Big Earth Data

ISSN: 2096-4471 (Print) 2574-5417 (Online) Journal homepage: www.tandfonline.com/journals/tbed20

# DACIA5: a Sentinel-1 and Sentinel-2 dataset for agricultural crop identification applications

**A. Băicoianu, I. C. Plajer, M. Debu, M. Ștefan, M. Ivanovici, C. Florea, A. Cațaron, R. M. Coliban, Ș. Popa, Ș. Oprișescu, A. Racovițeanu, Gh. Olteanu, K. Marandskiy, A. Ghinea, A. Kazak, L. Majercsik, A. Manea & L. Dogar**

**To cite this article:** A. Băicoianu, I. C. Plajer, M. Debu, M. Ștefan, M. Ivanovici, C. Florea, A. Cațaron, R. M. Coliban, Ș. Popa, Ș. Oprișescu, A. Racovițeanu, Gh. Olteanu, K. Marandskiy, A. Ghinea, A. Kazak, L. Majercsik, A. Manea & L. Dogar (09 Jun 2025): DACIA5: a Sentinel-1 and Sentinel-2 dataset for agricultural crop identification applications, Big Earth Data, DOI: 10.1080/20964471.2025.2512685

**To link to this article:** https://doi.org/10.1080/20964471.2025.2512685

BIG EARTH DATA
https://doi.org/10.1080/20964471.2025.2512685

DATA NOTE

OPEN ACCESS Check for updates

# DACIA5: a Sentinel-1 and Sentinel-2 dataset for agricultural crop identification applications

A. Băicoianu[a], I. C. Plajer[a], M. Debu[a], M. Ștefan[b], M. Ivanovici[a], C. Florea[a,c], A. Cațaron[a], R. M. Coliban[a], Ș. Popa[a], Ș. Oprișescu[a], A. Racovițeanu[a,c], Gh. Olteanu[a], K. Marandskiy[a], A. Ghinea[b], A. Kazak[a,d], L. Majercsik[a], A. Manea[a] and L. Dogar[a]

[a]Department of Mathematics and Informatics, Transilvania University, Brașov, Romania; [b]National Institute of Research and Development for Potato and Sugar Beet, Brașov, Romania; [c]Department of Image Processing and Analysis Laboratory, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania; [d]Department of Telecommunications and Electronic Systems, Technical University of Moldova, Chișinău, Republic of Moldova

**ABSTRACT**

Artificial intelligence and data analysis are essential in smart agriculture for enhancing crop productivity and food security. However, progress in this field is often limited by the lack of specialized, error-free labeled datasets. This paper introduces DACIA5, a multispectral image dataset for agricultural crop identification, complemented with Sentinel-1 radar data. The dataset consists of 172 Sentinel-2 multispectral images ($800 \times 450$ pixels) and 159 Sentinel-1 radar images, collected over Brașov, Romania, from 2020 to 2024, with precise, in-situ verified labels. Additionally, 6,454 Sentinel-2 and 5,995 Sentinel-1 rectangular patches ($32 \times 32$ pixels) were extracted, exceeding 6 million pixels in total. The cropland parcels considered in our dataset are used for research and are owned and cultivated by the National Institute of Research and Development for Potato and Sugar Beet, ensuring error-free labeling. The labels in our dataset provide detailed information about crop types, offering insights into crop distribution, growth stages, and phenological events. Furthermore, we present a comprehensive dataset analysis and two key use cases: crop identification based on a "past vs. present" approach and early crop identification during the agricultural season.

**ARTICLE HISTORY**
Received 24 February 2025
Accepted 18 May 2025

**KEYWORDS**
Sentinel-2 data; Sentinel-1 data; smart agriculture; artificial intelligence; crop identification; early crop identification

## 1. Introduction

With the growth of the world population (Ritchie & Rodes-Guiaro, 2024), the threat on food security induced by climate changes, conflicts, and the need for biodiversity protection, it becomes increasingly important to optimize the usage of resources for agricultural purposes (Anderson et al., 2020; Frona et al., 2019; Maja & Ayano, 2021). Precision

**CONTACT** A. Băicoianu a.baicoianu@unitbv.ro Department of Mathematics and Informatics, Transilvania University of Brașov, 500091, Romania

Supplemental data for this article can be accessed online at https://doi.org/10.1080/20964471.2025.2512685

agriculture, supported by the latest technological developments such as artificial intelligence (AI), offers a promising solution to these challenges (Ivanovici, Baicoianu, et al., 2024b; Marandskiy & Ivanovici, 2024). In this context, intense research on using AI and smart systems for crop monitoring has been done in recent years (Akkem et al., 2023). However, one of the main problems in using AI models in general, and machine learning (ML) and deep learning (DL) in particular, is the large amount of training data required for accurate detection and classification (Benami et al., 2021). In agreement with prior work (Alzubaidi et al., 2023), we observed that freely available labeled data is often insufficient. Moreover, in agricultural data sets, many labels rely on self-reported information from farmers, which sometimes is not accurate or misleading. This can induce serious biases in the resulting learning-based models (Cabrera et al., 2014). While ML and DL models can cope with erroneous labels and data, this robustness is limited (Song et al., 2022).

In this direction, while several datasets are available for land cover and slightly towards agricultural crop monitoring, they frequently exhibit limitations that prevent them from adequately addressing the specific requirements of precision agriculture.

One notable dataset is BigEarthNet (Sumbul et al., 2019), a large-scale benchmark archive comprising Sentinel-2 image patches spanning 10 European countries. Although comprehensive, the BigEarthNet dataset is primarily designed for multi-label remote sensing image classification, with a particular focus on land cover classes. Since it contains classes such as “agricultural land” and “pasture”, it lacks detailed crop type annotations, verified through in-situ validation, which represents a significant limitation for precise agricultural applications. Similarly, EuroSAT (Helber et al., 2019) provides a dataset based on Sentinel-2 imagery with a focus on land use and land cover classification. Although the dataset includes agricultural areas among its categories, it lacks the specific crop type information and multi-temporal data necessary for monitoring crop growth stages and phenological events. The SEN12MS dataset (Schmitt et al., 2019) provides also a selection of georeferenced multi-spectral Sentinel-1/2 imagery, suitable for use in deep learning and data fusion. While the SEN12MS dataset is useful for a range of remote sensing applications, it does not focus on agricultural parcels with accurate crop labels over multiple years. This limits its applicability for detailed crop monitoring tasks. Additionally, the CropHarvest dataset (Tseng et al., 2021) is a global repository for crop type classification, aggregating data from multiple sources. However, CropHarvest frequently relies on crowd-sourced labels and lacks the desired level of accuracy for precise AI modeling in agriculture. Additionally, the dataset may exhibit inconsistencies in temporal coverage and spatial resolution, which are essential for analyzing crop dynamics over time. A recent pixel-based dataset, TimeSen2Crop (Weikmann et al., 2021), includes over a million labeled multispectral Sentinel-2 pixels collected across Austria over two consecutive years, covering 16 different crop types. The labels were derived from farmers’ declarations. A notable feature of this dataset is the inclusion of reports on snow, shadows, and cloud coverage for each labeled unit.

The exhibited limitations, including the absence of accurate, in-situ validated crop type labels, insufficient multi-year data over the same parcels, the reliance on less precise crowd-sourced labels, and lack of regional specificity, underscore the need for a new dataset that effectively addresses the specific requirements of smart agriculture.

In order to help the international community and enrich the available data, we generated a new dataset for crop monitoring, certification, and identification/detection,

which has the important advantage of accurate labels, based on in-situ institutional validation. Moreover, the dataset was constructed on the same parcels over the interval of 5 years, thus including variability resulting from different climatic conditions. The construction and validation of the data set was done in collaboration with the National Institute of Research and Development for Potato and Sugar Beet from Brașov, Romania (INCDCSZ, n.d.). To enhance clarity and readability, the National Institute of Research and Development for Potato and Sugar Beet will be referred to as the *Potato Institute* throughout the remainder of this article.

## 2. Methods

The dataset presented in this study is based on data acquired by Sentinel-2 optical instruments (Drusch et al., 2012) as well as on Sentinel-1 radar data (Torres et al., 2012). This section provides a comprehensive description of the procedures and steps used in producing the dataset. It includes a detailed account of the experimental design and the computational processing involved. Specifically, we present the geographical and climate context, describe the experimental setting in relation to Sentinel-1 and Sentinel-2 data, and outline the process of dataset creation, starting from the engineering data, and detailing the tools utilized.

### *2.1. Geographical and climate setting*

The land of the Potato Institute from Brasov, where the metadata for our dataset was collected, is part of the Bârsa premontane plain, which is a vast depression in Romania, surrounded by mountains. Bârsa is a part of the so-called "Potato country" in Romania. From a geographical perspective, Bârsa area extends from 45°27′ to 46°00′ north latitude and from 26°10′ to 26°13′ east longitude and has a surface of 2406 $km^2$ in extent. The altitude varies from 550 m at Bod village to 722 m at Zărnești city. This region experiences a climate transitional between Mediterranean and continental types. According to the Köppen classification (Beck et al., 2023; Peel et al., 2007) it falls into the climate region Dfb (cold, without dry season, warm summer). This climate is characterized by cold winters with significant snowfall and moderate precipitation throughout the year, as defined by the Köppen climate symbols and criteria. The region enjoys relatively warm springs, mild summers with temperatures suitable for potato cultivation, and long, sunny autumns extending into November. The humidity, indicated by an aridity index ranging from 34.8 to 40.8, supports favorable growing conditions for crops like potato and sugar beet. Bârsa has relatively low-to-medium annual precipitation (548–782 mm) due to its mountainous surroundings which absorb the fallen precipitation, with most rainfalls occurring in winter and drier periods in spring. Summer rainfalls are 250–300 mm. The region's average annual temperature of 7–8°C and summer temperatures of 15–17°C are suitable for potato cultivation, though recent droughts have increased the need for irrigation.

The soils within the Potato Institute belong to the mollic, hydromorphic, and weakly developed soil classes. It predominates the chernozimoid soil type, with a high humus content in the A horizon and a very large reserve in the first centimeters, moderately acidic reaction, good nitrogen supply, nitrification conditions negatively influenced by the acid reaction, medium insurance with mobile

forms of assimilable phosphorus and potassium, the degree of saturation in bases is high (approx. 70%), thus determining a good microbial activity. Under these conditions, the soil is favorable for potato cultivation, with a quality score of 81 out of 100. As noted in Teaci (1980), this score is based on an evaluation of the main climatic parameters (temperature, precipitation), edaphic factors (physical, chemical, and relief characteristics), and their ecological favorability for various agricultural crops. This type of information is highly relevant for the agricultural domain and experts. Combined with the publicly available dataset we provide researchers and agricultural professionals with supplementary information, to allow a more in-depth analysis and the evaluation of the impact of various parameters on both the quantity and quality of agricultural production.

From the flora point of view, the institute is located in an area where the spontaneous vegetation is represented by forest associations where *Quercus* and *Fagus* species predominate, mixed with shrubs (*Crataegus sp., Cornus sp*.) and herbaceous species, like grasses (*Lolium sp., Poa sp*.) and legumes (*Trifolium sp*.). Main crops include potato, sugar beet, cereals, legumes, and fodder plants. The flat terrain and soil texture support the full mechanization of these crops.

