Journal of Big Data

RESEARCH Open Access

Transfer learning approach based on satellite image time series for the crop classification problem

Ognjen Antonijević[1*], Slobodan Jelić[1], Branislav Bajat[1] and Milan Kilibarda[1]

*Correspondence: oantonijevic@grf.bg.ac.rs

[1] Department of Geodesy and Geoinformatics, Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia

Abstract

This paper presents a transfer learning approach to the crop classification problem based on time series of images from the Sentinel-2 dataset labeled for two regions: Brittany (France) and Vojvodina (Serbia). During preprocessing, cloudy images are removed from the input data, the time series are interpolated over the time dimension, and additional remote sensing indices are calculated. We chose TransformerEncoder as the base model for knowledge transfer from source to target domain with French and Serbian data, respectively. Even more, the accuracy of the base model with the preprocessing step is improved by 2% when trained and evaluated on the French dataset. The transfer learning approach with fine-tuning of the pre-trained weights on the French dataset outperformed all other methods in terms of overall accuracy 0.94 and mean class recall 0.907 on the Serbian dataset. Our partially fine-tuned model improved recall of crop types that were poorly classified by the base model. In the case of sugar beet, class recall is improved by 85.71%.

Keywords: Transfer learning, Remote sensing, Encoder–decoder architecture, Domain adaptation, Crop classification, Attention mechanism

Introduction

In recent decades, machine learning techniques (ML) have played an increasingly important role in solving problems in various areas of the geosciences. This is largely due to the rapid development of spatial information acquisition techniques (remote sensing, global navigation satellite systems, etc.) and the integration of mathematical algorithms into spatial mapping and analysis software [1]. One of the areas of geoscience that has successfully harnessed these innovations is the field of agriculture and crop production, which has led to the development of new disciplines such as digital agriculture (smart agriculture) and precision agriculture [2, 3]. These new disciplines are making agriculture more productive, more competent, and more environmentally friendly [2]. Considering all that the new spatial data collection technologies entail, the volume of data collected, the temporal and spatial resolution of data collection, the diversity of data sources, data quality, and reliability, this type of data could be classified as Big Data [4].

Motivation

However, regardless of the benefits that new spatial data collection technologies bring in terms of their volume and diversity, the major limitations for ML applications in solving geoscience problems are the dependence on extensive labeled data and the training costs associated with in situ measurements [5, 6]. One way to overcome this problem is to apply models developed for larger data sets to a specific task where we have a smaller number of instances. A central postulate in many ML and data mining algorithms is that the training and test data must be in the same feature space and with the same distribution. However, in many geoscience applications, this assumption does not hold [7].

This problem is related to the nature of geospatial data. In particular, their properties are strongly dependent on their location in geographic space. In both regression and classification tasks, unsatisfactory results are often obtained when the training data used to build the model is collected from one image or geographic region and then applied to another region [8]. Obtaining training data that match the feature space, data distribution, and predicted data distribution of the test data can be difficult and expensive [9]. Therefore, it is necessary to build a powerful learning model for a target domain trained using a related source domain. In this case, the *transfer learning* (TL) approach [7], in which a model developed for one learning task is adapted as a starting point for building a model for another learning task, provides some strategies to overcome this challenge.

Crop classification is a good example of a problem that, because of its scale and complexity, is solved almost exclusively through the use of (geospatial) remote sensing data. Accurate crop type maps are a valuable and often essential source of information for various applications, such as food security, crop rotation, crop yield prediction, and disaster preparedness [10, 11].

Previous work

TL techniques have been successfully used in many real-world applications, such as image processing, human activity classification, software fault classification, multilingual text classification, etc. [9]. Especially in recent years, TL has become an important and useful tool for geoscience applications. The fields of application are diverse: geological and geotechnical studies [12–14], forestry [15–17], soil science [18, 19], agronomy and agriculture [20–23].

The basic feature of all these studies is that freely available datasets, which can be referred to as Big Data, are used as covariates (predictors or features) in ML models. These data are mainly remote sensing data collected from satellite platforms as multispectral optical [24–26] or radar images [23] and even RGB drone images [22]. In some studies, the data sources are also combined with covariates derived from the digital terrain models (slope length factors, aspects, topographic wetness indices) [18].

Although most of these applications in agriculture can be categorized as classification tasks, the TL approach has also found its application in regression tasks. For example, Bursać et al. [27] proposed an instance-based TL method for estimating soil organic carbon (SOC) on croplands in different European countries by adopting SOC-related knowledge from the global source domain (LUCAS 2015 survey [28]). When it comes to the problem of crop classification, there is a long history of research on the use of remote sensing for this task. Early classifiers were developed based on single images and simpler mathematical models [29]. The emergence of satellite missions with finer spatial, temporal, and radiometric resolution (especially open data missions such as Landsat and Sentinel-2), classical ML methods began to be applied to crop classification based on image time series [30, 31]. The current research focus is shifting to state-of-the-art deep learning algorithms [32–34], which achieve high accuracy in solving the of crop classification problem.

These algorithms require large amounts of training data to achieve high performance. Some countries make their farmers' data available through public portals that contain information on field boundaries and crops. However, in most areas, crop type data are not available or are of limited quality and spatial and temporal coverage. Here, we focus on TL to improve crop classification performance in the target domain with less data.

Russwurm et al. [35] compared different crop classification models and found that the TransformerEncoder model had the highest overall accuracy in most cases. A self-attention model for classification of raw satellite time series was presented in study [36]. The model architecture was developed by Vaswani et al. [37], who proposed a Transformer model based on an encoder–decoder architecture with an attention layer that eliminates the need for recurrence in sequence-to-sequence modeling. This approach is widely used in various areas of sequence-to-sequence modeling: machine translation, natural language processing [38], etc. Since the crop classification problem based on satellite image time series is a *sequence-to-label* problem, the authors in study [35] discarded the decoder part of the architecture from study [37] and added a fully connected layer followed by Softmax as an adaptation to the classification model.

