parts: a large-scale classification set and a fine land-cover classification set. The large-scale classification set contains 150 pixel-level annotated GF-2 images, and the fine classification set is composed of 30 000 multiscale image patches coupled with 10 pixel-level annotated GF-2 images. The training and validation data with 15 categories is collected and relabeled based on the training and validation images with five categories, respectively.

A recently published dataset that resembles the one presented in this study is PASTIS [11], which includes time series of multispectral images obtained from the Sentinel-2 satellite constellation. It contains the ten nonatmospheric bands resampled to the highest spatial resolution of 10 m and the labels expand over 18 crop types. However, contrary to Sen4AgriNet, PASTIS provides single-year observations in a single country, thus limiting the variance and diversity of the data.

TABLE I
LIST OF STATE-OF-THE-ART REMOTE SENSING DATASETS FOR ML APPLICATIONS

Dataset	Type	Area	Span	Use	Labels
BigEarthNet	Pixel	EU	None	LC	590k
Eurosat	Pixel	EU	None	LC	27k
so2Sat	Pixel	Global	None	LC	500k
CV4A	Object	Kenya	1 Year	Urban	4k
BreizhCrop	Object	France	1 Year	Agri.	610k
GID	Pixel	China	N/A	LC	30k
PASTIS	Both	France	1 year	Agri.	124k
Sen4AgriNet	**Both**	**Cat & Fr**	**Annual**	**Agri.**	**42m**

Sen4AgriNet is the only dataset that cumulatively supports both pixel- and object-based aggregations, spans across different countries and multiple years, its input satellite data are time-series, and contains several millions of annotations.
The proposed Sen4AgriNet dataset is marked in bold.

From the reported datasets, most of them do not take into account the time series of satellite image acquisition, while the majority also focuses on annotating tags instead of masks, which constrains the usage of the dataset in simple classification scenarios (see Table I). For example image segmentation, object detection, parcel counting, etc., are applications that cannot be tackled with this kind of aggregation. In addition, removing the temporal aspect of satellite time series restricts Deep Learning models from learning the temporal/seasonal dynamics of the classes. Of course, the temporal dynamics in classes are not always applicable, e.g., in land cover types and urban structures. On the contrary, crop type classes show significant spectral and spatial changes over time, thus it is essential to include the time aspect for such applications.

Adding to the complexity of the problem, crop types have variable spectral and spatial response at different geographic areas and cultivation techniques, which are usually strongly connected to the geographic regions. Phenology stages, different seeding dates and agricultural practices, and varying meteorological conditions are just a few parameters that affect the spectral signatures of the satellite data, even for the same crop type. Therefore, constructing efficient, accurate, and robust DL models requires a domain adaptation (DA) strategy. The Sen4AgriNet dataset was designed in order to foster the development of such ML approaches.

III. CROP TYPE CLASSIFICATION WITH DEEP LEARNING

Several DL approaches have been recently proposed for the classification of crop types on optical satellite imagery, outperforming established methods based on computer vision or traditional ML techniques, e.g., [12]. For example, in [13] an ensemble of 1-D and 2-D convolutional neural networks (CNN) is employed in order to discern 11 types of land cover from Sentinel-2 and Landsat-8 images over Ukraine. The predictions are further refined by fusing auxiliary information such as parcel boundaries, statistical data, vector geospatial data, and more. In [14], a more sophisticated architecture is proposed (FG-Unet) which is based on the popular U-Net model [15], extended with a second branch for independent classification of every image pixel enabling the model to produce coarser polygon boundaries. The input of this method is a window of three Sentinel-2 images over France captured at three different time steps.

A number of publications have also explored 3-D convolutions to simultaneously handle the spatial, spectral, and temporal components of the input images. For example, in [16], the convolutional layers of a VGG model are replaced with their 3-D equivalents in order to classify GaoFen-1/2 image pixels into nine land cover classes. Similarly, in [17], MODIS imagery along with normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) indices are fed to a simple CNN architecture with 3-D convolutions for the detection of winter wheat.

On a different note, several studies have regarded the problem of crop type/land cover classification as a time-series classification problem and have proposed architectures specifically designed to exploit the temporal aspect of the input data. Such models perform either in an N-to-N scheme producing a prediction for every input time step, or in an N-to-1 scheme producing a single prediction after examining the whole input sequence. Here, we will only consider the latter case as it is similar to the approach followed in this work. Previous studies have incorporated recurrent neural networks (RNNs) in their pipeline, such as the long short-term memory (LSTM) model [18], which are fed pixel vectors of multiple time steps [19]–[22]. In [23], a 2-branch model is proposed (DuPLO) whose first branch is a CNN spatial feature extractor, whereas the second branch is a gated recurrent unit (GRU) [24] temporal feature extractor. The outputs of both branches are then fused and fed to a final classification layer which makes a prediction for the central pixel of the input window.

Another popular technique for handling spatiotemporal data is the TempCNN model, a custom deep neural architecture which performs convolution simultaneously on the spectral and temporal dimensions, thus managing to extract useful information for the phenological stages and the spectral signature of the different land cover types. This model was first introduced in [25] for Formosat imagery classification and then subsequently used in [26] in comparison with LSTM, DuPLO, multiscale ResNet [27], and transformer [28] for Sentinel-2 input data. The transformer was also explored by [29] in a pipeline which includes three multilayer perceptrons (MLPs) for feature extraction and final classification.

Lastly, a number of studies have also employed convolutional recurrent layers which are essentially RNNs performing 2-D convolutional operations internally. An example of such method is shown in [30] where a convolutional gated recurrent unit (ConvGRU) takes an input sequence of Sentinel-2 images both in correct and reverse order and produces a classification map for 17 crop classes. A more recent module was proposed in [31], named Stackable Recurrent Cell (STAR), and then extended with convolutional layers in [32] (MS-ConvSTAR) for hierarchical crop type classification on Sentinel-2 imagery. This model achieves faster and more stable convergence than the LSTM/GRU equivalents.

IV. DOMAIN ADAPTATION

When there is a significant divergence between the train and test data distributions, a model trained on the former will likely fail to generalize on the latter, and thus, performance will be greatly degraded. A popular approach which attempts to tackle this problem is domain adaptation (DA), a special branch of transfer learning which aims to alleviate the variation between the source (train set) and target (test set) data distributions caused by *data shift*, *concept drift*, and *multimodal domain shift* often observed in remote sensing data [33]. Data shift describes the spectral differences between images captured under different conditions, i.e., different atmospheric effects, sun positions, sensor angles, etc. Concept drift refers to the variation of the intrinsic class characteristics over time and/or space. For example, the same crop displays different spectral signatures as the plant develops, since leaf characteristics, and biomass-to-soil ratios can vary both over time (seasonal, phenotypical changes) and space (regional agricultural policies). Finally, multimodal domain shift arises when multiple sensor types are utilized and differences in bands, resolution, etc., are observed. In the present study, multimodal domain shift is not an issue to be handled since all images were acquired by the Sentinel-2 satellite, but the other two challenges of data drift and concept drift are relevant. In particular, data drift can possibly be observed throughout the whole dataset since atmospheric and illumination conditions are unstable and may vary across time, whereas concept shift is prevalent both in a single examined time series and across regions.

Through DA techniques a model trained on the source domain can be carefully transferred to the target domain without suffering from the effects of the aforementioned distribution variations. There are three major types of DA depending on the availability of ground truth labels in the target domain: 1) supervised DA; 2) Unsupervised DA; 3) semisupervised DA [34]. In supervised DA, target data are fully labeled, but labels may be fewer and/or different than those of the source data. In unsupervised DA, no ground truth labels are available for the target data, whereas in semisupervised DA few labeled instances may be present in the otherwise unlabeled target data set. Several DA approaches have been proposed over the years and can be categorized into three families. Discrepancy-based DA methods attempt to fine-tune the model by minimizing some criterion between the source and target distribution. This criterion can be based on the target labels (e.g., [35]–[37]), the statistical shift between the distributions (e.g., [38]–[40]), the architecture of the model (e.g., [41]–[43]) or the geometrical properties of the distributions (e.g., [44]). Adversarial-based DA methods employ a domain discriminator to assess whether the generated features correspond to the source or the target domain, thus encouraging domain confusion and the production of more robust features (e.g., [45]–[47]). Finally, reconstruction-based DA aims to achieve feature invariance either by translating the target to the source domain (e.g., [48], [49]) or by mapping both source and target domains to a common latent space (e.g., [50], [51]).

Specifically for the task of crop type classification, a number of methods have been proposed for cross-region adaptation, such as [52] and [53] which take as input a time series of images and output a single segmentation map. A method published recently in [54] additionally accounts for the phenology shift observed for a single crop between different geographical regions and

with a time shift estimation procedure and a semisupervised learning scheme it manages to boost the performance of the model proposed in [29].

V. METHODOLOGY

A. Harmonized Crop Type Taxonomy

The ICC [55] scheme was developed by the United Nations FAO organization. It is an approach to produce a harmonized vocabulary and taxonomy for crops and plants that are used in food production. The CAP in Europe is an example that requires such normalization among the crop labels, since each country member uses different naming systems and languages. Therefore, Sen4AgriNet adopts and customizes an extended version of FAO ICC in order to create a universally applicable crop label nomenclature.

A harmonized crop type taxonomy which is designed to be used for satellite crop type classification should not contain solely crop classes, since agricultural parcels coexist with other unrelated classes in satellite images. Therefore, additional classes are needed to annotate pixels (or objects) that are present in the satellite imagery, but are out of scope from the agricultural domain context. In addition, in order to promote the adoption of the dataset under the CAP regulatory framework, relevant land use classes need to be added. To this end, we created this version of Sen4AgriNet by including some major land use classes from CLC [5], and a few additional classes to complete the semantic concepts ("fallow land" subclass, "barren land" subclass and "no data available" subclass). CLC classes related to agriculture and forestry were not included, since they are properly and in detail covered by FAO ICC.