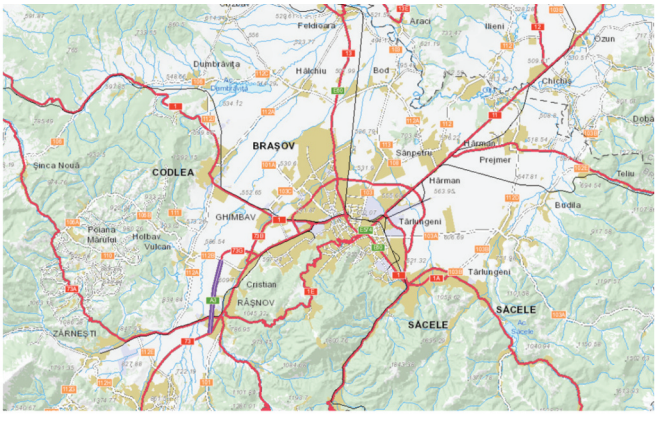
To provide a comprehensive view of the natural characteristics of the Brașov region, we have included a series of three thematic maps, each highlighting different aspects of the area, including height profile, soil types, and land cover, see Figure 1. For the creation of the height profile map, data were sourced from the official platform provided by the National Agency for Cadastre and Land Registration of Romania (https://geoportal.ancpi.ro/). The soil types map, depicted in Figure 1b, was generated utilizing information from (Liedekerke et al., 2006; Panagos, 2006; Panagos et al., 2022). The land cover map was produced using resources available on the Copernicus Land Monitoring Service platform (https://land.copernicus.eu/).

### 2.2. Dataset acquisition

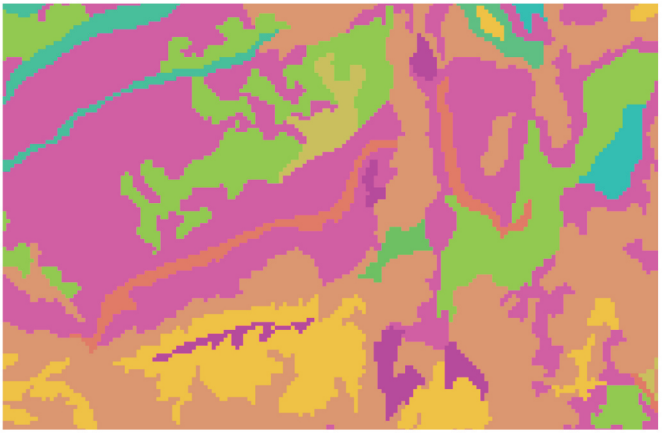
AI models, particularly in recent deep learning frameworks, require a lot of data (Borisov et al., 2024). In agriculture, a significant source of big data comes from remote sensing, especially through satellite imagery. More and more satellite missions offer free data for the international community. In this context, one of the most significant is the Copernicus Earth Observation Programme (Aschbacher, 2017). This is a European Union program using a constellation of satellites called "Sentinels" to gather extensive environmental data, which is freely available to users worldwide.

The Sentinel-2 satellites (Gascon et al., 2014), launched in 2015 and 2017, offer 13 spectral bands with spatial resolutions ranging from 10 m to 60 m (Gascon et al., 2017), enabling detailed analysis of vegetation, soil, and water. These satellites operate in a Sun-synchronous orbit, maintaining consistent sunlight angles for accurate, shadow-minimized imagery, with a 100-min orbit period at an altitude of 786 km and a 290 km swath width (n.d.).

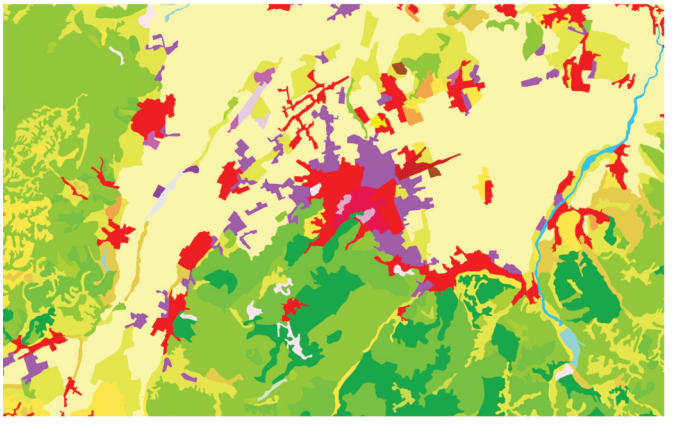
The dataset was obtained by downloading Sentinel-2 products from the Copernicus Browser platform (Copernicus Browser, 2024). The data is free and can be downloaded after creating an account. The platform provides data from all Copernicus missions. Users can view and download data specifically from the Sentinel-2 mission, at processing level




a) Height profile map




b) Soil types map




c) Land cover map

**Figure 1.** Illustrating maps with the characteristics of the Brașov area, including latitude and longitude, legend, and scale.

**Table 1.** Sentinel-2 bands with the central wavelengths (in nanometers) and their description.

| Band | Spatial Resolution (m) | Central Wavelength (nm) | Description |
|---|---|---|---|
| B1 | 60 | 443 | Aerosol |
| B2 | 10 | 490 | Blue |
| B3 | 10 | 560 | Green |
| B4 | 10 | 665 | Red |
| B5 | 20 | 705 | Vegetation Red Edge 1 |
| B6 | 20 | 740 | Vegetation Red Edge 2 |
| B7 | 20 | 783 | Vegetation Red Edge 3 |
| B8 | 10 | 842 | NIR |
| B8A | 20 | 865 | Narrow NIR |
| B9 | 60 | 945 | Water Vapor |
| B10 | 60 | 1375 | Cirrus Clouds |
| B11 | 20 | 1610 | SWIR1 |
| B12 | 20 | 2190 | SWIR 2 |

2A. We downloaded the data covering the period from 2020 to 2024. The downloaded images are from the area of the Research and Development Institute of Transilvania University of Brașov (Research and Development Institute of Transilvania University of Brașov, 2024) located at latitude: 45.669410 and longitude: 25.549550. The downloaded products are all at processing level 2A. This processing level is applied to the raw level 1C data with atmospheric correction. During atmospheric correction, the band number 10 (wavelength 1373 nm) is removed (Sentinel-2 processing, n.d.).

In Table 1 we provide an overview of the spectral bands specific to Sentinel-2 images as presented in (Pour et al., 2023). It should be noted that Band 10 was excluded from the level 2A products during the process of atmospheric correction. Since our dataset was generated using level 2A products, the images comprise Bands 1–9 and Bands 11–12. As can be observed in Table 1, the original Sentinel-2 bands have different spatial resolutions. In order to uniformize the bands, they were all sampled to the same spatial resolution of $10 \times 10$ m, using the Sentinel Application Platform (SNAP) default Nearest Neighbour algorithm (SNAP – Resampling Methods, n.d).

The image selection process was carried out using the Copernicus Browser platform, which allows filtering Sentinel-2 satellite images based on the maximum cloud coverage parameter (Max. cloud coverage). To ensure a balance between data availability and quality, we set a threshold of 30% cloud coverage, meaning that only images with a cloud cover below this percentage were considered. It is important to note that this threshold applies to the entire satellite product, not exclusively to our area of interest. Regarding the use of the Scene Classification Layer, while the initial filtering was performed using the Max. cloud coverage parameter, an additional manual selection step was necessary. This involved visually inspecting the pre-selected images to ensure that the area of interest was free from cloud shadows and other obstructions that could affect the analysis. The downloaded products were visualized with the Sentinel Application

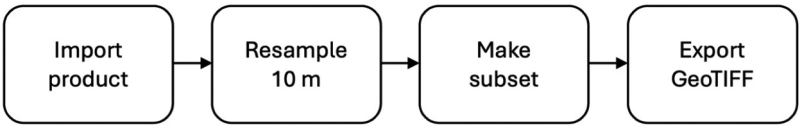



**Figure 2.** Preprocessing steps in Sentinel Application platform (SNAP) – European Space Agency (ESA).

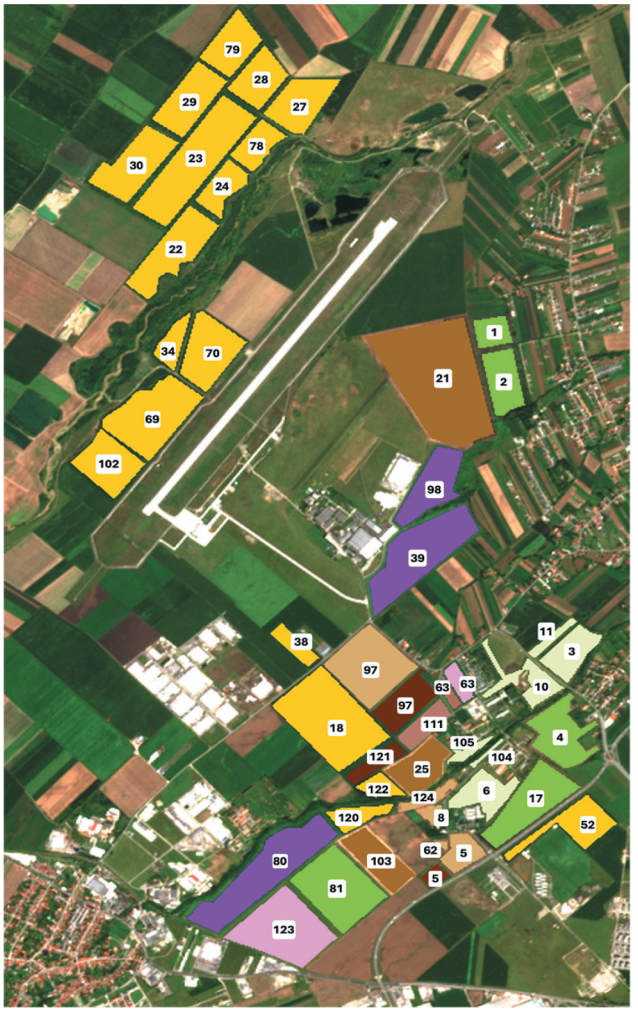



**Figure 3.** Example of the Potato Institute map of agricultural crops for 2023. The map presents the delimitation of each parcel. The parcels are identified by the corresponding parcel number and color. The colors were assigned using the color palette proposed by the US Department of agriculture (United States Department of Agriculture, 2024).

Platform (SNAP) (Sentinel Application Platform, n.d.) which is a free, open-source platform developed by the European Space Agency (ESA) (European Space Agency, n.d.). Using the ESA SNAP, the Sentinel-2 mission products were visualized, and the steps in Figure 2 were applied in order to obtain images that encompass the area of interest (AOI).

The subset of the initial product was realized using the coordinates inferred after the identification of the Research and Development Institute of Transilvania University of Brașov. The size of the images was chosen to encompass all parcels at the Potato Institute. The AOI was identified using maps from the Potato Institute. An example map can be seen in Figure 3. The maps were used to identify the parcels cultivated by the Potato Institute. Parcel marking was done manually using a software program we developed in MATLAB. The application provides the possibility to read GeoTIFF images from Sentinel-2. To

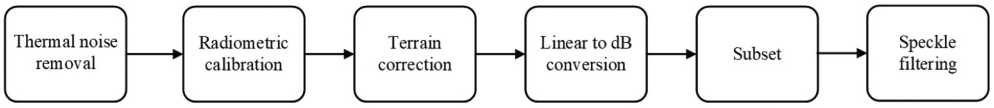



**Figure 4.** Processing steps for the Sentinel-1 data using the Sentinel-1 Toolbox.

visualize the image, the red, green, and blue (RGB) color channels were selected. In order to produce a color composite image to be displayed, the image values were scaled in the range [0, 255] and a brightness enhancement was performed.

To identify and save the parcel coordinates, an image was chosen where the parcel boundaries were clearly visible. Parcel delineation and coordinate extraction were done with *roipoly* MATLAB function on images acquired in summer, where the parcels can best be observed. Saving the parcel coordinates was done by clicking on the corners of the parcels in the RGB image and saving the coordinates of these points in .txt files. After saving the coordinates of all the parcels that exist on the Potato Institute's maps, masks were generated to cover each year's parcels. These masks were created based on the .txt files, which contained the necessary spatial information. The details regarding the number of these files have been provided in Table 3.