We refer to state-of-the-art TL models for crop classification and related tasks based on satellite image time series. Hao et al. [25] used a transfer learning approach with a random forest model on variable-length NDVI time series data with high confidence pixels sampled at 30-m and 15-day resolution from harmonized Landat-8 and Sentinel-2 data in the training dataset. The test datasets were selected for three different regions in China. Recently, Nowakovski et al. [22] used a transfer learning approach for crop classification based on high-resolution drone images. They applied the fine-tuning strategy of VGG16 and GoogLeNet CNN models to two datasets (Malawi and Mozambique) and reported high accuracy scores for the TL approach. A very interesting application of an unsupervised adaptive domain adversarial neural network for maize yield prediction was proposed by Ma et al. [24]. The authors used an adversarial network to reduce domain shift and learn domain invariant features. Another approach of unsupervised multi-source domain adaptation for crop mapping is proposed by Wang et al. [39], who based their model for classifying corn and soybean on the Multiple Feature Spaces Adaptation Network (MFSAN) architecture in combination with multiple layers of the (pre-trained) ResNet-50 model. The model was trained on the USA Cropland Data Layer (CDL) for 2018 and transferred to three provinces in Northeast China. Accuracy was assessed by comparing its outputs with previously published classification maps for the same region.

Contribution

Our contribution to crop classification based on satellite image time series consists of the following:

- a base model with modified self-attention architecture and preprocessing step with the improved overall accuracy on the BreizhCrops dataset [35],
- transfer learning approaches for the base model with improved overall accuracy compared to the base model for the Vojvodina dataset,
- publicly available code contributed to the project in study [35].

Paper organization

In the “Materials and methods" section, we present the data sets, the preprocessing strategy, the formulation of the transfer learning problem, and a description of the model architecture. The “Results” section provides an overview of the evaluation and an interpretation of the results in terms of evaluation metrics. Finally, the “Discussion" section summarizes the main results of the paper, while the “Conclusion” section suggests some possible directions for future work.

Materials and methods

This section is organized as follows. “Data” and “Preprocessing” sections describe the data set, covariates, target classes, interpolation and noise removal techniques etc. “Transfer learning approaches” section gives a self-contained presentation of the transfer learning and transfer approaches that we applied to the adopted self-attention model. Since “Scaled dot-product attention layer” and “Multi-head attention layer” sections are the crucial parts of the basic model used in the transfer learning process, they are described in separate subsections. We also dedicate a subsection to the architecture of the “TransformerEncoder model”. The last “Classification performance metrics” Subsection briefly describes the metrics used to evaluate model performance.

Data

We trained a machine learning model to identify which crop was grown in a given field using remote sensing data. Input features for the classifier were created from time series of Sentinel-2 images. Sentinel-2 [40] is a satellite mission launched by ESA as part of the Copernicus program. It consists of 2 twin satellites in the same orbit, but phased 180°, with a repetition frequency of 5 days or less. Each satellite carries a multispectral instrument (MSI) with 13 spectral channels (bands) in the visible/near-infrared (VNIR) and shortwave infrared (SWIR) part of the electromagnetic spectrum. MSI acquires images at 3 spatial resolutions: 10 m, 20 m, and 60 m. The radiometric and spectral resolutions of each band are listed in Table 1.

To train the base model (BMP), we used crop type data from the work of Russwurm et al. [35]. French farmers submitted data on plot geometries and crops grown as part of the subsidy application process. These data were anonymized and published by the RPG (Registre Parcellaire Graphique) with an open licensing policy. The authors used 2017 field data from the Brittany region—a total of **608,489** parcels divided into 9 crop categories: **barley**, **wheat**, **rapeseed**, **maize**, **sunflower**, **orchards**, **nuts**, **permanent meadows**, and **temporary meadows**. For each plot (parcel), the authors created a time series of Sentinel 2 images for 2017, which were then processed using the MAJA [41, 42] processing chain. MAJA is a cloud detection

Table 1 Sentinel-2 bands and their spatial and radiometric resolutions

Band	Resolution (m)	Central wavelength (nm)	Bandwidth (nm)	Description
B1	60	443	21	Ultra blue (coastal and aerosol)
B2	10	490	66	Blue
B3	10	560	36	Green
B4	10	665	31	Red
B5	20	705	15	Visible and near infrared (VNIR)
B6	20	740	15	Visible and near infrared (VNIR)
B7	20	783	20	Visible and near infrared (VNIR)
B8	10	842	106	Visible and near infrared (VNIR)
B8a	20	865	21	Visible and near infrared (VNIR)
B9	60	940	20	Short wave infrared (SWIR)
B10	60	1375	31	Short wave infrared (SWIR)
B11	20	1610	91	Short wave infrared (SWIR)
B12	20	2190	175	Short wave infrared (SWIR)

Table 2 Crop types and number of labels in two datasets

Crop	France	Serbia
Barley	36922	187
Wheat	89617	763
Rapeseed	14746	63
Maize	153995	2108
Soya	/	234
Sugar beet	/	14
Sunflower	19	1033
Orchards	3054	/
Nuts	49	/
Permanent meadows	127835	/
Temporary meadows	182252	/

and atmospheric correction tool for optical remote sensing imagery (from various satellite missions) specifically designed for use with multitemporal methods. The processed images have 10 spectral bands (the 60 m resolution bands were discarded during processing). For each imaging date, the reflections in each band were averaged at the field level. This resulted in 10 time series for each field, which were used as input for classification.

Plot labels for Serbia were collected during a field campaign in early summer 2022 at various locations in Vojvodina province. Most of the territory of Vojvodina consists of agricultural land (94.8%), of which 77.8% belongs to arable land and 6.8% to forest. Vojvodina province is an important agricultural area in Serbia for the cultivation of basic grain crops [43] (wheat, maize, sugar beet, sunflower, clover, etc.). Crop labels and parcel boundaries were recorded for 4409 fields. Table 2 contains a comparison of the two in situ datasets. The map with sampling locations and crop distribution between fields is shown in Fig. 1.