FAO ICC [55] has 4 hierarchies ("Group," "Class," "Subclass," "Order") with corresponding numeric ordering. From the 160+ entities in the hierarchy, the fourth one ("Order") contained only 15 entities. In order to reduce the complexity of any future classification attempts and different label groupings, the "Order" level was removed by upgrading the corresponding entities to "Subclass" level and removing their parent "Subclass." Adopting and customising this FAO/CLC crop classification scheme to create Sen4AgriNet provides the following benefits.

1) Single language (English) is used and naming for all classes across all participating countries.
2) Classes are normalized among different datasets.
3) Hierarchical class structure is adopted. Depending on the application different levels of classes can be used.
4) Additional nonagricultural classes are used to model RS spectral signatures to support agricultural applications.

The presented custom FAO/CLC classification scheme has a total of 9 groups, 168 classes, and subclasses. The 161 classes/subclasses are crop related, 4 are some major CLC classes (as subclasses in this hierarchy), 2 are the fallow and barren lands, and 1 is the no data subclass. Please refer to Fig. 2 for a visualization of the proposed taxonomy.

B. Crop Type Labels—LPIS

The LPIS datasets that are used to create the Sen4AgriNet dataset undergo a tailored process before using them to annotate Sentinel-2 imagery. Initially, the crop type classes are translated from the local language (Spanish/Catalan and French) into English and mapped to the semantically normalized FAO ICC crop classification scheme. This is a laborious and manual procedure, but essential in order to harmonize the different labels among different LPIS systems. The current version of Sen4AgriNet contains the following.

1) The LPIS data for the region of Catalonia provided by the "Agricultura, Ramaderia, Pesca i Alimentació" with an Open Data Commons Attribution License from 2016 to 2020 [56], with a total of 2.5 M parcels.
2) France LPIS data provided by the French Paying Agency with an Open Data Commons Attribution License from 2016 to 2019 [57], with a total of 40 M parcels.

Fig. 3 presents an example of the annotated Sen4AgriNet dataset, highlighting also the spatial and temporal variability of the crops and background classes. The first two patches in Fig. 3 depict the same patch at different years (2019 and 2020), while the next two patches are different patches for the same year and region. The *background* and *other crops* classes in the taxonomy are distinguished on purpose. *Background* refers to any noncrop-related label, while *other crops* is a label representing crops that either can not be matched to the existing taxonomy or are unknown. A more detailed view with visual examples about how data loading works can be found at [58]. Please note that the aforementioned code repository explains both how instances of the same patch transform through time and how empty-data months are treated. Additionally, code examples are provided to assist individuals with custom logic writing.

C. Dataset Structure

The core Sen4AgriNet dataset is used to generate two subdatasets, the object aggregated dataset (OAD), and the patches assembled dataset (PAD). These subdatasets are reduced in size to allow the training of DL models without worrying about the availability of computing resources becoming a bottleneck.

We first construct the PAD subdataset, and then, based on this, we subsequently build the OAD subdataset. PAD was built using an automated procedure that downloads and processes Sentinel-2 images. This process is tile oriented per available year of LPIS data. Overall, the available LPIS data from France and Catalonia extend 55 Sentinel-2 tiles. All Sentinel-2 L1C images with less than 10% cloud coverage are selected for download and each image is split into 900 nonoverlapping patches. A single patch contains 366×366 images for the 10-m bands, 183×183 for the 20-m bands and 61×61 for the 60-m bands. The size of the patches was chosen in order to have integer division of the size of the tile with all 3 different spatial resolutions of Sentinel-2.

As a next step, the patches that correspond to the same location but have different dates are stacked into a single netCDF file. The netCDF format was chosen because it is self-describing (compared to other formats like geotiff, jpeg2000, etc.), portable, flexible, and is a popular standard for storing geospatial data. Sen4AgriNet consists of thousands of patches/files, thus managing this huge volume of data is a major concern. Furthermore, since we run many different experiments on Sen4AgriNet
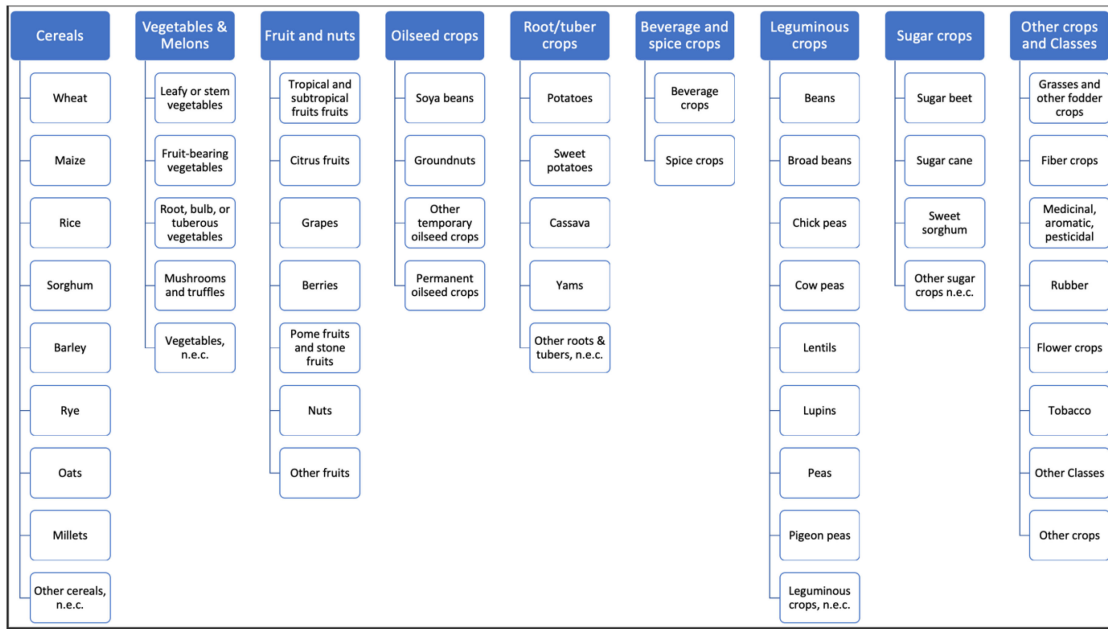
Fig. 2. Proposed Sen4AgriNet crop categorization taxonomy structure, inspired by FAO and based on the customization of the ICC [55] model.
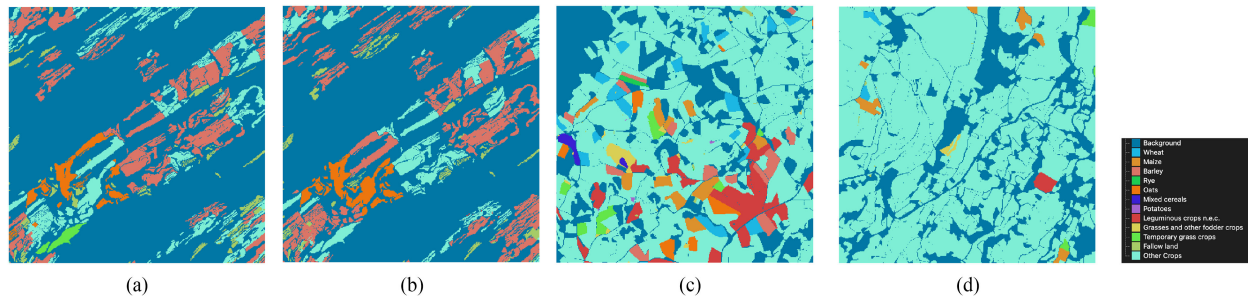

Fig. 3. Representative patches from the proposed dataset. (a) Catalonia 2019 (31TCG), (b) Catalonia 2020 (31TCG), (c) France 2019 (31TDK), (d) France 2019 (31TDK). Note that (a) and (b) refer to the same patch on different years, whereas (c) and (d) refer to different patches from the same year and region.

dataset, the self-descriptiveness flexibility of netCDF allows each experiment to be documented within the corresponding experimental dataset, without the need for auxiliary or explicit documentation tools.

After creating the patches, the labels need to be imported. For each patch the corresponding 366×366 LPIS with the Sen4AgriNet taxonomy applied was rasterized to match the specific pixels in that patch. Using this approach each patch includes a total of $13 * n_{\text{time}} + 2$ variables in the netCDF file. Each "spectral" variable contains all the n_{time} data acquisitions (depending on the region n_{time} varies from 30 to 50 timestamps for one year). The remaining two variables contain the rasterized LPIS masks. The first mask is a mask of integers corresponding to the crop codes of the taxonomy. The second mask is a 64-b integer that contains the parcels code number and is mainly useful for individual parcel extraction problems.

We must note here that the size of the resulting dataset is huge. A single Sentinel-2 tile stack that is transformed into patches for a one year span, results in ~ 900 patches in total and has volume ranging between 6 and 44 GBs.

The OAD part is built on top of PAD by aggregating to the parcel level. The pixels contained within each individual parcel are aggregated via the mean and standard deviation functions to two single numbers. This is repeated for each available timestamp within the netCDF variable. The final result is a CSV file containing the individual parcels as rows and the aggregated timestamps for spectral bands as columns.

France's and Catalonia's parcels are mainly present in 55 Sentinel-2 tiles (see Fig. 4). The total size of the Sen4AgriNet dataset is estimated to be 10 TB for the 2016–2020 time span, which corresponds to a total of 225 000 patches.