The Sentinel-1 mission, part of the European Space Agency's Copernicus programme, has been widely used in agricultural applications due to its all-weather, day-and-night Synthetic Aperture Radar (SAR) capabilities, and it operates in the C-band at 5.405 GHz. SAR uses the motion of the satellite platform to synthetically increase the aperture of the Radar, which allows high-resolution acquisition of images. SAR backscattering is affected by the structure and the water content of the soil surface and plants. In this dataset, we exclusively used Sentinel-1 images that are temporally close to corresponding Sentinel-2 images in our set. For each Sentinel-2 image, we selected a corresponding Sentinel-1 SAR image acquired within a three-day window, either up to 3 days before or after the Sentinel-2 acquisition date. We retrieved Sentinel-1 SAR images that are acquired in Interferometric Wide (IW) swath mode and included both VV (vertical transmit and vertical receive) and VH (vertical transmit and horizontal receive) polarizations with 10-m spatial resolution from Google Earth Engine's "COPERNICUS/S1_GRD" collection. These images have already been preprocessed using the Sentinel-1 Toolbox (n.d.). This includes detection, multi-looking, projection to ground range using an Earth ellipsoid model (e.g. WGS84), thermal noise removal, radiometric calibration, terrain correction (using SRTM 30 m or ASTER DEM for areas above 60° latitude) where images were also reprojected to a UTM coordinate system, and, finally, conversion of the terrain-corrected values to decibels ($10 \times \log_{10}(x)$). We further processed them after retrieval by applying a Lee Filter (Lee, 1980) to reduce speckle noise. Figure 4 illustrates the processing steps for the Sentinel-1 images.

## 2.3. Crop description and color scheme

A MATLAB script was used for the generation of masks based on the coordinates saved in .txt files. The files are saved in different directories for each year. Although the parcel

**Table 2.** The color palette for the masks. The first column lists the crop names, the second column provides the specific APIA identifier for each crop, the third column shows the labels assigned by us, and the final column indicates the crop color as defined by the United States Department of Agriculture.

| Name crops | APIA code | Label | Color | Color code (RGB) |
|---|---|---|---|---|
| Common winter wheat | 101 | 1 | | (165, 112, 0) |
| Common spring wheat | 1010 | 2 | | (217, 181, 107) |
| Corn | 108 | 3 | | (255, 211, 0) |
| Peas | 151 | 4 | | (84, 255, 0) |
| Late potatoes | 253 | 5 | | (112, 38, 1) |
| Other potato crop | 254 | 6 | | (125, 46, 7) |
| Potatoes for seed | 255 | 7 | | (180, 112, 91) |
| Potatoes for seed | 2557 | 8 | | (189, 125, 106) |
| Sugar beets | 3017 | 9 | | (167, 0, 228) |
| Temporal grassland | 450 | 10 | | (233, 255, 191) |
| Alfalfa | 9748 | 11 | | (255, 165, 226) |
| Permanent grassland | 606 | 12 | | (233, 255, 191) |
| Corn silage | 131 | 13 | | (255, 211, 0) |
| Soybean | 2037 | 14 | | (38, 112, 0) |
| Alfalfa | 9747 | 15 | | (255, 165, 226) |
| Winter rapeseed | 202 | 16 | | (209, 255, 0) |
| Sunflower | 123 | 17 | | (255, 255, 0) |

coordinates remain largely the same, the crops change from year to year. The masks were refined based on the information received from the Potato Institute. Each year, all the parcels were analyzed and checked on the maps received in order to avoid positioning errors or mistakes in specifying a crop.

All the files used in mask generation follow a naming convention. Each name is of the form "*a_b*" where *a* represents the parcel number and *b* the unique code given to a crop by the Payments and Intervention Agency for Agriculture (APIA) (n.d.) in Romania (see Table 2).

For example, the file with the name "1_151.mat" contains the multispectral data from parcel 1 and the crop with the code 151 corresponding to green peas. For each year, we generated masks for each parcel, with labels corresponding to the specific crops of that year. The visual representation of the masks enables users to easily identify the crops that were cultivated on the parcels by the Potato Institute in the respective year by

**Table 3.** Overview of acquired Sentinel-1 and Sentinel-2 images from DACIA5 dataset.

| Year | Number of images of Sentinel-1 | Number of images of Sentinel-2 | Number of text files | Number of crops | Number of patches of Sentinel-1 | Number of patches Sentinel-2 |
|---|---|---|---|---|---|---|
| 2020 | 40 | 40 | 51 | 10 | 1520 | 1520 |
| 2021 | 32 | 32 | 53 | 12 | 1380 | 1380 |
| 2022 | 25 | 30 | 47 | 12 | 914 | 1121 |
| 2023 | 28 | 30 | 47 | 12 | 1289 | 1410 |
| 2024 | 34 | 40 | 48 | 11 | 892 | 1023 |

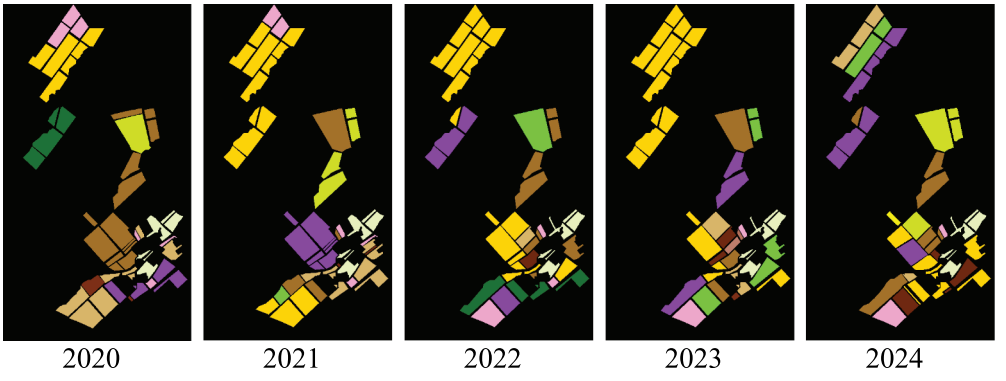



**Figure 5.** The masks with the parcels and the crops from 2020 to 2024, provided also as shapefiles in the DACIA5 dataset.

their specific colors. For the masks, we used the color palette proposed by the United States Department of Agriculture, which is detailed in Table 2. The colored masks allow us to observe the distribution and variety of the crops. The study of the masks also shows the crop rotation from year to year. The resulting masks are shown in Figure 5. The Potato Institute has dedicated parcels for research studies. These parcels are the same every year, but the crops are changed according to research needs. The APIA corresponding crop codes for the parcels from 2020 to 2024 are provided in the Appendix. It is important to mention that not all 17 crops are present every year.

In Table 3 we present the number of Sentinel-1 and Sentinel-2 images acquired in each year from 2020 to 2024, the number of text files containing the coordinates of the parcels, the number of crops in each year, as well as the number of patches generated from the Sentinel-1 and, respectively, the Sentinel-2 images. The number of text files differs in the different years, as some of the parcels were divided in some years into subparcels.

## 3. Data records

In this section, we provide detailed information about the data files, format (GeoTIFF), and other aspects that should ease the work with the data. The database contains all the available images from Sentinel-2 MSI and the corresponding Sentinel-1 SAR spanning over a time period of 5 years, from 2020 to 2024, covering the area of interest, as

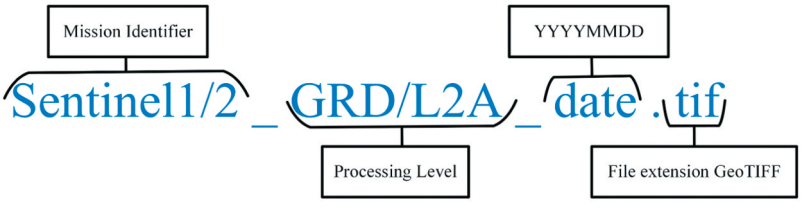



**Figure 6.** Image naming convention. YYYY is the four-digit year, MM is the two-digit month, and DD is the two-digit day.

described in Section 2.1. All the images respect a naming convention and are saved in GeoTIFF format. This GeoTIFF format can be read in MATLAB and Python. The naming convention is detailed in Figure 6. The Sentinel-2 images have dimensions of $800 \times 450 \times 12$, meaning each image consists of 800 pixels in height, 450 pixels in width, and 12 spectral bands, with a spatial resolution of $10 \times 10$ m. Similarly, Sentinel-1 images have dimensions of $800 \times 450 \times 2$, comprising 800 pixels in height, 450 pixels in width, and 2 bands, also with a spatial resolution of $10 \times 10$ m.

The Sentinel-1 data is stored in a folder named *Images_Sentinel1_2bands_GeoTIFF*, while folder *Images_Sentinel2_12bands_GeoTIFF* contains Sentinel-2 data. Within each of these folders, the acquired images are organized into subfolders corresponding to the year in which they were captured by the Sentinel-1 SAR sensors and the Sentinel-2 optical sensor, respectively. The folders containing these images have the names *Sentinel1_yyyy* and *Sentinel2_yyyy*, where *yyyy* represents the four-digit year. The directory *Images_Sentinel2_GeoTIFF* contains all the Sentinel-2 images as the previous Sentinel-2 folder, but these images have all the information provided by the Copernicus browser. Additional information is provided in Table 3, which displays the number of images from each year. The file naming convention is depicted in Figure 6.

In addition to the Sentinel-1 SAR data and the Sentinel-2 multispectral images, the database contains the ground truth of agricultural crops as RGB masks in PNG format and the masks with labels corresponding to each agricultural crop in both PNG and MAT formats. These are located in the *Masks_and_legend* directory. This directory also contains the legend for the masks in PDF format and the five subdirectories where the masks for each year are stored. The subdirectories are named *Masks_yyyy*, where *yyyy* represents the four characters for the corresponding year.

Using the aforementioned Sentinel-1 and Sentinel-2 imagery, we generated SAR and multispectral patches with dimensions of $32 \times 32 \times 2$, respectively, $32 \times 32 \times 12$,

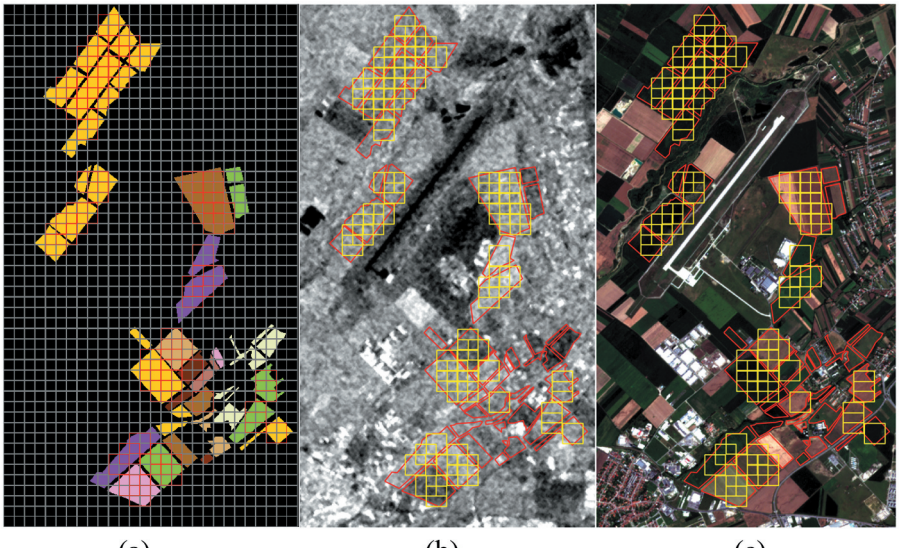



**Figure 7.** Selection grid for the patches.

which are systematically organized within the database in the *32×32_patches* folder. For automatic patch selection, the chosen rectangular areas of both the multispectral and SAR images were swept from the top-left corner to the bottom-right corner using a $32 \times 32$-pixel mask with a step of 16 pixels. Simultaneously, the label mask corresponding to each year was also traversed to count the number of pixels belonging to a specific crop within each identified patch. For each selected patch, a validity check was performed. A patch is considered valid if it contains at least 75% (768 out of 1024) pixels from a specific crop. The patches deemed valid were saved with a specific naming convention (which can be referenced). The position of the patches as a consequence of the traversal process can be observed in Figure 7(a) as well as the position of the resulting patches. In Figure 7(b) and (c), the overlay of patches on a Sentinel-1 image and a Sentinel-2 image, respectively, can be observed. This figure shows that the overlay is identical in both images, with each patch from Sentinel-1 corresponding to a patch from Sentinel-2, covering the same area.