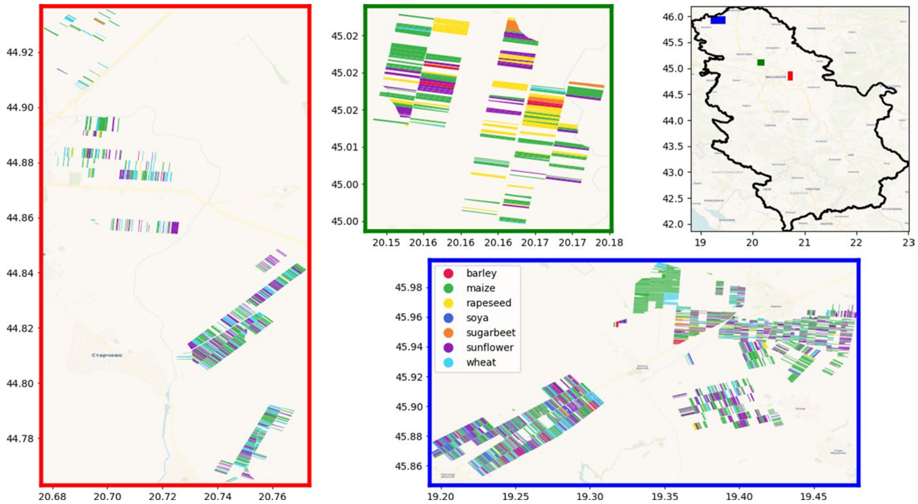


Fig. 1 Locations of crop labels in Vojvodina

Preprocessing

The length of the time series varies depending on the location of the parcel. Since most machine learning algorithms require fixed-length input data, the data had to be harmonized. Russwurm et al. [35] used sampling with replacement to harmonize the time series, with no filtering for cloudy data. We took a slightly different approach to preparing the data. First, the time series were filtered to remove cloudy dates, since reflectance values obtained from these images are blunders and do not contain useful information (they may even have a negative impact on model performance). After filtering, additional features were added to the data (besides the 10 Sentinel-2 spectral bands)—various spectral indices calculated from the original bands.

Finally, the time series of each parcel were harmonized with respect to the fixed time grid with a time step of 7 days starting on February 1 and ending on November 1. We chose these dates because they cover the growing and harvesting periods for most crop types. The linear interpolation method is used to impute missing data. Forward and backward padding is used in cases where data were missing at the beginning or end of the time grid. The preprocessing steps are shown in Fig. 2. The abrupt drops in NDVI values are due to clouds rather than to phonological development of the crop.

To justify the transfer learning approach, we confirmed the difference between the probability distributions of the inputs in the source and target domains using the Kolmogorov-Smirnov test. Figure 3 shows the *p*-value of the test and the time series of the reflectance values. For each of the basic input variables (Sentinel-2 reflectance bands), all training samples in each country were averaged and plotted. Both the *p*-values and the plots show that the distributions of the input data from the two data sets are different. Band 11 is the most similar between the two datasets, but the similarity is still very low.

Transfer learning approaches

Supervised learning is one of the most commonly used types of machine learning that requires a labeled dataset in both the learning and testing phases. Even when labels are available, the main assumption that both the feature space and the target

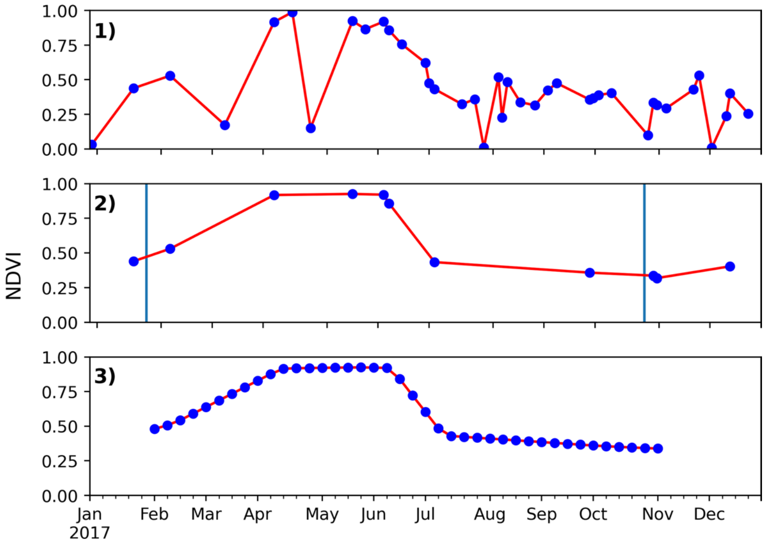


Fig. 2 Preprocessing workflow: (1) input time series (2) cloudy dates removed (middle) (3) final time series (cropped and resampled)

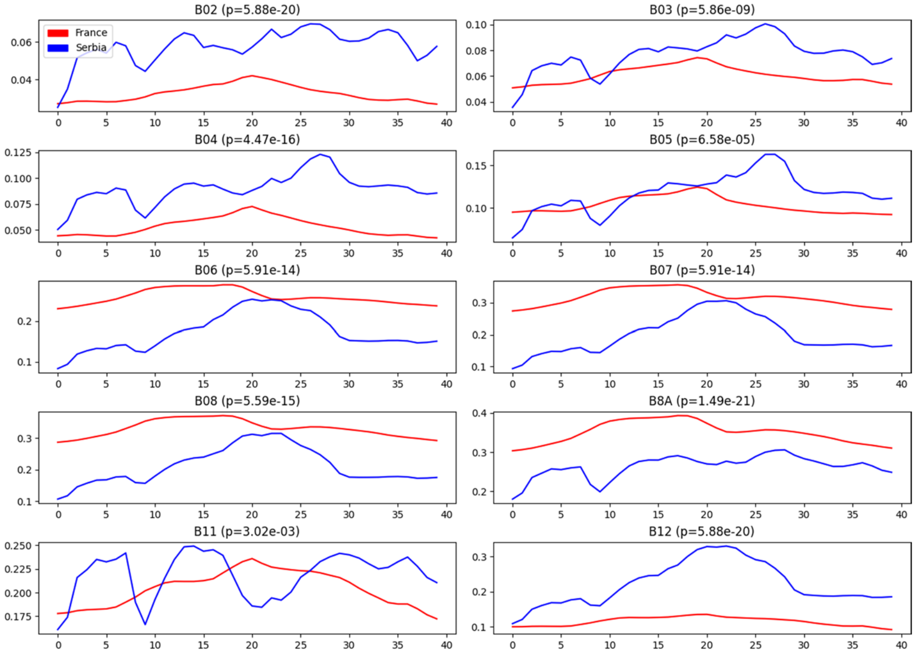


Fig. 3 Averaged reflectances of each band for two training datasets, with *p* values of Kolmogorov–Smirnov test