VI. EXPERIMENTS

In this work, we extend the experiments presented in [2] by using new DL architectures and new subdatasets. We expand the geographic and temporal coverage beyond using Catalonia's part of the Sen4AgriNet dataset for the year 2020 [2], by also including years 2019 and 2020 for the entire region of Catalonia and part of France. We sample 5000 patches from the entire dataset with label stratification, resulting in a reduced dataset of
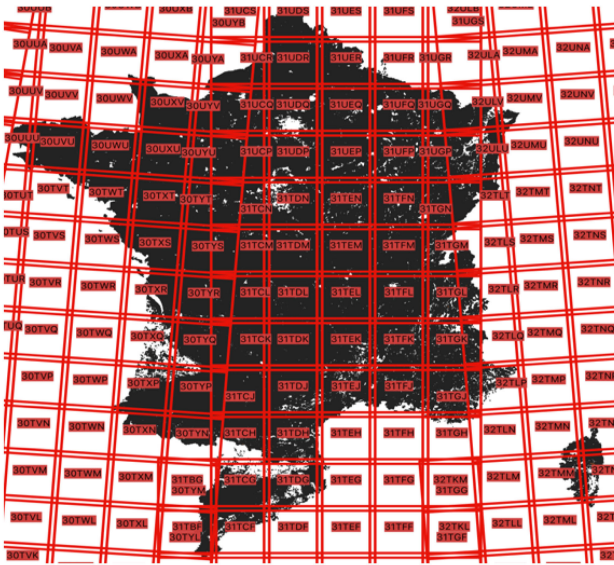
Fig. 4. France and Catalonia (Spain) LPIS dataset overlaid with Sentinel-2 tiles.
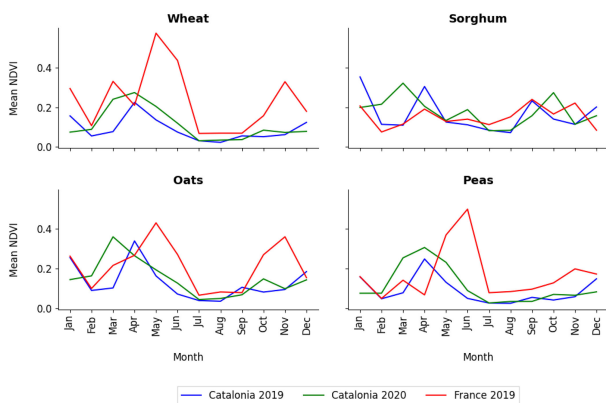

Fig. 5. Mean NDVI of different crops across space and time.

TABLE II
LIST OF SCENARIOS APPLIED IN THE EXPERIMENTS FOR PAD AND OAD

Scenario	Train	Test
1	Catalonia (2019, 2020) France (2019)	Catalonia (2019, 2020) France (2019)
2	Catalonia (2019, 2020)	France (2019)
3	France (2019)	Catalonia (2020)

wheat, sorghum, oats, and peas across the different years and regions. We can see that in Catalonia a similar trend emerges for both years (albeit somehow shifted in time), whereas in France a significantly different crop behavior is observed. This deviation is justified by both the different climatic conditions and agricultural practices between the different regions. However, crop classification is primarily based on the recognition of the unique growth patterns of the crops, therefore, such divergences pose a serious challenge on the design of a robust classifier.

The goal of these experiments is to train DL architectures for crop type classification (OAD) and cropland mapping (PAD) to shed some light into the effects of temporal and spatial crop variability on our models. Our approach is to create challenging scenarios and test the spatio-temporal generalization performance of state-of-the-art DL models. In Table II are the three basic scenarios we have identified.

1) Scenario 1: We use all 5000 patches together, and create a random split. Data from Catalonia 2019 and 2020, and from France in 2019 are used.
2) Scenario 2 (spatial generalization): We use the patches from Catalonia 2019 and 2020 for training, and test on the patches from France 2019.
3) Scenario 3 (spatio-temporal generalization): We use the patches from France 2019 for training and test on the patches from Catalonia 2020.

Due to the different representation of the parcels in the two subdatasets, different approaches need to be followed in each case. The first approach is based on PAD and considers the labels as rasterized pixel-based maps rather than polygons. Therefore, a semantic segmentation pipeline is employed and detailed pixel-level crop type maps are extracted. By further including the geometry of the parcels additional problems can be defined, such as parcel extraction, parcel counting, etc.

The second approach, based on OAD, considers the parcels as objects and the raster time series are aggregated per object via zonal statistics (mean, standard deviation, skewness, etc.), thus providing parcel-based insights to the model. This proves rather useful when the parcel geometries are already known (e.g., farmers declarations from EU paying agencies) and the need to simultaneously identify parcels and crop types becomes obsolete. Of course, when the initial geometries are not available then PAD trained models are the solution. Finally, OAD trained models can be used as input to the PAD dataset if the specific DL architecture utilizes pixel-based input (e.g., LSTM).

For the evaluation of our experiments, we use the accuracy, precision, and F1 scores. Due to high class imbalance, we weigh all metrics with the samples of each class. Also, the corresponding confusion matrices of each experiment and scenario are plotted, normalized across predictions.

size ~ 140 GB. In order to limit the timesteps for the input time series, we aggregated the data into 12 time-bins by calculating the median of each month, and used a fixed window including the medians of months 4 (April) through 9 (September) for model training. Only bands red, green, blue, and near-infrared were selected due to their higher spatial resolution.

As a second step, the number of labels was constrained in order to ensure proper number of available observations for training/validating/testing. The label distributions from each region (pixel-wise) were extracted. The common labels between regions were isolated, the cumulative sum of the labels was computed and then all common labels forming the 99.9% of the data were selected. This resulted in 11 classes (from a 168 total): wheat, maize, sorghum, barley, rye, oats, grapes, rapeseed, sunflower, potatoes, and peas. Based on this, the train-validation-test ratio is fixed to 60%–20%–20% for all experiments.

The spatiotemporal variability of the dataset is evident if we examine the observed mean NDVI of the selected crops. Fig. 5 showcases how the mean NDVI changes over time for
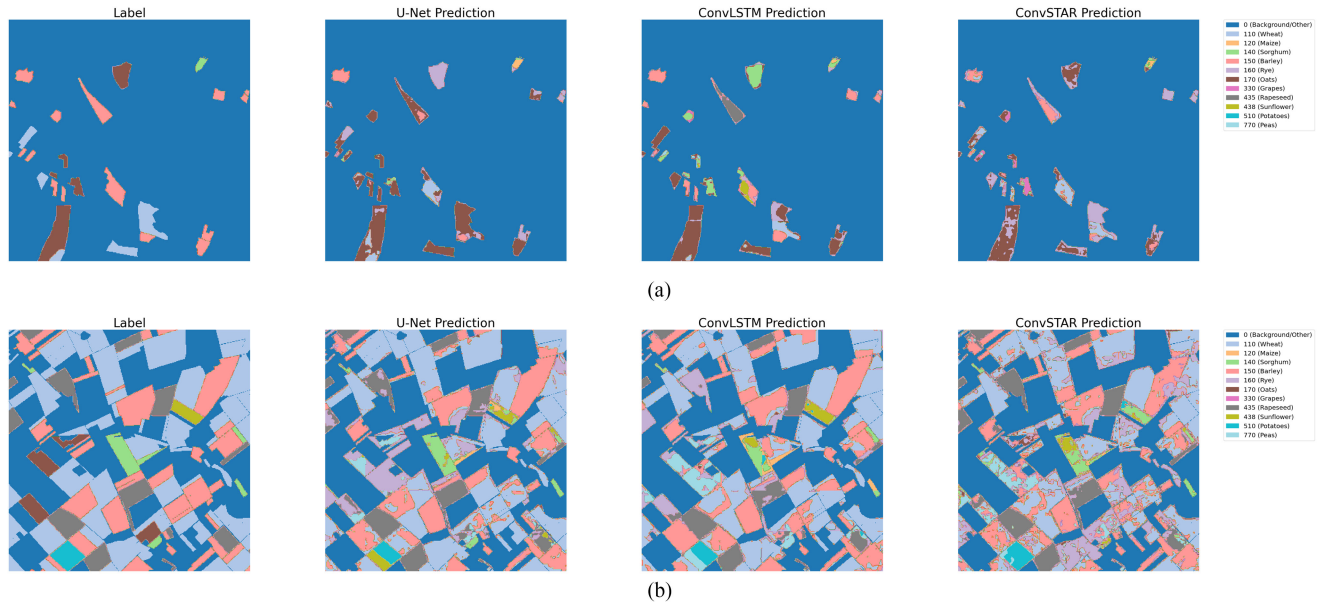
Fig. 6.   Scenario 1. Visual evaluation of the U-Net, ConvLSTM, and ConvSTAR PAD predictions for two different image patches.

A. Sen4AgriNet—PAD Experiment

In the first set of experiments, we employed three popular models for image segmentation: U-Net [15], ConvLSTM [59], and ConvSTAR [32]. First, a 3-layer U-Net (~ 1.9 m trainable parameters) with LogSoftmax activation and a weighted negative log-likelihood loss was employed as a simple and robust baseline for semantic segmentation. Then, a 3-layer ConvLSTM with an encoder–decoder structure (~ 9 m trainable parameters) and LogSoftmax activation at the last layer was trained. Intermediate layers utilize LeakyReLU activations, while the loss function is a weighted negative log-likelihood loss. Second, a 3-layer ConvSTAR model (~ 260 k trainable parameters) was also used in comparison, since it is considered more robust and shows more stable convergence. Similar to ConvLSTM, a final LogSoftmax activation was added for the final pixel classification and the loss function of choice was the weighted negative log-likelihood. Due to time constraints, both models adopt the architecture proposed in their corresponding publications.