After generating all the patches from the multispectral images, we examined their band selection RGB visualization, in order to filter the patches of interest for the training model, specifically those that contain visible crops (that have sprouted) and not bare soil. To generate the radar patches we used the same method as in case of multispectral patches. To filter them, we used the already examined RGB patches generated from multispectral images. An example of a radar and multispectral patch can be seen in Figure 8.

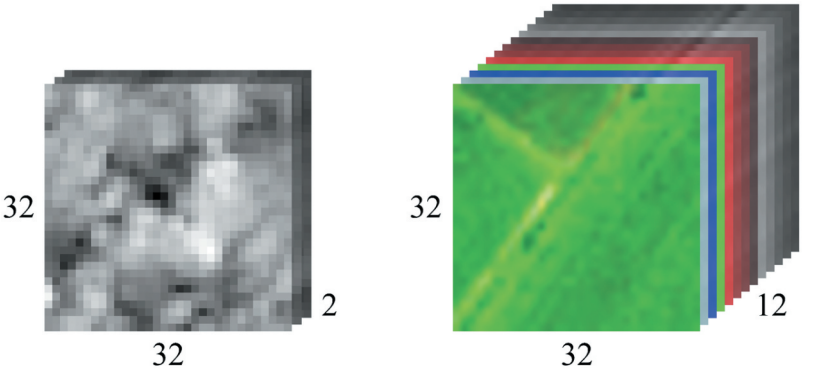



**Figure 8.** Visualization of a Sentinel-1 patch (left) and an RGB Sentinel-2 patch (right). The patches represent the same area on the ground.

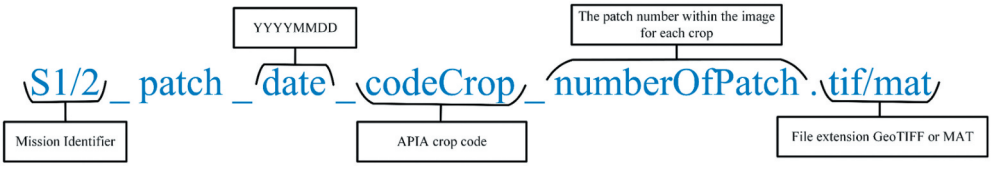



**Figure 9.** The naming convention for patches.

Each patch is named according to a structured convention, as exemplified in Figure 9. The practical use of these patches is demonstrated in Section 5, where they are employed to address two specific problems.

The *32×32_patches* directory is subdivided into two subdirectories: *32×32_SAR +MSI_patches* and *32×32_RGB_patches*, with the latter intended solely for visualization purposes. The first subdirectory contains radar and multispectral data utilized for addressing the two problems: *problem1* corresponds to the agricultural crop identification (ACI) with temporal generalization, while *problem2* focuses on the agricultural crop early identification (ACEI). For each problem, the data is further divided into training and test sets. The patches in these directories are georeferenced and are saved in both GEOTIFF and MAT formats, as indicated by the names of the subdirectories where they are stored: *sentinel1_patches_mat, sentinel1_patches_tiff, sentinel2_patches_mat,* and *sentinel2_-pacthes_tiff*. For the fused data scenario discussed in Section 5, we included a CSV file for training and testing, which maps the correspondence between Sentinel-1 patches and Sentinel-2 patches.

The directory *32×32_RGB_patches* contains the patches generated from the RGB masks, as a ground truth for the labels at pixel level. Based on these, we generated the patches from each image. Each patch contains an agricultural crop, which must occupy at least 75% of the pixels of the patch for it to be considered. For each year, there is a subdirectory where the patches from the masks are saved, named *patches_yyyy*, where *yyyy* represents the four characters corresponding to the respective year.

In our dataset, we also included the geographic information about the area and the crop field labels. *RoI_and_labels* folder has three sub-folders: *RoI* sub-folder includes files to locate the region of interest (*RoI*) as rectangle polygon, while *WGS84* and *UTM* sub-folders include polygons for crop field labels in geographic (WGS84) and projected coordinate system (UTM). They are further separated based on the year.

File names inside *RoI* and *WGS84* sub-folders which include "wgs84" represent the RoI and label polygons in the geographic coordinate reference system, specifically, World Geodetic System 1984 which is also known as EPSG:4326. This system represents Earth as a three-dimensional ellipsoid. File names that include "utm" in the RoI and UTM sub-folders represent the RoI and label polygons in a projected coordinate system, specifically the Universal Transverse Mercator (UTM), which provides a flat, two-dimensional representation of the Earth. We included both coordinate systems for convenience: WGS84 is more commonly used with Google Earth, Copernicus Browser, and similar applications, while the dataset itself is in the UTM coordinate system for more precise analysis. The primary files have the ".shp" suffix, which stores geospatial vector data. This type of file, known as a shapefile, contains the X and Y coordinates for the RoI and label polygons. Additional files with the suffixes ".shx", ".dbf", and ".prj" accompany the shapefile. The ".shx" file serves as an index for the shapefile, while the ".prj" file contains coordinate system and projection information for the RoI and labels. The ".dbf" file is a dBase database file that complements the shapefile by storing additional attribute data. For the RoI, the dBase file contains only the "NAME" field, while for labels, it includes the following fields: *crop_name, apia_code, label, rgb_r, rgb_g, rgb_b*, and *hex_color*. The *label* field holds non-zero positive integer values, *rgb_r, rgb_g*, and *rgb_b* store color values, and *hex_color* provides the HEX code equivalents of the RGB values. This color information can

be used to overlay crop field labels with specific colors for better visualization and inspection, consistent with the details provided in Table 2.

## 4. Technical validation

In the context of smart agriculture, where monitoring and automation are increasingly driven by machine learning models, the integrity and accuracy of data are critical. Machine learning algorithms rely heavily on large-size, high-quality data to make reliable predictions, optimize processes, and support decision-making. Any errors in labeling or inconsistencies in the dataset can lead to flawed models, deceiving accuracy, and sub-optimal outcomes, which is especially detrimental in agriculture, where precision is key to resource management, crop monitoring, and sustainability.

In Romania, APIA (https://apia.org.ro/) is responsible for the administration of agricultural subsidies and direct payments to Romanian farmers, thereby playing a pivotal role in the implementation of the European Union's Common Agricultural Policy (CAP) (ECA - European Court of Auditors, 2020), that is implemented with various tools, including the Single Area Payment Scheme (SAPS). This provides direct payments based on land size, along with other initiatives designed to advance environmental protection, rural development, and the modernization of agricultural practices. The agency has to ensure rigorous quality control measures that guarantee the effective and transparent utilization of funds.

The information provided by APIA ensures that our dataset is meticulously curated and free from labeling errors. As a national authority responsible for managing agricultural subsidies and interventions, APIA plays a crucial role in confirming and verifying the accuracy of information, ensuring that data used in our processes adheres to strict quality standards. The partnership with the Potato Institute has allowed us to create a clean, error-free dataset that meets the highest standards of data integrity. Through close collaboration with the Potato Institute, the dataset provided undergoes continuous validation and refinement, ensuring that the data used in the models is accurate, up-to-date, and highly relevant.

Furthermore, considering data collected over a period of 5 years, with measurements taken throughout the entire crop development cycle, from sprouting to harvest, allows for addressing significant issues such as the spectral overlap between different crops and the heterogeneity within a single crop (Rana et al., 2025), thus enhancing the generality of features learned by AI models. For the image patches, the applied filtering to produce clean data ensures that only the patches containing the agricultural crop, specifically, those in which the crop has sprouted but has not yet been harvested, are included.

## 5. Data set value

The integration of new datasets into precision agriculture offers valuable opportunities to enhance the monitoring, analysis, and optimization of farming operations. In this section, we present a series of practical use cases that demonstrate the applicability of the dataset within the context of precision agriculture. These use cases highlight scenarios where data-driven decision-making can lead to improved crop management, resource efficiency, and environmental sustainability.

By exploring these use cases, we aim to showcase the versatility and value of the dataset, providing insights into how it can address real-world challenges. Each use case is discussed in detail, outlining the experimental setups, methodologies employed, and the results obtained. Through these examples, we aim to emphasize the practical impact and innovative possibilities enabled by our dataset.

### 5.1. Crop identification: past vs present (problem 1)

For the first problem, we consider as being the "present" moment, "*the beginning of 2024*", and therefore all data before (2020, 2021, 2022, 2023) is viewed as available for training. All the data from 2024 is in the "present", thus not available for training the model and is consequently placed in the testing set. In this way, we can study how relevant is the data that *was* acquired in the past, to identify the data that *will be* presently acquired.

To facilitate the AI-based identification of agricultural crops, we used the patches corresponding to the known agricultural crops within the DACIA5 dataset, generated as described in Section 3. The multispectral patches of $32 \times 32 \times 12$ pixels from Sentinel-2, as well as the Sentinel-1 patches of $32 \times 32 \times 2$ were selected from the Potato Institute parcels. The resolution $32 \times 32$ has been popularized by the CIFAR 10 dataset (Krizhevsky & Hinton, 2009), which is one of the most used benchmarks in image classification (Brigato et al., 2022) (this report shows that it is by far most used, but data is limited to 2022).

All patches from the years 2020 to 2023 from Sentinel-2 result in a training set of 5431 images, each with a size of $32 \times 32 \times 12$. The test set, based on patches from 2024, comprises 1023 images of the same dimensions. For Sentinel-1, the number of train patches from the years 2020 to 2023 is 5103, from timestamps slightly different than those of the Sentinel-2 patches, while the number of samples in the test set from year 2024 is 892. As previously explained, a matching between the timestamps of the Sentinel-2 to the closest ones from Sentinel-1 data was performed, to allow the fusion of both types of satellite data.

Although the dataset comprises 17 types of crops (in total), some of the parcels were too small to provide patches of the required size, and as a result, these crops are not included. This leaves us with 12 crops, corresponding to 12 classes for our models. The patches were carefully filtered to ensure they predominantly contained pixels representing the target crop and were captured during periods of the year when the crop had sprouted and was clearly visible on the respective parcel. It is important to note that, while the dataset encompasses 12 distinct crop classes, only 8 of these classes were represented in the test set, reflecting variations in crop availability during the selected time period. This refined dataset enables precise and consistent input for our models, ensuring that they receive high-quality, relevant data for crop identification and classification tasks.

Two types of classification experiments were conducted. The first experiment utilized only the Sentinel-2 patches, while the second experiment employed fused data, where each $32 \times 32 \times 14$ image contains Sentinel-1 information in the first two channels and Sentinel-2 data from the corresponding patch in the remaining 12 channels.

### 5.1.1. Deep learning approach: identification using ResNet18

One of the most prevalent challenges in the literature is the classification and identification of agricultural crops and our dataset can be seen as addressing this issue. Initial experiments that will form a baseline indicating the problem difficulty and exploring this application were conducted using ResNet18 (He et al., 2016). ResNet18 was selected due to its ability to handle complex and diverse visual patterns efficiently. The architecture uses residual connections to mitigate the challenges of the vanishing gradient problem, making it particularly suited for datasets with nuanced agricultural features.