are the same and come from the same probability distribution usually fails. In many real-world applications such as machine translation, natural language processing, medical imaging, and remote sensing, the availability of labeled datasets is limited. This phenomenon has motivated the development of a very popular area of machine learning called *transfer learning*. We distinguish two main concepts: *domain* and *task*. In [7] we have a concise overview of transfer learning with a commonly accepted classification. *Domain* consists of the feature space $\mathcal{X}$ and the probability distribution $P(\mathbf{X})$, where $\mathbf{X}$ is a learning sample consisting of n instances, i.e., $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $n \in \mathbb{N}$, $\mathbf{x}_i \in \mathbb{R}^d$, $i \in \{1, \ldots, n\}$. On the other hand, *task* consists of a set of labels $\mathcal{Y}$ and an objective loss function f. In our use case, the domain consists of a feature space composed of d different spectral bands and vegetation indices [44] over satellite image time series datasets. The set of labels $\mathcal{Y}$ for the crop classification problem consists of different crop classes, while the objective is a cross-entropy loss function. Transfer learning is generally about transferring knowledge between different domains and tasks. For simplicity, we consider two domains: *source domain* $(\mathcal{X}_S, P(\mathbf{X}_S))$ and *target domain* $(\mathcal{X}_T, P(\mathbf{X}_T))$. The sets of labels are denoted by $\mathcal{Y}_S$ and $\mathcal{Y}_T$ in the source and target domains, respectively. In this paper, we focus on a particular type of transductive transfer learning, which we call **supervised domain adaptation**, where $\mathcal{X}_S \neq \mathcal{X}_T$ or $P(\mathbf{X}_S) \neq P(\mathbf{X}_T)$, while the task in the source and target domains is the same (crop classification), although the sets of labels $\mathcal{Y}_S$ and $\mathcal{Y}_T$ are different [45–47].

As $\mathbf{X}_S$, we use BreizhCrops, a publicly available, tagged, large-scale satellite image time series dataset extracted from Sentinel-2 for mapping harvests in the region of Brittany, France [35]. The target dataset $\mathbf{X}_T$ extracted from the time series Sentinel-2 for the Vojvodina region is also labeled. The set of labels $\mathcal{Y}_S$ in the source domain consists of 9 crop categories: barley, wheat, rapeseed, maize, sunflower, orchards, nuts, permanent grassland, and temporary grassland. The set of designations in the target domain $\mathcal{Y}_T$ consists of 7 crop categories: barley, wheat, rapeseed, maize, soya, sugarbeet, and sunflower. Our main goal in this work is to use the larger, publicly available dataset BreizhCrops to improve crop classification in a smaller dataset for Vojvodina, where the number of labeled instances is limited.

In our use case, the source and target domains have similar labels, while the probability distributions of the input data are different. Supervised domain matching is typically used when the target domain dataset is much smaller and more expensive to label, while the source domain dataset is larger and cheaper (i.e., publicly available, etc.).

We start with the model proposed by Russwurm et al. [35], based on the TransformerEncoder architecture (explained in detail in the next section), and refer to this model as the base model (**BM**). To improve the performance of the model, we perform several preprocessing steps on the input data. We refer to this model improved by preprocessing as **BMP**. In the first preprocessing step, cloudy timestamps are removed from the input time series, followed by interpolation across the time axis so that all instances contain the same number of observed signals at the same timestamps. Raw satellite spectral bands are used to compute task-specific vegetation indices, which are added as additional input features [44]. Unifying the time axis across the training samples, allowed us to remove the positional encoding layer from the originally proposed architecture in [36, 37] without compromising model accuracy. We compare the performance of the original model (BM) with our model (BMP).

In this paper, we also propose three transfer learning approaches to solve the supervised domain adaptation problem mentioned above, all based on the BMP model.

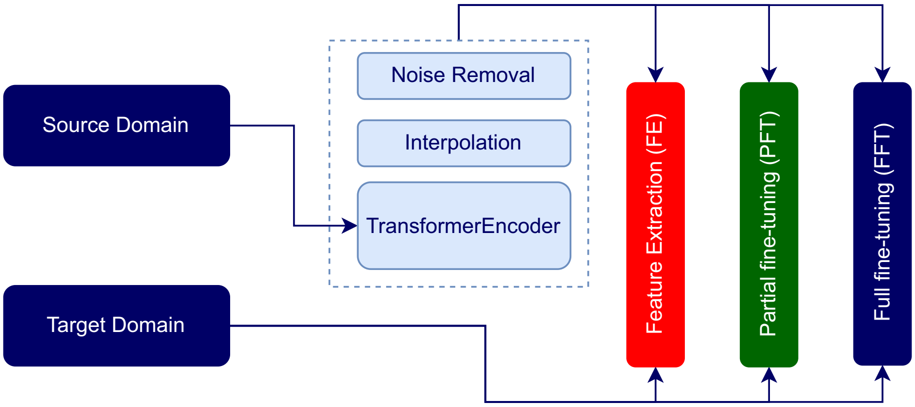


Fig. 4 Domain adaptation approach

Feature extraction approach (FE)

In this approach, all parameters of the model are set to the values of the corresponding parameters of the pre-trained BMP. In addition, all parameters are frozen except for the last linear output layer. In this way, we extract the learned features from the pre-trained model (without optimizing these parameters) and train only the linear regression parameters in the last output layer.

Partial fine-tuning (PFT)

While in the feature extraction approach we freeze all model parameters (except in the output layer), in fine-tuning only the parameters in the input layer are frozen while the other parameters are trained. Typically, the parameters in the input layer are trained to extract some general features, while the layers closer to the output extract task-specific features. We call this approach partial fine-tuning because we do not train all model parameters, but only the parameters that are expected to extract task-specific features.

Full fine-tuning (FFT)

In the third method, we unfreeze all layers and re-train all model parameters. This means that both the general feature extraction parameters and the task-specific feature parameters are re-trained.

Since the TransformerEncoder architecture is the basic building block of our transfer learning approaches, we give a self-contained description of this architecture at the technical level. For more details, the interested reader can refer to the seminal work of Vaswani et al. [37] and the application of this technique to the problem of crop classification in the paper of Russwurm et al. [36]. In BMP, our contribution to the model architecture is the incorporation of the preprocessing steps described earlier. This includes removing noise (e.g., clouds), enriching the input dataset with additional features, and interpolating missing values on the time axis. Due to the applied interpolation, our time series contain measurements at the same time points, eliminating the need for positional encoding and slightly reducing the complexity of the model. On the other hand, the proposed transfer learning approaches further improve the performance of the BMP on the Vojvodina data (Fig. 4). First, we give a mathematical description of the scaled attention layer, which is one of the attention mechanisms where the input vector is transformed by a linear operator represented by the attention matrix. Second, we give an overview of the multi-head attention layer. At the end, we summarize the overall architecture of TransformerEncoder.