Input patches were divided into nonoverlapping subpatches of size 61×61 for faster computation, so the input is a $6 \times 4 \times 61 \times 61$ time series (TxCxHxW, T: timesteps, C: channels, H: height, W: width). Specifically for the U-Net model, images from different timesteps were concatenated along the channels dimension resulting in an input of shape $24 \times 61 \times 61$ (T*CxHxW). During training, we masked all nonparcel or unknown pixels and let the models learn only from the known 11 labels. Similarly for the inference stage, all evaluation metrics were calculated on the known parcel labels only. Finally, the Adam optimizer with an initial learning rate of 0.001 was used and a learning rate reduction scheme was employed on validation loss plateaus.

1) Scenario 1: In the first scenario, where all regions and years are randomly sampled both in the training and test datasets, ConvLSTM performed significantly better than the other two

TABLE III
RESULTS ON ALL SCENARIOS FOR PAD

Scenario	Model	Acc. W (%)	F1 W. (%)	Precision W. (%)
1	U-Net	93.70	82.61	86.64
	ConvLSTM	**94.72**	**85.18**	**86.86**
	ConvSTAR	92.78	80.38	83.33
2	U-Net	**83.12**	**57.85**	**61.57**
	ConvLSTM	82.53	56.56	60.57
	ConvSTAR	79.52	52.15	58.98
3	U-Net	**72.11**	**43.54**	**68.42**
	ConvLSTM	69.86	40.47	66.17
	ConvSTAR	69.07	34.45	67.43

Best results are marked in bold.

models in all three summary metrics (see Table III). The confusion matrices (see Fig. 11) and the visual inspection of the predictions (see Fig. 6) support the high valued summary metrics. An interesting observation from the confusion matrices (see Fig. 11) is that all models seem to confuse the same crop types, with barley, rye, oats, and wheat being the most characteristic case. This is somewhat expected since they are all cereal crops with similar phenology.

2) Scenario 2: In the second scenario, all models were evaluated for their adaptation from one region to another. As shown in the results, all architectures achieve lower accuracy metrics with significantly reduced precision and F1 scores. Again, the corresponding confusion matrices (see Fig. 11) show similar misclassification patterns between the three models, with wheat, rapeseed, and sunflower being the most correctly classified crops. Fig. 7 shows a selection of patches to visually appreciate correct and wrong classifications of the models, with respect to the ground truth.

3) Scenario 3: In the third scenario, all metrics were considerably degraded and all architectures showcased similar performance. From the confusion matrices (see Fig. 11), only maize and rapeseed display the highest accuracy. Similarly, in Fig. 8
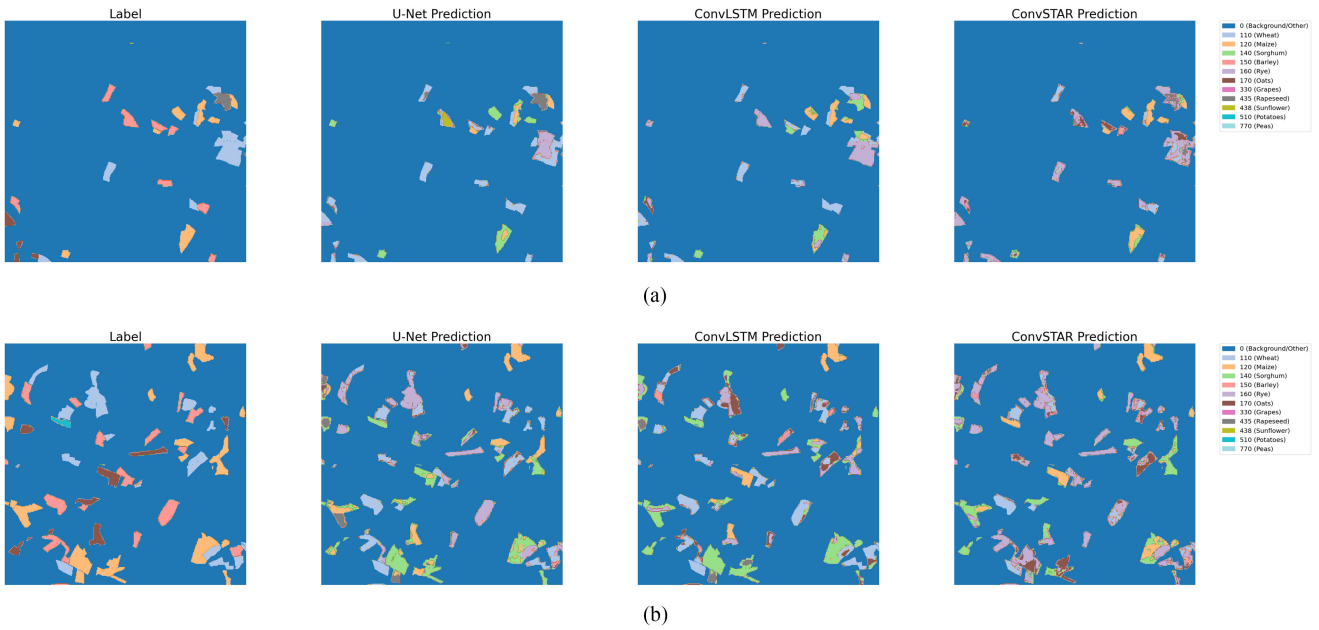
Fig. 7. Scenario 2. Visual evaluation of the U-Net, ConvLSTM, and ConvSTAR predictions for two different image patches.
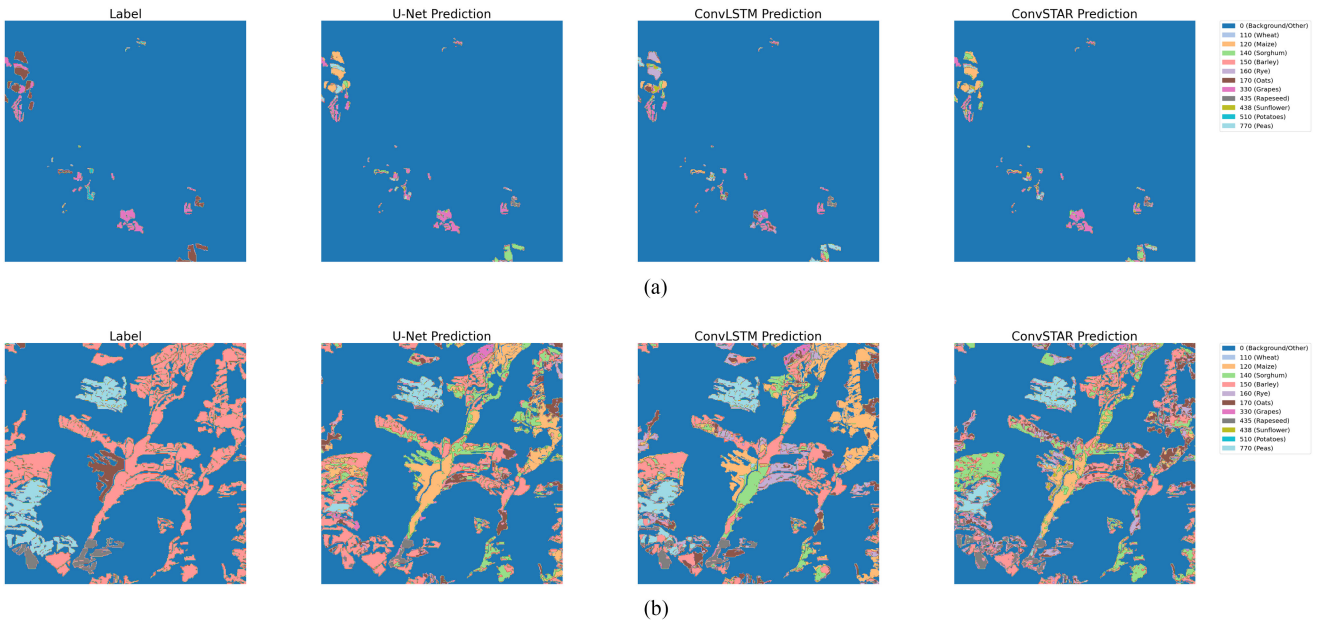

Fig. 8. Scenario 3. Visual evaluation of the U-Net, ConvLSTM, and ConvSTAR predictions for two different image patches.

misclassifications are widespread to all parcels in the selected patches.

B. Sen4AgriNet—OAD Experiment

In the object-based set of experiments, three architectures were utilized in order to evaluate model performance for the three different scenarios. The same patches as in the PAD section were selected in order to limit the number of differentiating factors among the two sets of experiments and focus on understanding the generalization performance of the two approaches.

The first architecture is a network consisting of three bidirectional LSTM [18] layers with a hidden layer of size 1024, alongside one linear layer (applied to the last hidden state), an ReLU, and a final classification linear layer. The entire model includes ~ 60 m trainable parameters.