**5.1.1.1. Model description and training procedure.** For crop identification, we started a pretrained ResNet18 model provided by the PyTorch library. This architecture is well suited for handling image classification tasks involving small-sized inputs, such as the patches in our dataset. Since the Sentinel-2 patches were derived from multispectral images with 12 spectral bands, we modified the input layer of the original ResNet18 model to accommodate the new input dimensions; this was done by replicating the green plane initial weights. Additionally, the network's final layer was replaced with a fully connected layer containing 12 neurons, each corresponding to one of the crop classes. For the experiment with the fused data, the number of input channels was set to 14.

For optimization, we employed the AdamW optimizer (Loshchilov & Hutter, 2019), a variation of the Adam optimizer that effectively handles weight decay, leading to improved generalization and reduced overfitting. The loss function used was the classical cross-entropy loss, as implemented in the PyTorch library.

During training, the model processed the training set in batches of 128 images. However, the dataset was inherently imbalanced, as some crop classes were significantly more represented than others. Such class imbalances can bias the model towards the overrepresented classes, reducing its ability to accurately classify the underrepresented ones. Addressing this issue was critical to ensure the model achieved balanced performance across all crop classes.

To mitigate the effects of class imbalance, we employed the *Imbalanced Data Sampler* from the *torch.utils.data* module in PyTorch, see Imbalanced Dataset Sampler (2022). This sampling strategy dynamically adjusts the composition of batches, ensuring that underrepresented classes are sampled more frequently during training. By doing so, the sampler effectively balances the contribution of each class to the training process, helping the model to learn features from all classes more equitably. This approach is particularly beneficial for datasets where the minority classes are crucial yet insufficiently represented, as it avoids the need for data duplication or excessive augmentation, which could otherwise lead to overfitting. For further details on such techniques, see Buda et al. (2018), which discusses strategies for dealing with imbalanced datasets in deep learning.

The learning rate was set to a standard value of $10^{-3}$, which is commonly used for deep learning models in similar tasks. To further prevent overfitting, a weight decay of $10^{-3}$ was applied. This regularization technique penalized excessively large weights, encouraging the model to rely on more robust features and improving its generalization capability.

**5.1.1.2. Results and discussion.** For both types of experiments, we ran the training process several times. For the Sentinel-2 set, the accuracy achieved on the test set varied between 60% and 65%, with an average of 62.67%, while for the training set, we

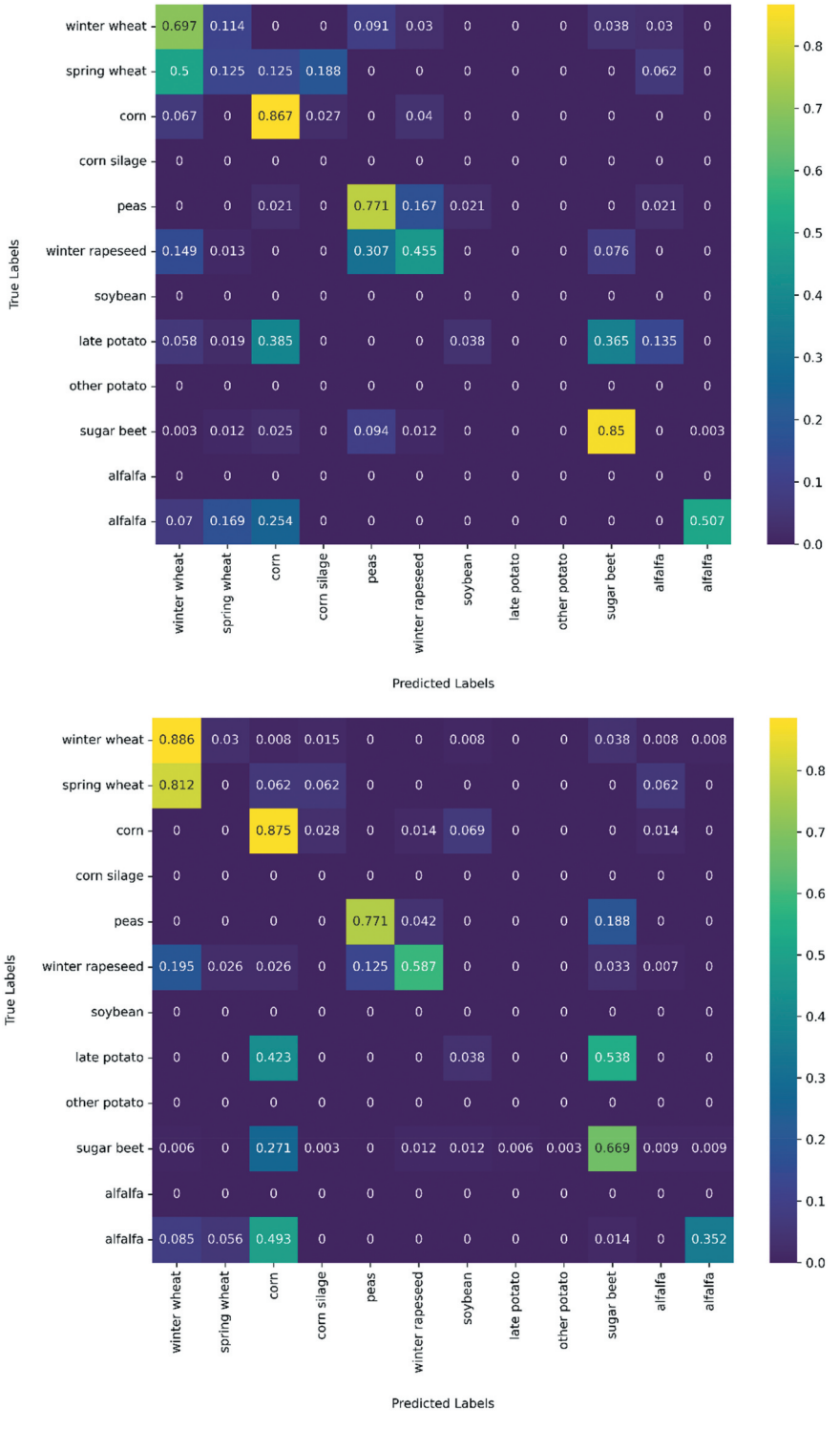



**Figure 10.** The confusion matrices obtained for two cases are detailed. For the Sentinel-2, the overall accuracy was 62.13% (top), while for the fused Sentinel-1 and Sentinel-2, it was 62.56% (bottom).

consistently obtained accuracies above 90%. To illustrate the model's performance, we selected one example for detailed analysis. For this example, the model achieved an accuracy of 91.72% on the training set and 62.13% on the test set.

For the experiment using the fused data, the best accuracy on the test set was around 55%, obtaining for the train set also a large accuracy above 91%. In this experiment, the training was run several times. The performance in this case is illustrated by one example, in which the accuracy on the train set was 97.33% and, on the test set 62.56%.

The confusion matrices for these two cases are presented in Figure 10. Each cell *(i, j)* of the matrix represents the fraction of images from class *i* that were classified as class *j*. This allows for a detailed evaluation of the model's performance across the different crop classes, revealing potential areas for improvement.

The results from the two datasets – one using Sentinel-2 data and the other combining Sentinel-1 and Sentinel-2 – are similar. The confusion matrices for both tests show nearly identical distributions of true positives, false positives, true negatives, and false negatives. Additionally, the accuracy values are almost the same, indicating that the model performs consistently across different data sources. While the overall accuracy is moderate, the consistency of the results suggests that the model's performance is not significantly influenced by the choice of the dataset. To enhance performance, further model adjustments or data preprocessing may be necessary.

One can notice that several agricultural crops are better identified, such as corn and peas, with true positive rates ≥0.75. Other crops, like wheat and sugar beets, are acceptably-well identified, with rates between 0.5 and 0.7. While for the rest of the crops, the identification exhibits very low accuracy rates.

However, it is worth mentioning that some crops are better classified by adding the Sentinel-1 bands. For example, wheat, potato, and rapeseed present a performance gain, whereas sugar beet and alfalfa show a lower performance. Radar (Sentinel-1) is sensitive to crop structure, and water content, which helps distinguish crops with unique structural characteristics. Optical (Sentinel-2) focuses on leaf reflectance and chlorophyll content. Adding the two bands can make a difference for plants with a more different structure. If the crops are sufficiently different at the spectral level, but similar at the structural level, the use of radar information can downgrade the results.

The classification results can be influenced by several factors, including hyperparameters (such as *learning rate*, *batch size*, *weight decay*, etc.), but also by the inherent imbalance in the training set, where certain classes are underrepresented. Another challenge arises from the presence of similar-looking crop types, which can make accurate classification more difficult.

To address these issues, one potential solution could be to cluster similar crops together, which might help the model focus on distinguishing between the most distinct features of each class. This approach could potentially reduce classification errors and improve the model's overall performance by mitigating the effects of both class imbalance and visual similarity between certain crop types.

There is potential for further improving the performance of the trained model, as well as exploring the use of other models for classification.

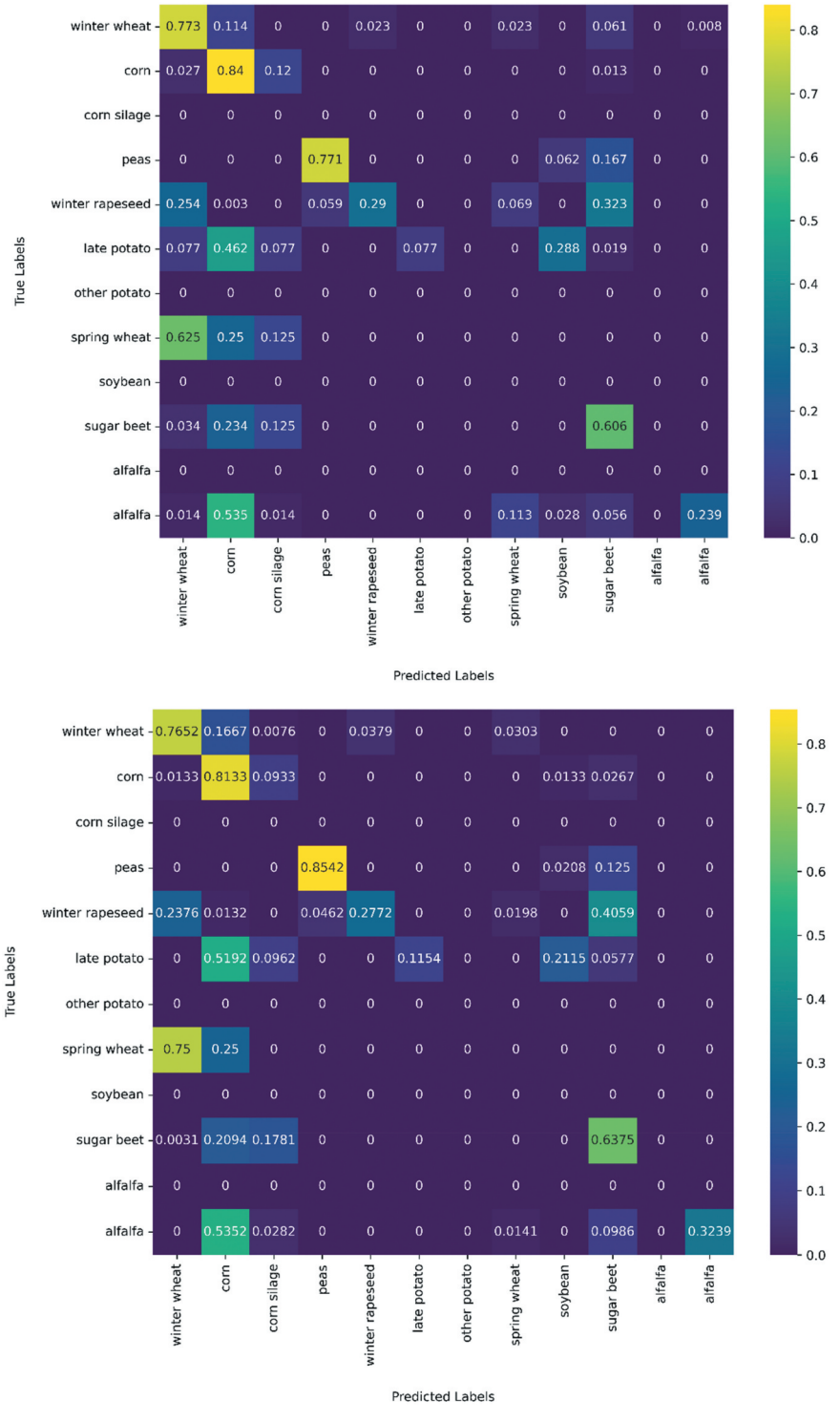



**Figure 11.** The confusion matrices obtained for the two experiments are detailed. For Sentinel-2 (top), the overall accuracy was 49.6%, while for the fused Sentinel-1 and Sentinel-2 (bottom), it was 54.09%.