Scaled dot-product attention layer

We use this type of attention mechanism as a cornerstone for other attention-based (sub) layers. The attention layer maps a query and a set of key-value pairs to an output [37]. The input of the layer is defined by a set of vectors $\mathbf{h}_t \in \mathbb{R}^{d_h}$, $t \in \{1, \ldots, T\}$, where T is the length of the time series of vectors of hidden dimension d_h. The hidden dimension d_h is given as the hyperparameter of the model. Since we have T input vectors (one for each timestamp), the input can be expressed as a matrix $\mathbf{H} \in \mathbb{R}^{T \times d_h}$ whose rows are vectors $\mathbf{h}_t^T$, $t \in \{1, \ldots, T\}$. The output vector $\mathbf{h}'_t \in \mathbb{R}^{d_h}$, $t \in \{1, \ldots, T\}$ are given as the weighted sum of the input vectors $\mathbf{h}_t \in \mathbb{R}^{d_h}$, $t \in \{1, \ldots, T\}$ considering the weights given in the vector $\mathbf{a}_t = [a_{t1} \ldots a_{tT}]^T \in \mathbb{R}_+^T$ as follows:

$$\mathbf{h}'_t = \sum_{j=1}^{T} a_{tj} \mathbf{h}_j, \tag{1}$$

where $\sum_{j=1}^{T} a_{tj} = 1$, for $t \in \{1, \ldots, T\}$. From (1) it follows that $\mathbf{h}'_t = \mathbf{H}^T \mathbf{a}_t$ or, equivalently, $\mathbf{h}'^T_t = \mathbf{a}_t^T \mathbf{H}$. If we represent T output vectors $\mathbf{h}'_t \in \mathbb{R}^{d_h}$ by the matrix $\mathbf{H}' \in \mathbb{R}^{T \times d_h}$, (1) can be expressed by the following matrix equation:

$$\mathbf{H}' = \mathbf{A}^T \mathbf{H}. \tag{2}$$

On the other hand, the attention matrix $\mathbf{A} \in \mathbb{R}^{T \times T}$ is defined by query and key matrices, i.e., $\mathbf{Q} \in \mathbb{R}^{d_h \times d_k}$ and $\mathbf{K} \in \mathbb{R}^{d_h \times d_k}$, $d_k \in \mathbb{N}$ (d_k is a hyperparameter of the model), as follows:

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_h}}\right), \tag{3}$$

where the matrix product $\mathbf{Q}\mathbf{K}^{\mathbf{T}}$ is divided by $\sqrt{d_h}$ to reduce scalar products, which tend to grow rapidly and bring gradients to extremely small values, making training difficult. The function Softmax is taken column-wise to normalize the attention weights.

If we replace $\mathbf{A}$ in (2) with (3), we get:

$$\mathbf{H}' = \left(\mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_h}}\right)\right)^T \mathbf{H}. \tag{4}$$

The equation (1) gives some insight into the attention mechanism. We see that $\mathbf{h}'_t$ is expressed as a linear combination of $\mathbf{h}_j$ with coefficients $a_{tj} \in (0, 1)$, $j \in \{1, \ldots, T\}$. Thus, the coefficient a_{tj} can be interpreted as the “fraction of attention” that the input vector $\mathbf{h}_j \in \mathbb{R}^{d_h}$ contributes to the output vector $\mathbf{h}'_t \in \mathbb{R}^{d_h}$.

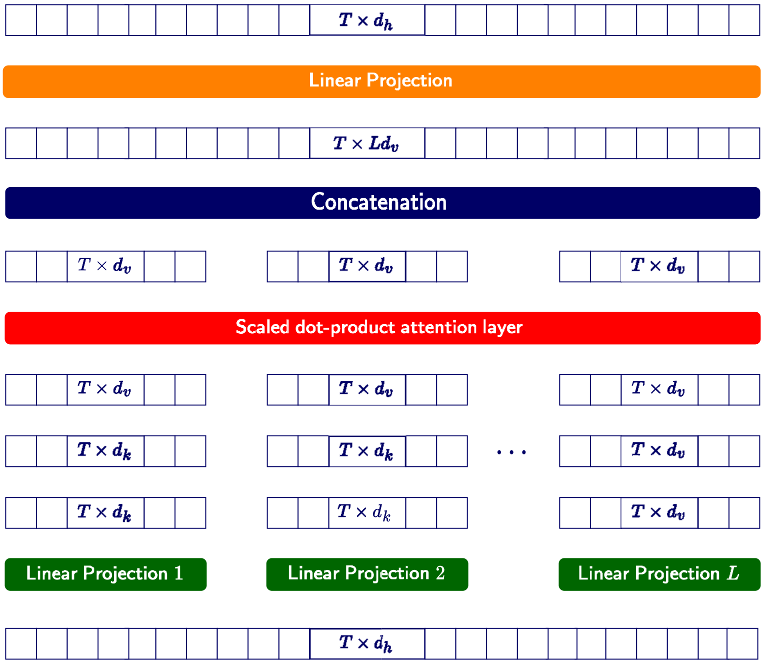


Fig. 5 Architecture of multi-head attention layer

Multi-head attention layer

This is the crucial part of the attention mechanism in the architecture of Transformer-Encoder. To avoid sequential execution, this layer performs in parallel the L transformations defined in (4) on the input $\mathbf{H} \in \mathbb{R}^{T \times d_h}$. Each computation has its own version of the scaled dot product attention, where the query and key matrices are computed as linear projections of the input matrix $\mathbf{H}$ through the matrices $\mathbf{Q}_i \in \mathbb{R}^{d_h \times d_k}$ and $\mathbf{K}_i \in \mathbb{R}^{d_h \times d_k}$, for $i \in \{1, \ldots, L\}$. The input matrix $\mathbf{H}$ is also projected by the matrix $\mathbf{V}_i \in \mathbb{R}^{d_h \times d_v}$, for $i \in \{1, \ldots, L\}$, where $d_v \in \mathbb{N}$ is also a hyperparameter of the model. After computing the L operations in (3), the resulting matrices are concatenated and projected by the matrix $\mathbf{W} \in \mathbb{R}^{Ld_v \times d_h}$ (Fig. 5).