Since their first introduction, transformer networks [28] are considerably valued as an extraordinary opponent to the LSTM networks. Thus, the second DL architecture is a transformer-encoder classifier that was designed to be tested against the well
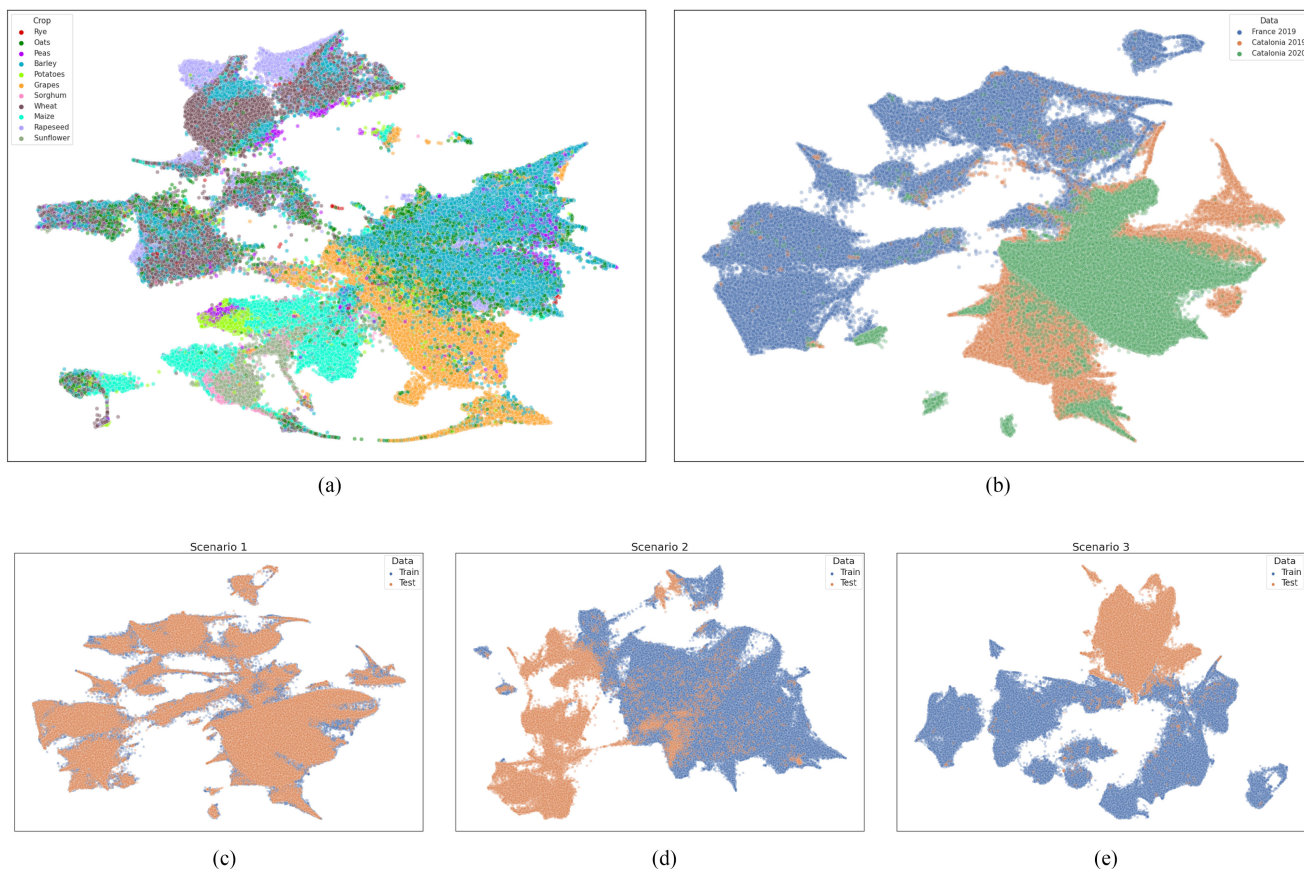
Fig. 9. UMAP visualizations for the examined datasets. (a) Distribution of the selected crops in all regions and years, (b) distribution of data across the three different datasets, (c) and (d) distribution of the train and test data in the three scenarios. Better viewed in colour.
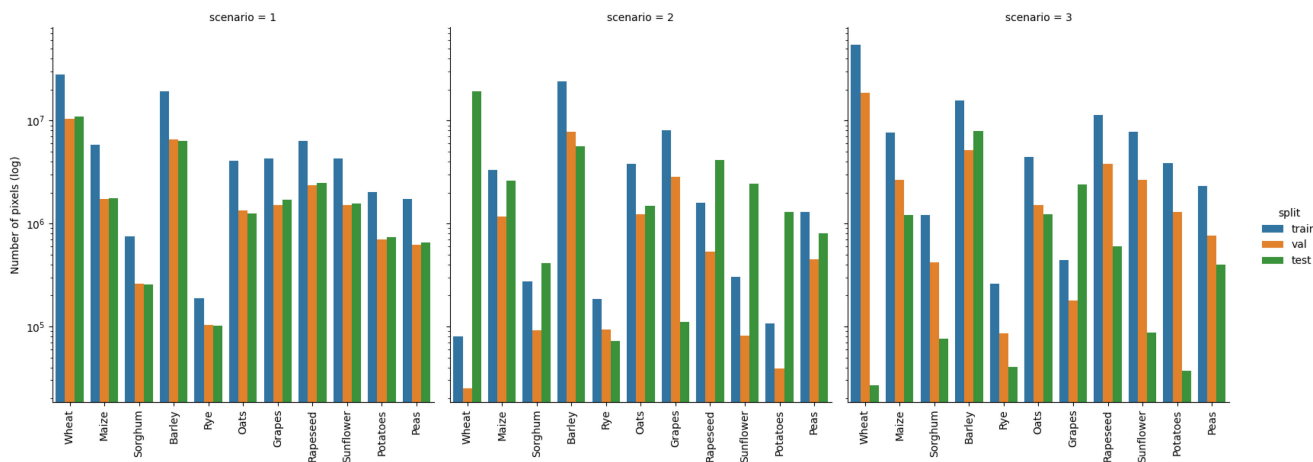

Fig. 10. Pixel distribution of the selected crop classes across the different scenarios. Vertical axis is in logarithmic scale.

known LSTM. Instead of calculating an embedding space, the precomputed parcel statistics are used as input. The dimensionality of the encoder-transformer type network uses a 26 features input (2 statistics for each band channel) and 2 heads, as heads should divide the number of features. For the forward function, a customized positional encoder was utilized. This positional encoder features the vanilla sine and cosine functions proposed by the authors of [28] for encoding each token position. After

encoding the position, the transformer encoder is applied and then a final sequence of two linear layers is used. The final model contains a total of ~ 270 k trainable parameters.

Other hyperparameters such as the number of encoder layers and the hidden dimension size are kept the same as in the LSTM runs (3 layers, 1024-D hidden size). Both experiments use the Adam optimizer with initial learning rates 0.001 and 0.0001, respectively, as well as a step learning rate scheduler
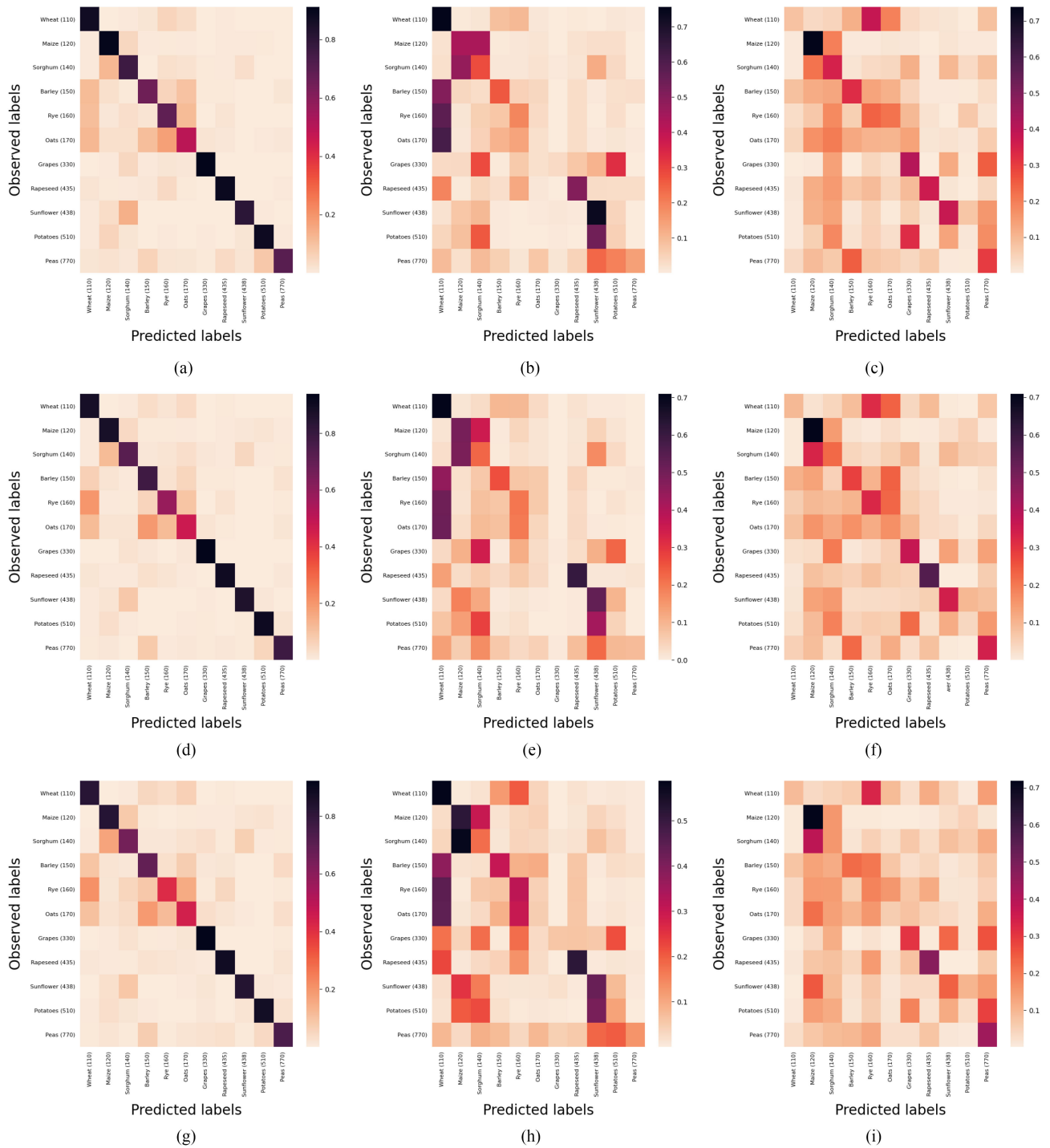
Fig. 11. PAD results. The first row shows the U-Net results, the second one the ConvLSTM results whereas the third the ConvSTAR results. Each column corresponds to a different scenario.

of 0.1 every 5 epochs. The cross entropy is used as the loss function.