### 5.1.2. Classical approach: identification with Random Forests

Random Forests (RF) (Breiman, 2001) are widely used for classification tasks due to their robust and versatile nature. They consist of an ensemble of decision trees, each trained on a random subset of the data and features. To prevent overfitting and improve generalization, the trees are regularized with injection of randomness, which came two-fold: trees are trained within bootstrapping framework (i.e. on randomly selected subsets) and each node selects a random subset of the input dimension for the split. As a result, Random Forests strike a good balance between bias and variance and can model non-linear relationships in the data without requiring prior assumptions about the data's distribution.

For a classification problem like the one here, the final result is determined by the majority (plurality) vote of the ensemble trees. Considering all these advantages, we conducted a classification experiment for our dataset, training a Random Forest to classify crops.

***5.1.2.1. Model description and training procedure.*** To train the model, we used in the first experiment all Sentinel-2 patches (i.e. 5431) having the full size $32 \times 32 \times 12$. In the second experiment, the fused Sentinel-1 and Sentinel-2 patches (i.e. 5103) of size $32 \times 32 \times 14$ were considered. The test data consisted of all the respective patches from 2024. The input to the model consists of multispectral pixels, with each pixel assigned a label corresponding to the patch from which it was derived. The selected hyperparameters were as follows: 40 decision trees and an InBagFraction of 0.8. The model produced 12 output classes, each representing a crop type present in the training set.

***5.1.2.2. Results and discussion.*** The confusion matrix constructed from the results obtained on the Sentinel-2 test set is shown in Figure 11 (top). Overall accuracy is 49.6%. With respect to individual crops, the best obtained classification score is 84% for corn (code 108), followed by the result for wheat at 77.3%, peas at 77.1% and sugar beet at 60.6%.

The results obtained for the second experiment with the fused test (Sentinel-1 + Sentinel-2) set are presented in Figure 11 (bottom). Overall accuracy is 54.09%. In comparison with the first experiment, we can observe that some classes are better classified, like peas, with an accuracy of 85.42%, while for other classes, like winter wheat (76.52%), the results are slightly worse.

Comparing the results obtained by RF (pixel-based) with those achieved by ResNet18 (patch-based) reveals certain similarities in classification, but at the same time certain discrepancies. Some crops are well identified by the deep model and poorly by the RF and the other way around. Other crops, such as winter wheat, are well separated by both approaches.

The two identification solutions presented here are aimed at offering baseline perspectives for future tests. This use case demonstrates the potential of our dataset in training a CNN model to identify different types of agricultural crops based on past data, a task that is critical for applications in precision agriculture, crop monitoring, and resource management.

### 5.2. Early crop identification (problem 2)

A second experiment focuses on the early identification of agricultural crops. To achieve this, we selected patches of interest from the refined dataset. We chose 20 May as the date for splitting the dataset into train and test. The date is motivated by the fact that at that moment the sowing process ended and the accumulation of data with respect to areas cultivated by various crops can be done. For Romania, it is also the date when APIA asks for self-reports about crops cultivated and it will start the verification process.

Given this date, we explored two working scenarios. We selected $32 \times 32 \times 12$ patches from the period 2020–2024 up to 20 May only for the crops that had already sprouted, resulting in 1176 patches. Based on the crops present before 20 May, we selected patches from after 20 May in the same period, yielding 1073 patches. Six crops were selected that appeared in both the training and test sets.

In the first working scenario, the model was trained on the patches from before 20 May and tested on those from after 20 May. That would correspond to the standard problem of after harvest identification: recognize older crops from young ones.

In the second scenario, the training and test sets from the first scenario were swapped: the training is done on mature plants, and the testing is on young plants. As mentioned in a previous section, experiments were also performed on the merged data between Sentinel-1 and Sentinel-2, with 1162 patches before 20 May and 1025 after this date. This is motivated from a machine learning perspective to study the dataset compactness and is the standard early crop identification. For the implementation of both scenarios, we used the same model to ensure a fair comparison between the two approaches.

#### 5.2.1. Deep learning approach with ResNet18

**5.2.1.1. Model description and training procedure.** The ResNet18 model, provided by the PyTorch library preparation, is similar to the Past vs Present crop identification: the input layer of the pretrained is changed to accommodate the input dimensions of $32 \times 32 \times 12$ of Sentinel-2 patches. For the experiment with the Sentinel-1 and Sentinel-2 data, the input dimensions were set to $32 \times 32 \times 14$. The network's final layer was replaced with a fully

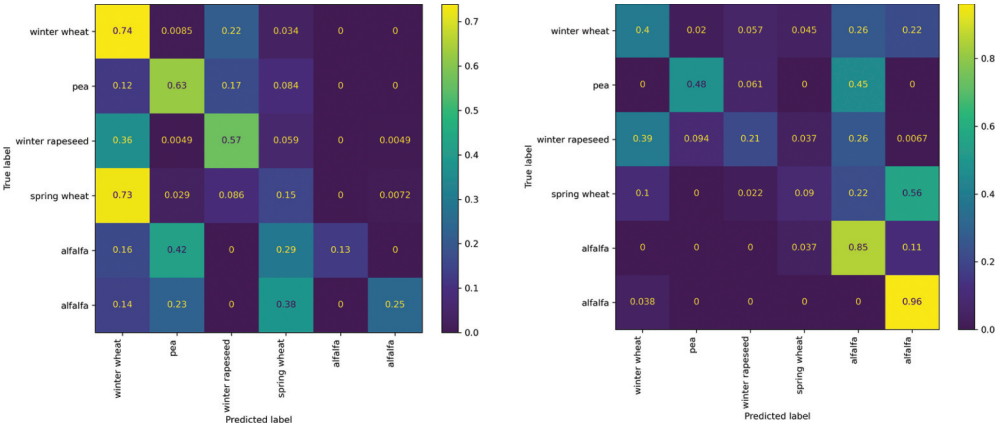
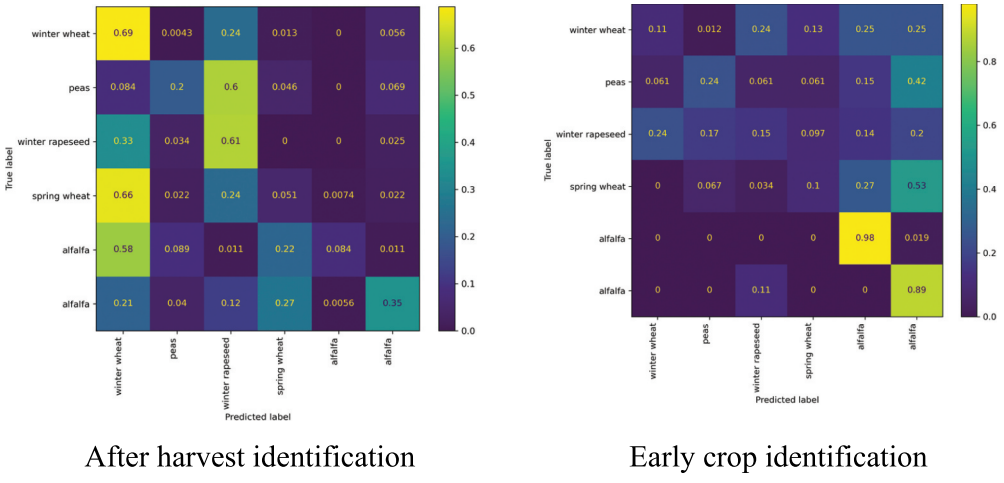



**Figure 12.** The confusion matrices obtained for the two scenarios, considering only Sentinel-2 data.




After harvest identification Early crop identification

**Figure 13.** The confusion matrices obtained for the two scenarios with fused data.

connected layer containing six neurons, each corresponding to one of the crop classes. The optimizer used was Adam, and the loss function employed was cross-entropy loss.

During training, the model processed the training set in batches of four images. The dataset was quite unbalanced, some crops were better represented than others. This aspect influenced the model to learn certain agricultural crops better than others, with those that were in a higher proportion achieving better results in identification on the test set.

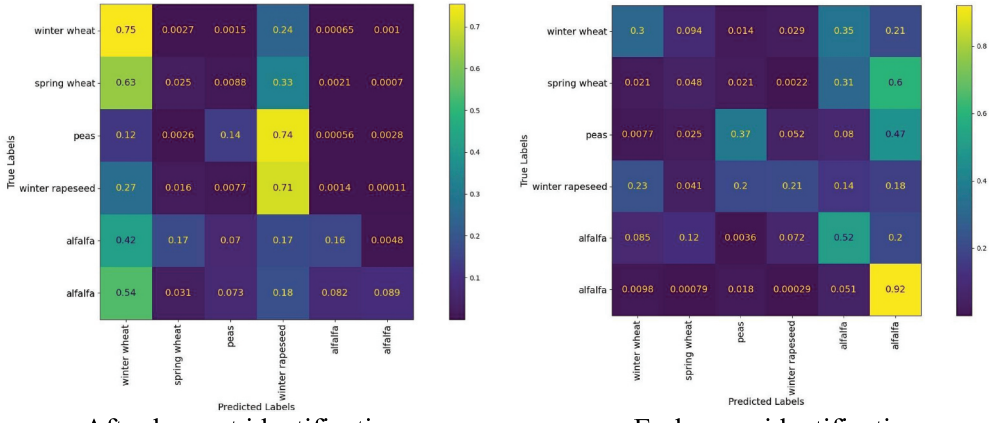
The model was trained for 40 epochs. The learning rate was set to a standard value of $3 \times 10^{-4}$. To further prevent overfitting, a weight decay of $3 \times 10^{-4}$ was applied. The smaller learning rate was chosen in parallel with the smaller batch and to ensure enough iterations as to allow the model to be able to learn.

***5.2.1.2. Results and discussion.*** For the experiment using Sentinel-2 data, in the first working scenario (mature crop identification), the overall accuracy on the test set was 41.01%, while on the training set, it was 91.79%. In the second working scenario (early crop identification), the overall accuracy on the test set was 49.96%, while on the training set, it was 94.25%. The confusion matrices of the two working scenarios can be seen in Figure 12.

For the experiment using the fused data, in the mature crop identification, the overall accuracy on the test set was 33.01%, while on the training set, it was 90.31%. For early crop identification, the overall accuracy on the test set was 41.16%, while on the training set, it was 95.17%. The confusion matrices of the two working scenarios with fused data can be seen in Figure 13.

Through the experiments and scenarios addressed, we found that the model is influenced by the unbalanced amount of data of each agricultural crop, for example in the first scenario we have a higher proportion of wheat crops and in the second scenario the proportion of alfalfa is higher. These aspects are also observed in the confusion matrices, by high accuracy of wheat and of alfalfa, respectively.