This calculation can be expressed as follows:

$$\mathbf{H}' = \left(\mathrm{Conc}_{i=1}^{L} \ \mathrm{softmax} \left(\frac{\mathbf{HQ}_{\mathrm{i}}\mathbf{K}_{\mathrm{i}}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}}{\sqrt{\mathrm{d}_{\mathrm{h}}}} \right) \mathbf{HV}_{\mathrm{i}} \right) \mathbf{W},$$

where $\mathbf{Q}_i \in \mathbb{R}^{d_h \times d_k}$, $\mathbf{K}_i \in \mathbb{R}^{d_h \times d_k}$, $\mathbf{V}_i \in \mathbb{R}^{d_h \times d_v}$, for $i \in \{1, \ldots, L\}$ and $\mathbf{W} \in \mathbb{R}^{Ld_v \times d_h}$ are trainable matrices and $\mathbf{H}' \in \mathbb{R}^{T \times d_h}$ is the output matrix.

TransformerEncoder model

The model used in transfer learning follows an architecture described in [35, 36] with some differences. Since the input data is interpolated with respect to the time axis so that each data instance contains a fixed number of timestamps during the season, positional encoding layer is omitted here. The positional encoding layer provides information about the absolute and/or relative position of the data in the sequence, since the architecture of TransformerEncoder does not include convolutional or recurrent layers [37]. However, the timeline in the preprocessed dataset contains input features at predefined times, so there is no need to explicitly encode the position. Although input embedding is typically used in natural language processing and machine translation to embed discrete variables in a continuous space, we use a fully connected layer that embeds the input data $\mathbf{X} \in \mathbb{R}^{T \times d_i}$, where d_i is the input dimension, in the space of dimension d_h. After the fully connected layer, we pass the input sequence to the TransformerEncoderLayer, which consists of Multi-head attention layer and the fully connected layer. The multi-head attention layer is followed by a normalization layer that takes the original input and the input obtained after the multi-head attention layer, adds the two together, and normalizes the result. The normalized output is passed to the fully connected layer, which contains a hidden layer of dimension d_f. After the fully connected layer, the same addition and normalization steps are performed as after the multi-head attention problem. This layer is repeated N times.

After the TransformerEncoderLayer, we apply the maximum pooling layer to reduce the time axis. The fully connected layer is used to reduce the output of dimension d_h to the number of classes. In the end, the function Softmax maps the outputs to probabilities for each class (Fig. 6).

Classification performance metrics

We used the following metrics to evaluate the performance of the models: overall accuracy (OA), balanced accuracy (BA) or class-mean recall, weighted F-score (WF), macro (class-mean) F-score (MF) and Cohen's kappa (κ).

Overall accuracy is a basic performance measure that is not sensitive to low minority class performance in unbalanced data sets such as ours. More informative is balanced accuracy, which gives equal importance to all classes. For the same reasons, we included the macro F-score in addition to the weighted F-score, while using Cohen's kappa as the standard measure for classifier comparison.

Results

We implemented our models in Python 3.10.8 and Pytorch 1.13.1 using the publicly available code repository referenced in [35], while our code and list of additional computed input features are publicly available in the Github repository.[1] The strategy for loading the data into memory was changed when we preprocessed our data by unifying the number of timestamps. Since we pre-trained the model on the BreizhCrops dataset, we used the same architectural hyperparameters referenced in [35]. Although we made additional adjustments to our hyperparameters, there are no significant differences from the hyperparameters chosen in [35] for the TransformerEncoder model. The dimension of the hidden layer d_h is 64, while the dimension of the inner fully connected layer d_f is set to 128. The evaluation metrics for the test data are based on a stratified 4-fold cross-validation. The number of TransformerEncoder layers N is 5, while the number L of multi-head attention

[1] https://github.com/sjelic/vojvodina_crop_classification.git.

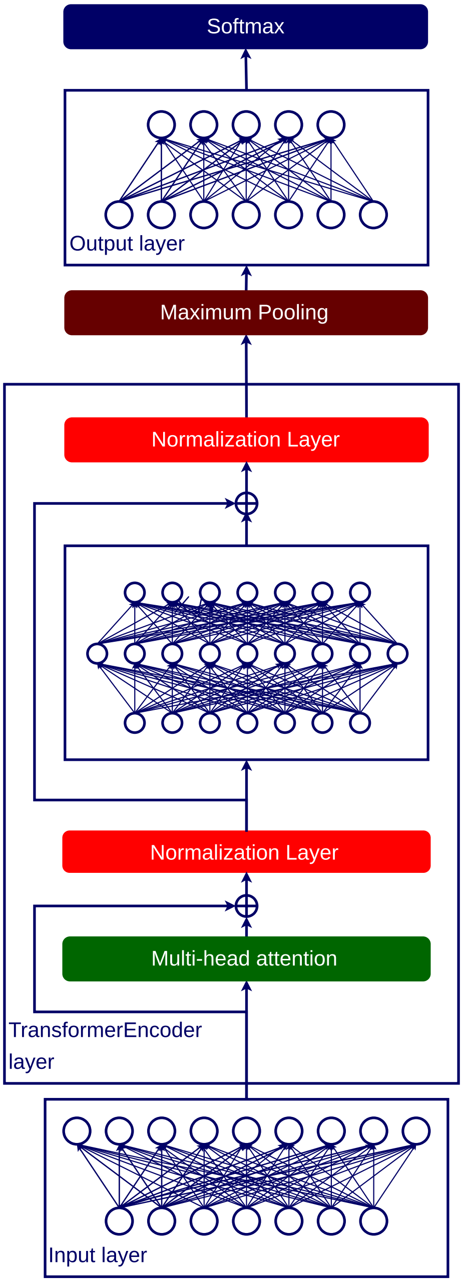


Fig. 6 TransformerEncoder model architecture

Table 3 Classification metrics for BM and BMP models trained on BreizhCrops dataset

Model	Overall Accuracy	Balanced accuracy	Macro f-score	Kappa
BM	0.81	0.59	0.81	0.76
BMP	**0.83**	**0.61**	**0.83**	**0.78**