Finally, a quite famous deep learning architecture in the remote sensing field is being marshaled. TempCNN [25], [9] with a total of ∼ 719 k trainable parameters, where convolution stages are designed to capture the time aspect of the OAD dataset.

Results obtained from model will be compared against each other for every test scenario.

1) Scenario 1: In the first scenario, the three different architectures (LSTM, transformer, TempCNN) were evaluated. All architectures displayed quite similar performance as shown in the metrics Table IV, both in the overall classification evaluation and in the individual classes (confusion matrices). The TempCNN architecture achieves slightly better results. As seen in the corresponding confusion matrices [see Fig. 12(a)–(g)] most of the crop type categories are classified correctly. Though, a
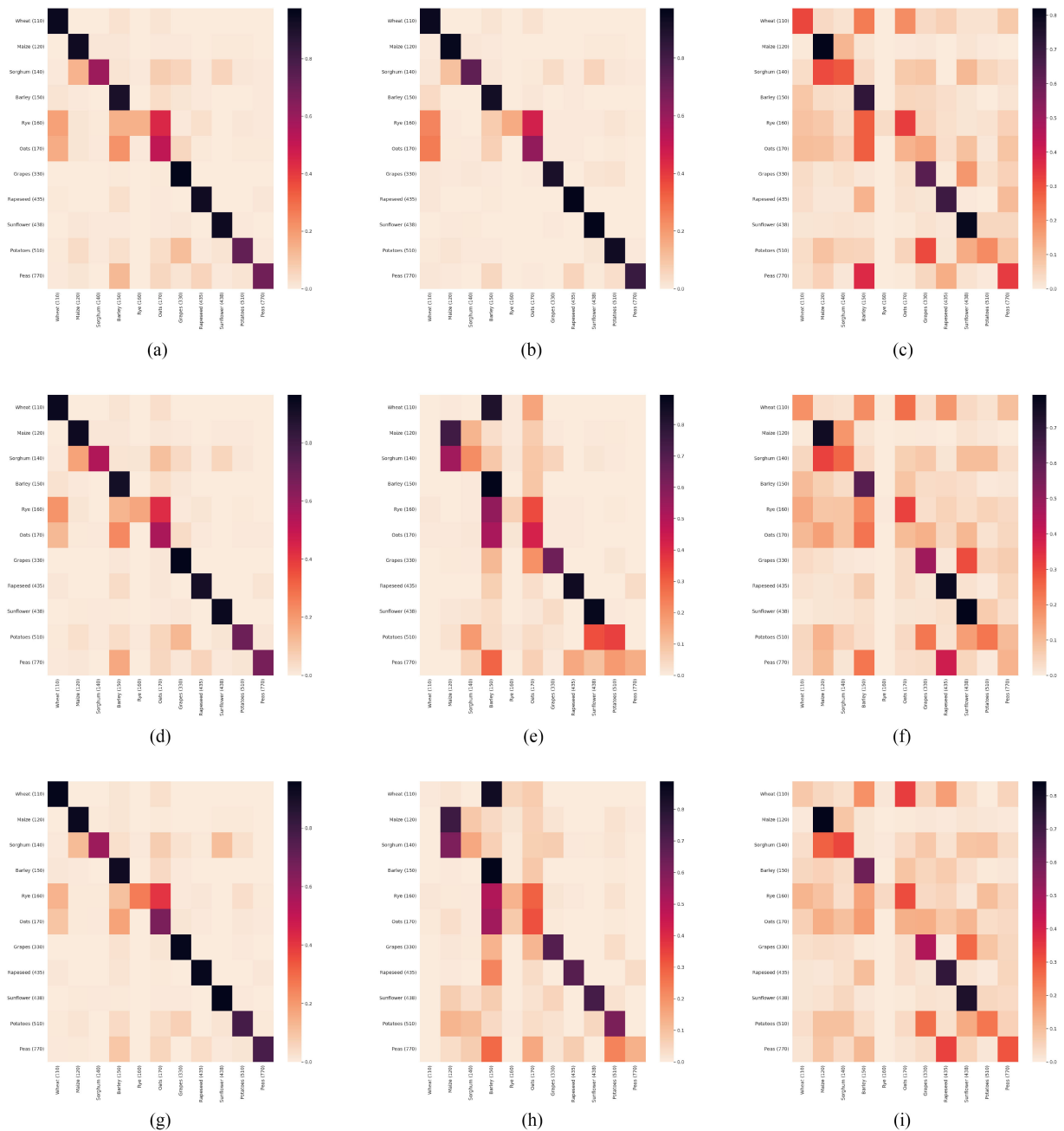
Fig. 12. OAD results. The first row shows the LSTM results whereas the second row the transformer-encoder classifier results. The third row illustrates results obtained for the TempCNN architecture. Each column corresponds to a different scenario.

specific misclassification pattern among some classes is observed in both architectures. Rye is mostly confused with wheat, barley, and oats, while oats are confused with rye, barley, and wheat. On the contrary, wheat is hardly confused with other classes. This behavior can be first attributed to the fact that wheat, rye, barley, and oats are crops with very similar spectral and temporal signatures. Second, wheat is a dominant class (39.4%), while the other three classes together account for 27.6% of the training data.

2) Scenario 2: In the second scenario, the training was performed on Catalonia for both available years and tested on France 2019 labeled data. Impressively, the LSTM model

significantly outperformed both the TempCNN and the transformer on all metrics (see Table IV). In the transformer and TempcNN confusion matrices [see Fig. 12(e) and (h)] high misclassification rates among the classes is apparent. In this scenario the LSTM architecture presented the same pattern with scenario 1, while the transformer performance was severely degraded. The adaptation of LSTM can be attributed to the fact that time is encoded as part of the feedback process, while transformers handle data as a sequence. This subtle difference on the time representation can have significant impact on the classification performance (transferring from one region to another), because the phenological cycles do not simply shift across time, but are

TABLE IV
RESULTS ON ALL SCENARIOS FOR OAD

Scenario	Model	Acc. W (%)	F1 W. (%)	Precision W. (%)
1	LSTM	88.52	88.03	87.85
	Transformer	88.36	88.10	87.90
	TempCNN	**90.08**	**89.97**	**90.01**
2	LSTM	**91.55**	**91.34**	**91.31**
	Transformer	39.17	31.45	58.52
	TempCNN	36.90	30.14	60.71
3	LSTM	**60.60**	**63.96**	**70.55**
	Transformer	51.21	56.71	67.76
	TempCNN	52.32	57.38	68.35

Best results are marked in bold.

also stretched or shrunk down depending on the soil-climate conditions of the area.

3) Scenario 3: In the third scenario, training was performed on France 2019 data and tested on Catalonia 2020 data. Similarly to the previous scenario, the LSTM model performed better than the transformer and TempCNN models, though not with the same performance gap (see Table IV). This degradation in performance can be attributed to the fact that in this scenario the training samples are drawn from a different distribution with respect to the test samples. The variation includes two components: the first component is that the growing season across the regions is different for the same crops (mainly due to different climatological conditions). The second component is that different years have slightly different starting dates for crop cultivation. Jointly, these factors may significantly affect the classification performance.

VII. DISCUSSION

Agricultural monitoring mainly includes tasks like crop type classification, parcel extraction and counting, and crop phenology evolution. Consistent acquisition of satellite multispectral time series of images are the stepping stone toward a complete agricultural monitoring process. The opening to the public of various LPIS systems, which contain crop type declarations and parcel geometries, bundled with the availability of consistent satellite measurements and advances in DL architectures enabled the creation of Sen4AgriNet. The first part of the challenge in order to materialize the dataset was the harmonization of the contextual information contained in the different LPIS. This challenge was solved by adapting a crop type classification structure inspired by FAO. The second part of the challenge was to build the dataset for multiple countries, bundling thousands of images with the corresponding labels in order to create a Machine Learning ready dataset. The challenge was met by creating a custom pipeline that automatically identifies the Sentinel-2 images that need to be downloaded from ESA's Scihub [60] and/or AWS Sentinel-2 bucket [61], downloads them, splits them to nonoverlapping patches, stacks the different time stamps, and finally attaches the labels. This procedure runs in an end to end fashion until the generation of the complete standalone netCDF4 files.