When we use Sentinel-2 data, in the after-harvest identification, we observe good identification of wheat, winter rapeseed, and peas. In the early crop identification, the resulting accuracies are different: the model identifies well the alfalfa and peas. Peas are



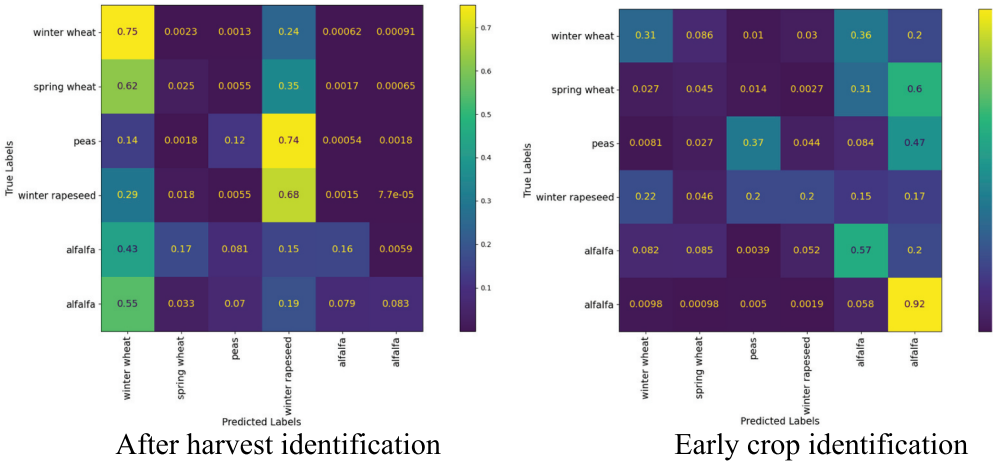

After harvest identification Early crop identification

**Figure 14.** The confusion matrices were obtained with random forest for the two scenarios.

identified quite well in both scenarios. A major impact on the early identification of crops is also the choice of the date we use to split the dataset; this has to take into account the growing season of the agricultural crops. The use of fused data does not greatly influence the obtained results, the confusion matrices do not change significantly by adding Sentinel-1 data. In the first work scenario wheat and rapeseed are well identified, and in the second work scenario alfalfa is best identified. A difference occurs in the peas crop identification, where a decrease is observed.

### *5.2.2. Early crop identification using Random Forest*

As in the Past vs Present crop identification problem, where ResNet18 and Random Forests were compared, we conducted a similar analysis in this second experiment focused on early crop identification, thus complementing the baseline result with Random Forests.




After harvest identification Early crop identification

**Figure 15.** The confusion matrices were obtained with random forest for the two scenarios with fused data.

***5.2.2.1. Model description and data.*** The considered dataset was identical to that presented in the previous section, while the Random Forest ensemble is the same as in the previous problem.

***5.2.2.2. Results and discussion.*** In the mature crop identification, where the training set contains images before May 20, and the test set contains images after this date, the total accuracy was 30.64%. Among the very well-identified crops are winter wheat and rapeseed with a percentage of 70–75%. The highest confusions are between peas and rapeseed and the two types of wheat. The confusion matrix for this scenario is represented in Figure 14, left subfigure.

In the second scenario, where the training set contains images after May 20, and the test set contains images before this date, the total accuracy was 35.65%. In this case, the crops correctly identified are two types of alfalfa. The confusions are also present for most of the crops with alfalfa classes. The confusion matrix for the second case can be seen in Figure 14, right subfigure.

As mentioned in Section 5.2.1, a major impact on the results is the uneven distribution of classes. Also, the division of the data set containing crops in different vegetation phases can be an explanation. For example, in the first case (Figure 14 -left) many confusions are made with winter wheat because it is the crop with a more advanced vegetation phase in spring. The test images, being selected from 20 May onwards, contain crops with more vegetation, which is why they can be confused with the most developed crop that was present in the training set.

For the two fused data experiments, the results are quite similar in terms of overall performance and confusion matrices, as illustrated in Figure 15. The accuracies are 30.4% for after harvest identification and 35.28% for early identification, respectively. It can be observed that in the case of early identification there is an increase for the alfalfa crop, but otherwise the results are similar. The results for Random Forest are more stable (there are no major differences between the Sentinel-2 data only and the merged data) because it treats each band separately, avoids overfitting, and is less sensitive to texture differences. CNNs have high capacity and might overfit to radar features for some crops while failing to generalize well for others.

***5.2.2.3. Comparison between deep learning and Random Forest approaches.*** When the two approaches are compared, as expected, the deep learning model presents better results. Having a convolutional layer and more parameters, the ResNet18 model recognizes better local structure of each crop (such as plant spatial organization, and texture) and is therefore able to outperform the Random Forest. The latter, due to its nature, does only local and linear approximations of the crop texture and structure.

## 6. Conclusions

In the context of smart agriculture, the proposed DACIA5 dataset can contribute to the development of AI-based tools for agricultural crop identification. The dataset presented in this paper is publicly available on Zenodo, to further help the community in developing algorithms and methods for crop identification and monitoring. The provided dataset comprises Sentinel-1 SAR data and Sentinel-2 multispectral images. This ensures a balanced representation of different observational modalities, making the dataset more versatile and

adaptable to various scenarios and use cases within precision agriculture. By integrating both radar (Sentinel-1) and optical (Sentinel-2) imagery, we aim to provide a comprehensive dataset that facilitates model generalization across diverse agricultural crops, improving its applicability in real-world settings. This complementarity is a key strength of our dataset, allowing it to be leveraged in different methodological approaches and machine learning frameworks. Furthermore, the dataset is highly reliable, with accurate labels verified in-situ by accredited institutions like the National Institute of Research and Development for Potato and Sugar Beet, Brașov, Romania. The data set is spanning over 5 years (2020–2024) and comprising 17 types of agricultural crops on 47 parcels. With its 172 Sentinel-2 multispectral images (each of them 800 × 450 pixels), 159 Sentinel-1 radar images, together with its 6,454 Sentinel-2 and its 5,995 Sentinel-1 rectangular patches (of size 32 × 32 pixels) the dataset comprises over 6 M pixels. Ground truth data is essential, particularly for machine learning approaches, as it helps reduce errors in crop identification applications. The 32 × 32-pixel size patches allow for addressing two problems (use cases). We introduce two use cases for our dataset: past versus present agricultural crop identification and late versus early crop identification using machine learning models. Our dataset can be further used for new models and results, thereby contributing to the global efforts towards improved resource and land management.

Additionally, the dataset can be integrated or concatenated with other datasets through recently developed spectral image data aggregation methods (Luca et al., 2025). Current research has demonstrated how combining datasets with interpolation techniques can enhance machine learning models' generalization capabilities across different geographic regions. Such methodologies allow researchers to effectively utilize geographically constrained datasets while preserving the benefits of highly accurate labeling.

However, there are several limitations to consider. Although the dataset includes a substantial amount of multispectral pixel data, its size is still relatively small for training deep neural networks. Cloudy satellite images were excluded during preprocessing, and future work could explore including such cases to improve model robustness. The five-year timespan, while valuable, may not fully capture longer-term climatic variations. Furthermore, the dataset lacks complementary contextual information such as weather conditions, disease occurrences, or agricultural interventions, which could impact plant development. While the data was collected from a limited geographic region and may reflect local-specific patterns, the data aggregation approaches mentioned above can help researchers address this spatial limitation by combining our highly accurate dataset with others that offer greater geographic diversity. By making this dataset publicly available, we hope to support the field of precision agriculture and the development of more accurate, efficient, and sustainable farming practices through improved crop identification and monitoring technologies.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Funded by the European Union. The AI4AGRI project entitled "Romanian Excellence Center on Artificial Intelligence on Earth Observation Data for Agriculture" received funding from the European Union's Horizon Europe research and innovation program under grant agreement

no. 101079136. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them; AI4AGRI project received funding from the European Union's Horizon Europe research and innovation programme [101079136].

## Notes on contributors

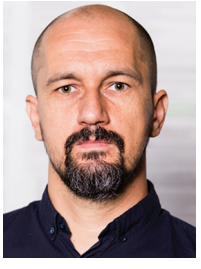
***Alexandra Baicoianu*** holds a PhD in Computer Science from Babeș-Bolyai University, Cluj-Napoca. She currently serves as an Associate Professor in the Department of Mathematics and Computer Science at Transilvania University of Brașov, where she teaches various courses and seminars on Algorithms, Formal Languages and Compilers, Data Mining, Data Warehousing, etc. With a strong academic background, she has published over 30 peer-reviewed scientific papers in prominent journals and conferences. She has also contributed to the academic literature as a co-author of six books, reflecting her commitment to advancing research and education in the field.

***Ioana Cristina Plajer*** received her PhD in Computer Science from Transilvania University of Brasov, Romania, in 2011 and is currently a Lecturer with the Faculty of Mathematics and Computer Sciences of the Transilvania University. She is also a member of the Artificial Intelligence and Earth Observation for Romania's agriculture (AI4AGRI) European project. Her research interests include machine learning, image processing, spectral imaging and remote sensing.

***Matei Debu*** received a bachelor's degree in computer science from the Transilvania University of Brașov in 2023 and is currently pursuing a master's degree in Modern Technologies in Software Systems Engineering at the same university. He is actively involved as a research assistant in several projects, including notable initiatives such as AI4AGRI, which focuses on the integration of artificial intelligence in agriculture, and other projects like AI4RiSK, IMINT and SEEN all of them focus on using artificial intelligence in different scenarios. His research interests span artificial intelligence, software systems design, and the application of modern technologies to solve complex, real-world problems.

***Floriana Maria Ștefan*** is the Head of Potato Genetic Breeding and Selection Lab at NIRDPSB Brasov and also serves as an Associate Professor at the Faculty of Food and Tourism of Transilvania University Brasov. As a horticultural engineer, her expertise lies in potato genotyping and phenotyping. Her research interests further extend to the development of temporal and spatial data.

***Mihai Ivanovici*** received his PhD from the Politehnica University of Bucharest, Romania, in 2006. He is currently a full professor at Transilvania University of Brasov, Romania. His research interests include signal and image processing and analysis, as well as remote sensing and Earth Observation data analysis.



***Corneliu Florea*** earned his master's degree from the National University of Science and Technology Politehnica of Bucharest, Romania, in 2004 and a PhD from the same university in 2009. After a stint with digital still camera software industry, he currently serves as a professor at the university's Image Processing and Analysis group. There, he teaches courses on machine learning, computer vision, and introductory statistical information processing. His research interests include statistical approaches to machine learning and computer vision, yielding over 80 peer-reviewed papers and of more than 25 U.S. patents.

***Angel Cațaron*** obtained his PhD in 2004 from the Politehnica University of Bucharest. He is currently affiliated with Transilvania University of Brașov. His research interests focus on data analysis, machine learning and their applications in addressing complex real-world challenges, with an emphasis on the practical use of predictive modeling and data-driven decision-making.

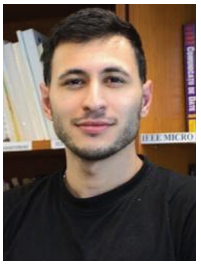
***Radu-Mihai Coliban*** is an Associate Professor in the Department of Electronics and Computers at Transilvania University of Brasov, Romania. His research interests include hyperspectral image processing, signal processing and digital electronics.

***Ștefan Popa*** received his PhD in Electronics, Telecommunications and Information Technology in 2021 from the Transilvania University of Brasov, Romania. Currently, a researcher in the university's Electronics and Computers Department, his work spans FPGA and ASIC design, data structures and algorithms, artificial intelligence, signal and image processing.

***Serban Oprișescu*** got his PhD in Electronics and Telecommunications from the Politehnica University of Bucharest, Romania, in 2007. He is currently a lecturer with Transilvania University of Brașov, Romania. He holds a B.S. degree in Electrical Engineering and Computer Science (2002) and an M.S. degree in Biomedical Engineering (2003) from the University Politehnica of București. His research focuses on image and video processing, ToF cameras, biomedical image processing and analysis and computer science. He has authored two ISI journal papers and more than 23 conference papers.