Highest performing values are highlighted in bold

Table 4 Classification metrics for different models tested on Vojvodina dataset

Model	Overall accuracy	Class-mean recall	Macro f-score	Weighted f-score	Kappa
BMP	0.924	0.767	0.764	0.923	0.889
FE	0.936	0.896	0.896	0.936	0.906
PFT	**0.940**	**0.907**	**0.906**	**0.939**	**0.911**
FFT	0.939	0.896	0.905	**0.939**	**0.911**

Highest performing values are highlighted in bold

Table 5 Models' F1 score for each crop class, calculated on Vojvodina test dataset

	Barley	Maize	Rapeseed	Soya	Sugar beet	Sunflower	Wheat
BMP	0.856	0.946	0.857	0.838	0.0	0.913	0.936
FE	0.842	0.955	0.915	**0.879**	0.815	0.928	0.939
PFT	0.868	**0.958**	0.924	0.856	**0.857**	**0.934**	0.943
FFT	**0.873**	0.956	**0.93**	0.853	0.846	0.932	**0.947**

Highest performing values are highlighted in bold

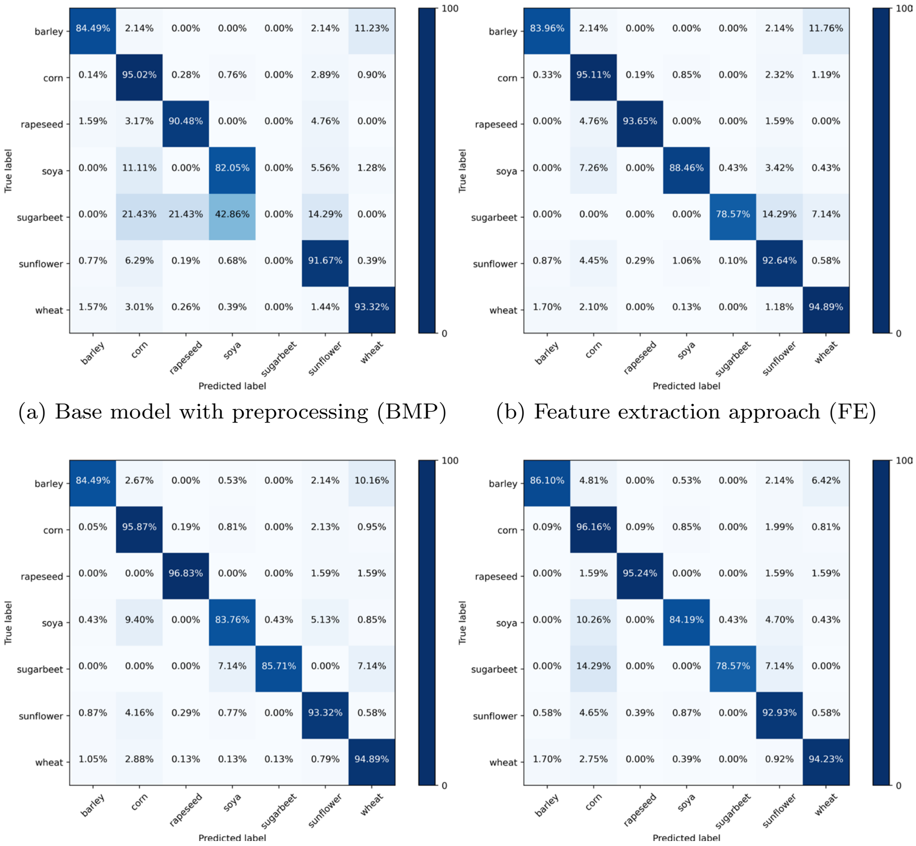
layers computed in parallel is 2. The training process is performed using the stochastic gradient descent minibatch algorithm on a workstation with Intel(R) Core(TM) i7-10700K CPU @ 3.80GHz, 2 x NVIDIA GeForce RTX 3070 and 130 GB RAM.

First, we report on the comparison of the BM and BMP models on the BreizhCrops dataset (Table 3).

Table 4 shows the performance of the transferred models and the BMP model trained on Vojvodina data. All models were evaluated using the same train-test splits from the 4-fold stratified cross-validation, with performance metrics calculated using the aggregated results from the 4 folds. The PFT model showed the highest performance among all tested models. The performances of the above models per class are shown in Table 5. The confusion matrices for all models tested on the Vojvodina dataset are shown in Fig. 7. They provide an overview of the performance of the classifiers for each of the crop classes.

Figure 8 gives a more insightful representation of the results of the confusion matrix by comparing the classifiers BMP and PFT in terms of the errors they make on each crop class. In addition to the percentage labels shown on the y axis, the bars of the graph show the total number of correctly and incorrectly classified samples (except for small errors, i.e. $< 3\%$ samples).

To statistically test whether the PFT model performs significantly better than the BMP model on the Vojvodina dataset, we used McNemar's [48] test. It is used to compare the performance of two classifiers based on a 2 by 2 contingency table of the predictions of the two classifiers. The results of the overall and class comparison are shown in Table 6. At a significance level of $\alpha = 0.05$, the values of the test statistic between -1.96 and 1.96 show that there is no significant difference between the performances of the classifiers, while values < -1.96 and $>$ 1.96 show that one



(a) Base model with preprocessing (BMP) (b) Feature extraction approach (FE)

(c) Partial fine-tuning approach (PFT) (d) Full fine-tuning approach (FFT)

Fig. 7 Confusion matrices

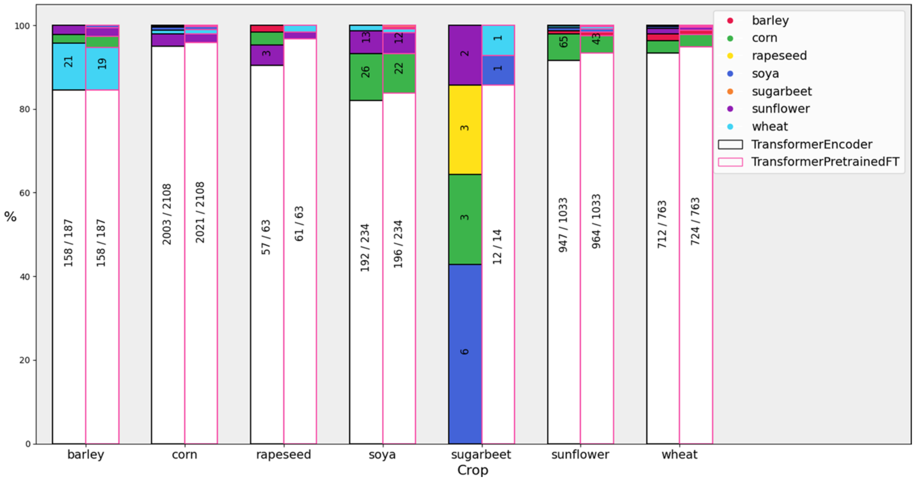


Fig. 8 Comparison of class errors and its distribution, for BMP and PFT models

Table 6 McNemar test statistic comparing PFT and BMP on Vojvodina dataset, overall and for specific crops