A subsequent step involves the creation of appropriate splits for training, validation and test, which is a nontrivial task. We implemented a stratified sampling procedure, based on the location, the class, and cultivation year of the parcels in our dataset,

selecting those that have a significant number of observations and are common in both countries. However, this balanced selection of parcels in our splits, which is different for each experimental scenario, is not necessarily mapped to a balanced selection of pixels instead of parcels. In Fig. 10, we show the pixel distribution of the different crop classes, for the three splits and for the three scenarios. Interestingly, in scenario 1 the balance parcel distribution is also mapped in the pixels distribution. In scenarios 2 and 3 though, this is not the case. For example, in scenario 2, the number of wheat pixels are two orders of magnitude less in the train set than in the test set. The inverse is observed in scenario 3. These discrepancies are due to the different agricultural practices between Catalonia and France. In Catalonia agriculture is more fragmented and the average parcel area is less in hectares compared to France. Therefore, balancing the splits based on parcel number is a good selection for the OAD set, while balancing the splits based on pixel count is a good selection for the PAD set.

Investigating the generalization potential of DL models vis-à-vis the three different scenarios and the two different datasets, there are a couple of observations. First, in scenario 1, the probability density functions of the train and the test set should be nearly identical, since in both sets the samples experience the same spectral, spatial, geographic, and temporal variability. This does not apply for scenarios 2 and 3. In scenario 2, the temporal variability is captured but not the geographic/climatic. In scenario 3 both the yearly and the climatic variability information is lost. Therefore, the probability density functions of the train and the test sets differ significantly. This explains the observed degradation in performance for both PAD (see Table III) and OAD (see Table IV) datasets. In order to fortify this observation, different UMAP [62] visualizations are provided in Fig. 9 which were computed using the OAD data. In Fig. 9(a), a sample of the selected crops from all regions and years is visualized and we can see that most crops form distinct clusters, with some observed and anticipated confusion between those with similar phenology, e.g., wheat and rapeseed, oats and barley, etc. Fig. 9(b) provides a visualization of the three different scenarios, where we observe a partial overlap between data from both years in Catalonia and a clearer divergence between Catalonia and France clusters. This indicates that both the temporal and more importantly the spatial diversity in Sen4AgriNet result in different data distributions. Finally, in Fig. 9(c)–(e) the deviations between the train and test sets of the different scenarios are illustrated. As expected, the two splits in scenario 1 are heavily overlapping, in contrast to those in scenarios 2 and 3 where separate clusters are formed with some minor overlap, due to the different year and/or location used in the train and test splits.

Second, the OAD-based experiments provide better results with respect to PAD, for all scenarios, especially for the second and third scenario. The enhanced classification performance and better generalization of the OAD was expected. OAD aggregates several pixels for each parcel and the mean and standard deviation of this aggregation are included as input features. On the contrary, the PAD architectures used pixel based processing, thus, the spatial heterogeneity is not well captured. In principle, our object based analysis smooths out the probability distributions in the spatial and temporal domain,

removing high-frequency components. In the extreme case of training on 2019 France data and testing on 2020 Catalonia data (scenario 3), this translates to bringing the two data distributions closer. However, PAD and OAD solve different DL problems and direct comparison is not entirely fair. OAD's increased classification accuracy comes at the expense of spatial resolution, while it completely misses out on capturing interparcel variability. Finally, it is understood that the main obstacle in further improving the generalization capacity of a trained model is the inclusion of additional regions and with higher yearly coverage, to capture the characteristics of nondominant classes with more observations and extract more meaningful features.

Regarding the PAD experiments, top accuracy was achieved as anticipated in the first scenario where regions and years are mixed, allowing the models to learn all temporal crop variations for both Catalonia and France. The most challenging classes to distinguish in scenario 2 are cereal-based crops, such as wheat, maize, sorghum, barley. This is attributed to their similar spectral content in the satellite time series. Finally, the third scenario experienced the worst performances, especially for ConvSTAR, which seems to be unable to discern the different classes with the sole exception of maize and rapeseed. Overall, U-Net shows more stable performance and manages to outperform the other two models in the most difficult scenarios (2 and 3) while achieving competitive results in the first scenario. This implies that the recursive nature of the ConvLSTM and ConvSTAR models does not offer much improvement in the studied cases. In addition, although ConvSTAR has considerably fewer trainable parameters and converges faster, this seems to inevitably come at the expense of accuracy. Lastly, ConvLSTM and U-Net employ a more complex encoder–decoder architecture, whereas ConvSTAR is a simple structure of three stacked ConvSTAR cells. According to the results of the different scenarios, this encoder–decoder scheme seems to be more efficient and achieve more robust results. ConvLSTM and U-Net manage to predict more compact crop areas with less internal variability, whereas ConvSTAR is more prone to predict multiple classes in a single parcel (see Figs. 6–8). The experiments suggest that ConvSTAR could benefit from an encoder–decoder architecture similar to ConvLSTM or a more thorough hyperparameter tuning. More complex models like DuPLO [23], BCDU-Net [63] or TempCNN [25] may be more robust to geographical/temporal differences and are left as a future plan for experimentation. As a last thought, further exploration and finetuning of the models' hyperparameters could potentially boost the accuracy of the predictions and help produce higher quality classification maps.

On the OAD experiments, all model architectures have unique structures. This results in different training challenges; RNN networks and transformers are notorious for their high training time requirements. The training time required for each epoch is quite lower for the transformer-encoder classifier due to 1) the tremendous difference in the number of total trainable parameters and 2) the limited parallelism in the nature of LSTM networks. However, despite the lower training time of transformer, in all training runs the LSTM model seemed to converge a lot faster requiring fewer epochs. Training time is not a issue when dealing with CNN-based models, as parallelism is quite

ubiquitous. This enables for far better training times, and thus, quicker convergence of the overall architecture.

Concerning the OAD reported results, the tests indicate that LSTM seems to perform better than the transformer. The transformer-encoder classifier seems to match the results of the LSTM in the first scenario, but it degrades dramatically in scenario 2. Transformers are considered as especially "data hungry" models, therefore using more data from Sen4AgriNet core dataset would provide a boost in the metrics. Finally, scenario 3 seems to be the middle ground for OAD, which was unexpected. This may be attributed to the "bidirectional" element of the LSTM cells which results in better performance since it enables them to learn in both directions.

The above analysis for the experiments conducted in this study suggests that direct application of a trained DL model in a different setting hinders performance and greatly affects the generalizability of the model, thus some adjustment is required beforehand. The discrepancies in the label distributions, along with the intrinsic spatiotemporal variation observed between train and test sets in the different scenarios, provide an ideal setting for the exploration of DA techniques. Especially for scenarios 2 and 3, the direct knowledge transfer between source and target domains results in a significantly degraded performance and low quality predictions. In addition, the interested researcher can further design new scenarios with different types of spatiotemporal variation, e.g., by varying only the temporal dimension for the same region or by employing different sets of labels for train and test. In any case, careful adaptation of the model from one domain to another will certainly bridge this data distribution gap and improve performance and generalizability.

Last but not least, the extension of Sen4AgriNet to other types of segmentation tasks is straightforward. Apart from the semantic segmentation task presented in this study, instance segmentation and panoptic segmentation problems can also be addressed. In our case, instance segmentation refers to the localization of a parcel with a surrounding box and the pixels belonging to this parcel and specific crop type, managing to differentiate between adjacent parcels with the same crop. Panoptic segmentation extends the tasks of instance and semantic segmentation by including labeling of the surrounding environment, e.g., forests, water bodies, wetlands, etc. Sen4AgriNet comes with a total of 168 labels including noncrop related classes which can serve as the broad "stuff" category in panoptic segmentation terminology, whereas parcel boundaries can assist the identification of each individual object in an image for both tasks. We believe that the proposed dataset will encourage further research in this area and provide the required momentum for the development of more models and approaches, especially as far as panoptic segmentation is concerned where research studies focused on remote sensing are few and far between (e.g., [11] and [64]).

VIII. CONCLUSION

The lack of harmonized labeled data among the paying agencies, which operationally gather country-wide labels every year, initiated the creation of Sen4AgriNet. Inspired by the FAO ICC nomenclature and adapted from the CAP and the remote sensing perspective, we proposed in this work the Sen4AgriNet crop type

classification scheme. We build the benchmark dataset by leveraging the availability of Sentinel-2 multitemporal, multicountry labeled data exploiting the recent opening up of LPIS parcel data. We construct a dataset that consists of 225 000 patches with corresponding pixel-based crop type maps. Based on this, we extracted two Sen4AgriNet subsets (PAD and OAD) for tackling different sets of classification problems, for unknown and known parcel geometries, respectively. The experiments were divided into three different scenarios to investigate the impact of diverse agricultural practices, climatic zones, phenology phases, crop spectral signatures across different regions and cultivation years. As expected, the changes of the probability distribution functions experienced when moving between geographic regions for training and testing DL models, has a significant impact on classification performance and limits the models' capacity to adapt and generalize.

We believe that Sen4AgriNet can be regarded as a labeled benchmark dataset, tailored for CAP and the use of Sentinel-2 imagery that come at no cost, and can spur numerous DL-based applications for crop type classification, parcel extraction, parcel counting, and semantic segmentation. More importantly, the dataset can be extended to include other input data sources, including Sentinel-1 synthetic aperture radar data, and meteorological data, allowing a new family of applications on early warning risk assessment and agricultural insurance.

References

[1] P. W. Owen *et al.*, "The land parcel identification system: A useful tool to determine the eligibility of agricultural land—But its management could be further improved.," Special Rep. 25, Oct. 2016.

[2] D. Sykas, I. Papoutsis, and D. Zografakis, "Sen4AgriNet: A harmonized multi-country, multi-temporal benchmark dataset for agricultural earth observation machine learning applications," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 5830–5833.

[3] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904. [Online]. Available: http://dx.doi.org/10.1109/IGARSS.2019.8900532

[4] M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon, "Sen2Cor for sentinel-2," in *Proc. Int. Soc. Opt. Photon. Image Signal Process. Remote Sens. XXIII*, vol. 10427, 2017, Art. no. 1042704.