***Andrei Racovițeanu*** is a lecturer at the National University of Science and Technology Politehnica Bucharest. He is also a researcher within the Image Analysis and Processing Laboratory. His research interests include image processing and analysis, machine learning and remote sensing.

***Gheorghe Olteanu*** is an agronomy expert at Transilvania University of Brașov. His research focuses on in-situ measurements to monitor vegetation status, contributing to advancements in precision agriculture and sustainable land management.



***Kamal Marandskiy*** received a B.Sc. in Electronics, Telecommunication and Radio Engineering from the National Aviation Academy of Azerbaijan in 2020, and his M.Sc. in Advanced Electrical Systems from Transilvania University of Brasov in 2022. He is currently pursuing a PhD in Electronics and Telecommunications. His research focuses on machine learning techniques for Earth observation data analysis.

***Adrian Ghinea*** is a network engineer at NIRDPSB Brasov. He holds a master's degree in Computer Science, specializing in Algorithms and Software Products from the Faculty of Mathematics and Computer Science at Transilvania University of Brasov. His research focuses on remote sensing.

***Artur Kazak*** is a PhD student in Electronics, Telecommunications, and Information Technologies at Transilvania University of Brașov. His research focuses on AI-based analysis of Earth Observation data, aiming to advance techniques for environmental monitoring and decision-making.

***Luciana Majercsik*** received the B.S. degree in Mathematics from the University of Bucharest, Romania, and completed her PhD in Computer Science at the Transilvania University of Brasov, Romania. Her research interests include machine learning, multi- and hyperspectral image analysis and visualization, as well as graph-based methods in remote sensing.

***Adrian Constantin Manea*** earned a PhD in Computers and Information Technology from Transilvania University. His research focuses on the legal and security aspects of Earth Observation, exploring the intersection of technology, law, and data protection.

***Liviu Doru Dogar*** is a PhD student in Computer and Information Technologies at Transilvania University. His research focuses on surveillance camera systems and data acquisition techniques for monitoring applications.

## ORCID

A. Băicoianu http://orcid.org/0000-0002-1264-3404

## Data availability statement

Data is available on Zenodo at https://doi.org/10.5281/zenodo.14915950.

## References

Akkem, Y., Biswas, S. K., & Varanasi, A. (2023). Smart farming using artificial intelligence: A review. *Engineering Applications of Artificial Intelligence*, *120*, 105899. https://doi.org/10.1016/j.engappai.2023.105899

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-Dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., Duan, Y., Abdullah, A., Farhan, L., Lu, Y., Gupta, A., Albu, F., Abbosh, A., & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46. https://doi.org/10.1186/s40537-023-00727-2

Anderson, R., Bayer, P. E., & Edwards, D. (2020). Climate change and the need for agricultural adaptation. *Current Opinion in Plant Biology*, *56*, 197–202. https://doi.org/10.1016/j.pbi.2019.12.006

Aschbacher, J. (2017). ESA's earth observation strategy and Copernicus. In *Satellite earth observations and their impact on society and policy* (pp. 81–86). Springer Open. https://doi.org/10.1007/978-981-10-3713-9_5

Beck, H. E., McVicar, T. R., Vergopolan, N., Pehrsson, R., Huang, X., Nijssen, B., Zhang, Y., Ukkola, A. M., Wood, E. F., Schumacher, D. L., Weerasinghe, I., & Sheffield, J. (2023). High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data*, *10* (1), 724. https://doi.org/10.1038/s41597-023-02549-6

Benami, E., Jin, Z., Carter, M. R., Ghosh, A., Hijmans, R. J., Hobbs, A., Kenduiywo, B., & Lobell, D. B. (2021). Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nature Reviews Earth and Environment*, *2*(2), 140–159. https://doi.org/10.1038/s43017-020-00122-y

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(6), 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brigato, L., Barz, B., Iocchi, L., & Denzler, J. (2022). Image classification with small datasets: Overview and benchmark. *Institute of Electrical and Electronics Engineers Access*, *10*, 49233–49250. https://doi.org/10.1109/ACCESS.2022.3172939

Buda, M., Maki, A., & Maki, D. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Computers in Biology and Medicine*, *98*, 107–118. https://doi.org/10.1016/j.neunet.2018.07.011

Cabrera, G. F., Miller, C. J., & Schneider, J. (2014). Systematic labeling bias: De-biasing where everyone is wrong. *2014 22nd International Conference on Pattern Recognition* (pp. 4417–4422). IEEE. https://doi.org/10.1109/ICPR.2014.756

Copernicus Browser. (2024). Retrieved July 10, 2024, from https://dataspace.copernicus.eu/explore-data

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, *120*, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

ECA - European Court of Auditors. (2020). *Special report: Using new imaging technologies to monitor the common agricultural policy: Steady progress overall, but slower for climate and environment monitoring*, https://www.eca.europa.eu/en/publications/SR20_04

European Space Agency. (n.d.). Retrieved July 11, 2024, from https://www.esa.int/

Frona, D., Szenderák, J., & Harangi-Rákos, M. (2019). The challenge of feeding the world. *Sustainability*, *11*(20), 5816. https://doi.org/10.3390/su11205816

Gascon, F., Bouzinac, C., Thépaut, O., Jung, M., Francesconi, B., Louis, J., Lonjou, V., Lafrance, B., Massera, S., Gaudel-Vacaresse, A., Languille, F., Alhammoud, B., Viallefont, F., Pflug, B., Bieniarz, J., Clerc, S., Pessiot, L., Trémas, T. … Martimort, P. (2017). Copernicus Sentinel-2A calibration and products validation status. *Remote Sensing*, *9*(6), 584. https://doi.org/10.3390/rs9060584

Gascon, F., Cadau, E., Colin, O., Hoersch, B., Isola, C., Fernández, B. L., & Martimort, P. (2014). Copernicus Sentinel-2 mission: Products, algorithms and Cal/Val. In Prasad S. Thenkabail (ed), *Earth observing systems XIX* (Vol. 9218, pp. 455–463). SPIE. https://doi.org/10.1117/12.2062260

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(7), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

Imbalanced Dataset Sampler. (2022). *GitHub repository*. Retrieved July 11, 2024, from https://github.com/ufoym/imbalanced-dataset-sampler

INCDCSZ. (n.d.). National Institute of Research and*Development for potato and sugarbeet*. Retrieved July 11, 2024, from https://potato.ro/

Ivanovici, M., Baicoianu, A., Plajer, I. C., Debu, M., Ștefan, F.-M., Florea, C., Cațaron, A., Coliban, R.-M., Popa, S., Oprisescu, S., Racoviteanu, A., Olteanu, G., Marandskiy, K., Ghinea, A., Kazak, A., Majercsik, L., Manea, A., & Dogaru, L. (2024). *AI4AGRI Sentinel-2 Brasov area 2020-2024 multi-spectral dataset for crop monitoring and identification [data set]*. Zenodo. https://doi.org/10.5281/zenodo.14283243

Ivanovici, M., Olteanu, G., Florea, C., Coliban, R. M., Ștefan, F., & Marandskiy, K. (2024). Digital transformation in agriculture. https://doi.org/10.1007/978-3-031-63337-9_9

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Lee, J. S. (1980). Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-*PAMI-2*(2), 165–168. https://doi.org/10.1109/TPAMI.1980.4766994

Liedekerke, M. V., Jones, A., & Panagos, P. (2006). *ESDBv2 raster library - a set of rasters derived from the European soil database distribution v2.0*. European Commission and the European Soil Bureau Network.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations* (pp. (ICLR)). https://doi.org/10.48550/arXiv.1711.05101

Luca, R. I., Baicoianu, A., & Plajer, I. C. (2025). Spectral image data aggregation for multisource data augmentation. *European Journal of Remote Sensing*, *58*(1). https://doi.org/10.1080/22797254.2025.2492295

Maja, M. M., & Ayano, S. F. (2021). The impact of population growth on natural resources and farmers' capacity to adapt to climate change in low-income countries. *Earth Systems and Environment*, *5*(2), 271–283. https://doi.org/10.1007/s41748-021-00209-6

Marandskiy, K., & Ivanovici, M. (2024). Early identification of potato fields using data fusion and artificial neural network. *2024 International Symposium on Electronics and Telecommunications (ISETC)*, Timisoara, Romania (pp. 1–4). IEEE. https://doi.org/10.1109/ISETC63109.2024.10797316

Panagos, P. (2006). The European soil database. *GEO: Connexion*, *5*(7), 32–33.

Panagos, P., Van Liedekerke, M., Borrelli, P., Köninger, J., Ballabio, C., Orgiazzi, A., Lugato, E., Liakos, L., Hervas, J., Jones, A., & Montanarella, L. (2022). European soil data centre 2.0: Soil data and knowledge in support of the EU policies. *European Journal of Soil Science*, *73*(6), e13315. https://doi.org/10.1111/ejss.13315

Payments and Intervention Agency for Agriculture. (n.d.). Retrieved July 11, 2024, from https://apia.org.ro/

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, *11*(5), 1633–1644. https://doi.org/10.5194/hess-11-1633-2007

Pour, A. B., Parsa, M., & Eldosouky, A. M. (Eds.). (2023). *Geospatial analysis applied to mineral exploration: Remote sensing, GIS, geochemical, and geophysical applications to mineral resources*. Elsevier.

Rana, S., Gerbino, S., & Carillo, P. (2025). Study of spectral overlap and heterogeneity in agriculture based on soft classification techniques. *MethodsX*, *14*, 103114. https://doi.org/10.1016/j.mex.2024.103114

Research and Development Institute of Transilvania University of Brașov. (2024). Retrieved July 11, 2024, from https://icdt.unitbv.ro/ro/contact.html

Ritchie, H., & Rodes-Guiaro, L. (2024). *Peak global population and other key findings from the 2024 UN world population prospects*. *Our world in data*. Retrieved October 18, 2024, from https://ourworldindata.org/un-population-2024-revision

Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X. (2019). SEN12MS – a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-2*(W7), 153–160. https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019

Sentinel-1 Toolbox. (n.d.). Retrieved February 12, 2025, from https://earth.esa.int/eogateway/tools/sentinel-1-toolbox

Sentinel-2 bands. (n.d.). Retrieved July 17, 2024, from https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/bands/

Sentinel-2 processing. (n.d.). Retrieved July 11, 2024, from https://sentinels.copernicus.eu/web/sentinel/sentinel-data-access/sentinel-products/sentinel-2-data-products/collection-1-level-2a

Sentinel Application Platform. (n.d.). Retrieved July 11, 2024, from https://step.esa.int/main/download/snap-download/

SNAP - Resampling methods. (n.d.). Retrieved February 12, 2024, from https://step.esa.int/main/wp-content/help/versions/11.0.0/snap/org.esa.snap.snap.help/general/overview/ResamplingMethods.html

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J. G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(11), 8135–8153. https://doi.org/10.1109/TNNLS.2022.3152527

Sumbul, G., Charfuelan, M., Demir, B., & Markl, V. (2019). BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS, 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 5901–5904). https://doi.org/10.1109/IGARSS.2019.8900532

Teaci, D. (1980). *Bonitarea terenurilor agricole (Bonitarea și caracterizarea tehnologică a terenurilor agricole)* (p. 296). Editura Ceres, București.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Navas Traver, I., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R. . . . Rostan, F. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *12*, 0034–4257. https://doi.org/10.1016/j.rse.2011.05.028

Tseng, G., Zvonkov, I., Nakalembe, C. L., & Kerner, H. (2021). CropHarvest: A global dataset for crop-type classification. In *NeurIPS, 2021 datasets and benchmarks track*. https://openreview.net/forum?id=JtjzUXPEaCu

U.S. Department of Agriculture. (2024). Retrieved July 12, 2024, from https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php

Weikmann, G., Paris, C., & Bruzzone, L. (2021). TimeSen2Crop: A million labeled samples dataset of Sentinel-2 image time series for crop-type classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 4699–4708. https://doi.org/10.1109/JSTARS.2021.3073965