[5] Copernicus, CORINE Land Cover. 2018. [Online]. Available: https://land.copernicus.eu/pan-european/corine-land-cover

[6] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019.

[7] X. X. Zhu *et al.*, "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [Software and Data Sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.

[8] R. E. F. (2020), "CV4A competition Kenya crop type dataset," Radiant MLHub, Version 1.0, Accessed: Dec. 2020. [Online]. Available: https://doi.org/10.34911/RDNT.DW605X

[9] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner, "BreizhCrops: A time series dataset for crop type mapping," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci. ISPRS*, 2020.

[10] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[11] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4872–4881.

[12] V. Sitokonstantinou, I. Papoutsis, C. Kontoes, A. Lafarga Arnal, A. P. Armesto Andrés, and J. A. Garraza Zurbano, "Scalable parcel-based crop identification scheme using sentinel-2 data time-series for the monitoring of the common agricultural policy," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 911.

[13] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[14] A. Stoian, V. Poulain, J. Inglada, V. Poughon, and D. Derksen, "Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 1986. [Online]. Available: https://www.mdpi.com/2072-4292/11/17/1986

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3D convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 75. [Online]. Available: https://www.mdpi.com/2072-4292/10/1/75

[17] L. Zhong, L. Hu, H. Zhou, and X. Tao, "Deep learning based winter wheat mapping using statistical data as ground references in Kansas and northern Texas, U.S.," *Remote Sens. Environ.*, vol. 233, 2019, Art. no. 111411. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425719304304

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[19] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sens. Environ.*, vol. 221, pp. 430–443, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425718305418

[20] T. He, C. Xie, Q. Liu, S. Guan, and G. Liu, "Evaluation and comparison of random forest and A-LSTM networks for large-scale winter wheat identification," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1665. [Online]. Available: https://www.mdpi.com/2072-4292/11/14/1665

[21] J. Xu *et al.*, "DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping," *Remote Sens. Environ.*, vol. 247, 2020, Art. no. 111946. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425720303163

[22] M. M. G. de Macedo, A. B. Mattos, and D. A. B. Oliveira, "Generalization of convolutional LSTM models for crop area estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1134–1142, Mar. 2020.

[23] R. Interdonato, D. Ienco, R. Gaetano, and K. Ose, "DuPLO: A dual view point deep learning architecture for time series classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 91–104, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271619300115

[24] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, 2014, pp. 1724–1734. [Online]. Available: https://doi.org/10.3115/v1/d14-1179

[25] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 523. [Online]. Available: https://www.mdpi.com/2072-4292/11/5/523

[26] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 421–435, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301647

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, Eds., vol. 30. Red Hook, NY, USA: Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[29] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, Jun. 2020, pp. 12 322–12331. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01234

[30] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS Int. J. Geo- Inf.*, vol. 7, no. 4, p. 129, 2018. [Online]. Available: https://www.mdpi.com/2220-9964/7/4/129

[31] M. O. Turkoglu, S. D'Aronco, J. Wegner, and K. Schindler, "Gating revisited: Deep multi-layer RNNs that can be trained," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3064878.

[32] M. O. Turkoglu *et al.*, "Crop mapping from image time series: Deep learning with multi-scale label hierarchies," *Remote Sens. Environ.*, vol. 264, 2021, Art. no. 112603. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425721003230

[33] B. Kellenberger, O. Tasar, B. Bhushan Damodaran, N. Courty, and D. Tuia, "Deep domain adaptation in earth observation," in *Deep Learning for the Earth Sciences*, 1st ed., G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, Eds., New York, NY, USA: Wiley, Sep. 2021, pp. 90–104. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/9781119646181.ch7

[34] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231218306684

[35] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5716–5726.

[36] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1358–1367.

[37] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425719303414

[38] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 945–954.

[39] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Computer Vision - ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 443–450.

[40] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 447–463.

[41] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1510–1519.

[42] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[43] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1249–1258.

[44] S. Chopra, S. Balakrishnan, and R. Gopalan, "Dlid: Deep learning for domain adaptation by interpolating between domains," in *ICML Workshop Challenges Representation Learn.*, vol. 2, no. 6. 2013.

[45] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971.

[46] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 95–104.

[47] M. Long, Z. CAO, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Red Hook, NY, USA: Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/ab88b15733f543179856008600245108dd8-Paper.pdf

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[49] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[50] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[51] H.-Y. Lee *et al.*, "DRIT: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, no. 10/11, pp. 2402–2417, Nov. 2020. [Online]. Available: http://link.springer.com/10.1007/s11263-019-01284-z

[52] Z. Wang, H. Zhang, W. He, and L. Zhang, "Phenology alignment network: A novel framework for cross-regional time series crop classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 2940–2949.

[53] M. Martini, V. Mazzia, A. Khaliq, and M. Chiaberge, "Domain-adversarial training of self-attention-based networks for land cover classification using multi-temporal Sentinel-2 satellite imagery," *Remote Sens.*, vol. 13, no. 13, Jun. 2021, Art. no. 2564. [Online]. Available: https://www.mdpi.com/2072-4292/13/13/2564

[54] J. Nyborg, C. Pelletier, S. Lefèvre, and I. Assent, "TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation," Nov. 2021. [Online]. Available: http://arxiv.org/abs/2111.02682

[55] Food and A. O. of the United Nations, *A System of Integrated Agricultural Censuses and Surveys. Volume 1, World Programme for the Census of Agriculture 2010*, (ser. FAO Statistical Development), no. 11. Rome, Italy: FAO, 2005.

[56] Government of Catalonia, Department of Agriculture, Livestock, Fisheries and Food, "Crop map with DUN origin in Catalonia," 2020. [Online]. Available: http://sac.gencat.cat/sacgencat/AppJava/organisme_fitxa.jsp?codi=4163

[57] F. Levavasseur *et al.*, "RPG explorer: A new tool to ease the analysis of agricultural landscape dynamics with the land parcel identification system," *Comput. Electron. Agriculture*, vol. 127, pp. 541–552, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169916305117

[58] D. Sykas, I. Papoutsis, and D. Zografakis, "Sen4agrinet: A harmonized multi-country, multi-temporal benchmark dataset for agricultural earth observation machine learning applications," 2021. [Online]. Available: https://github.com/Orion-AI-Lab/S4A

[59] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 802–810, 2015.

[60] Copernicus Open Access Hub, "Api Hub," 2021. [Online]. Available: https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/APIHub Description

[61] AWS Amazon, "AWS Sentinel-2 bucket," 2021. [Online]. Available: https://registry.opendata.aws/sentinel-2/

[62] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, 2018.

[63] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-net with densely connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 406–415.

[64] O. L. F. de Carvalho *et al.*, "Panoptic segmentation meets remote sensing," Nov. 2021. [Online]. Available: http://arxiv.org/abs/2111.12126

Dimitrios Sykas received the Rural, Surveying and Geoinformatics Engineering Dip.Ing. degree from the National Technical University of Athens (NTUA), Athens, Greece, in 2010, and the Ph.D. degree in hyperspectral and synthetic aperture radar (SAR) remote sensing from National Technical University of Athens (NTUA), Athens, Greece, in 2015.

He has been a Researcher Associate with the Laboratory of Remote Sensing, NTUA, from 2010 to 2015. Later on, he worked on several companies and institutions as R&D Researcher and Technical Manager. In 2020, he joined the National Observatory of Athens as PostDoc Researcher. He has significant expertise in the fields of photointerpretation, remote sensing, digital image processing, deep learning, and their applications. He has worked in 16 research programs related to remote sensing, published more than 20 peer-reviewed papers and is inventor of European patent on hyperspectral image compression.

Maria Sdraka received the M.Sc. degree in electrical and computer engineering (ECE) from the National Technical University of Athens (NTUA), Athens, Greece, in 2016. She is currently working toward the Ph.D. degree in focused on the application of AI techniques on Remote Sensing data for natural disaster monitoring with the Institute of Astronomy, Astrophysics, Space Applications and Remote Sensing (IAASARS), National Observatory of Athens (NOA), Athens, Greece, since 2021.

She is participating in two European research programmes related to damage assessment and monitoring of forest wildfires. Her research interests include the application of AI techniques on remote sensing data for earth observation tasks. She has worked on signal processing, data fusion, image enhancement, classification, and segmentation as well as change detection through the assistance of deep learning algorithms.

Dimitrios Zografakis received the B.Sc. degree in informatics from the University of Piraeus, Piraeus, Greece, in 2019 and the M.Sc. degree in data science and machine learning from the National Technical University of Athens, Athens, Greece, in 2021.

He has worked at various data engineering positions. His main research interests include large-scale data processing systems and computer vision algorithms.

Ioannis Papoutsis (Member IEEE) received the Diploma in electrical and computer engineering (ECE) from the National Technical University of Athens (NTUA), Athens, Greece, in 2002, the M.Sc. degree in technologies for broadband communications from the Department of Electronic and Electrical Engineering, University College London, London, U.K., in 2003, and the Ph.D. degree in remote sensing from NTUA in 2014.

In 2019, he was elected Associate Researcher with the Institute of Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece, where he leads OrionLab, a research unit of artificial intelligence for big earth observation (EO) data. He has been the Operations Manager with the Greek node of European Space Agency (ESA) Hubs that distribute Sentinel data. He has also acted as the Copernicus Emergency Management Services Manager for Risk and Recovery activations. He has participated and coordinated several research projects funded by the European Commission and ESA. His research interests include the exploitation, management and processing of big satellite data, and machine learning for knowledge extraction and fusion of multimodal EO data.