	Overall	Barley	Maize	Rapeseed	Soya	Sugar beet	Sunflower	Wheat
McNemar	**4.371**	0	**1.988**	1.633	0.784	**3.464**	**2.380**	**2.058**

Test statistic values that indicate a difference between the performance of the classifiers are highlighted in bold

classifier performed significantly better (< -1.96 directly trained model (BMP) performed better, > 1.96 transferred model (PFT) performed better).

Discussion

Although supervised deep learning models are data hungry and perform well on raw data by extracting and learning task-specific features, specific preprocessing steps and enrichment with provably important features are very important to improve the accuracy of the underlying deep learning model. It is not easy to systematically compare crop classification studies because of the variations in the quality of the training and testing samples that determine the performance of the classifier. Therefore, it is not possible to explicitly compare our results with other crop classification studies. However, we can make a comparison with the results published in [35], listed in Table 3. The overall accuracy of the model from [35] improves from 0.81 to 0.83, i.e., by 2.5%, while the balanced accuracy improves from 0.59 to 0.61, as can be seen in Table 3. This is undoubtedly due to the application of preprocessing steps that remove cloudy values from the time series, but also due to the included vegetation indices—in contrast to the work of Russwurm et al. [35].

On the other hand, the use of transfer learning proved to be very important for case studies where less in situ data or even lower quality data are available. Figure 7 shows that all models perform worst in predicting sugar beet, soya and barley. However, it is interesting to note that PFT has the highest improvement compared to BMP (Table 4). The results of the McNemar test from Table 6 confirm that the classifier PFT performs significantly better than BMP in classifying 4 (out of 7) classes. For example, in Fig. 7a, c we see that the recall for sugar beet improved from 0 to 85%.

Even more interesting is the fact that sugar beet was not present in the BreizhCrops dataset. This proves that the adapted model succeeded in learning domain-invariant features in the first layers through a partial fine-tuning strategy, which contributed significantly to the compilation of features in the last output layer (before Softmax). To the best of our knowledge, this is the first crop classification study based on satellite remote sensing that uses transfer learning based on labeled in-situ samples in both the source and target domains. All previous studies from this field [25, 39, 49] consider the models trained on existing crop classification maps such as CDL [50]. Most of these studies evaluated the performance of the transferred models by comparing the output to previously published classification maps from the target domain. In addition, most of these studies used images without atmospheric corrections (Sentinel-2 LIC) and focused on fewer crop types, often developed to distinguish between two classes, such as corn and soybean [39] or corn and rice [49].

Conclusion

We confirmed that the knowledge transfer from the source to the target domain was successful from both accuracy and computational resource perspectives. The simplicity and efficiency of the obtained model are promising for extending this research in two directions. More specifically, we ask whether the problem of optimal sampling design can be formulated and solved in such a way that the optimal size of the labeled data set in the target domain can be determined while maintaining the overall accuracy of the chosen model. The positive answer to this question would be of great impact on other fields of geoscience, including both classification and regression problems. On the other hand, the applicability of the model to distant regions that differ significantly in terms of climate and vegetation would also be an interesting research topic.

Acknowledgements
We would like to thank Russwurm et al. for making their data and code publicly available at https://breizhcrops.org/.

Author contributions
OA preprocessed and prepared the data. BB performed the primary literature review. SJ carried out the conception and design of the research, and code implementation. MK verified the relevance of the bibliography and the consistency of the results. All authors participated in analysis for this work and the writing of the manuscript. All authors read and approved the final manuscript.

Funding
This research was partially funded by the Science Fund of the Republic of Serbia—Program for Development of Projects in the Field of Artificial Intelligence, Grant No. 6527073 (project acronym CERES), and by the European Union's Horizon 2020 AgriCaptureCO2 project (Grant Agreement No. 101004282).

Availability of data and materials
Code from the paper is made publicly available at https://github.com/sjelic/vojvodina_crop_classification.git Names, formulas and paper references for the vegetation indicies used are available at https://github.com/sjelic/vojvodina_crop_classification/blob/master/veg_indicies.xlsx.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 30 January 2023 Accepted: 14 April 2023
Published online: 29 April 2023

References

1. Lee S. Application of artificial neural networks in geoinformatics. MDPI; 2018.
2. Chergui N, Kechadi MT. Data analytics for crop management: a big data view. J Big Data. 2022;9(1):1–37.
3. Wolfert S, Ge L, Verdouw C, Bogaardt M-J. Big data in smart farming—a review. Agric Syst. 2017;153:69–80.
4. Chergui N, Kechadi M-T, McDonnell M. The impact of data analytics in digital agriculture: a review. In: 2020 international multi-conference on "Organization of knowledge and advanced technologies" (OCTA). London: IEEE; 2020. p. 1–13.
5. Kovačević M, Bajat B, Gajić B. Soil type classification and estimation of soil properties using support vector machines. Geoderma. 2010;154(3–4):340–7.
6. Iman M, Rasheed K, Arabnia HR. A review of deep transfer learning and recent advancements. Preprint arXiv:2201.09679; 2022.
7. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
8. Tuia D, Persello C, Bruzzone L. Domain adaptation for the classification of remote sensing data: an overview of recent advances. IEEE Geosci Remote Sens Mag. 2016;4(2):41–57.
9. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big data. 2016;3(1):1–40.
10. Wardlow BD, Egbert SL. Large-area crop mapping using time-series Modis 250 m NDVI data: an assessment for the U.S. central great plains. Remote Sens Environ. 2008;112(3):1096–116. https://doi.org/10.1016/j.rse.2007.07.019.

11. You L, Wood S, Wood-Sichra U, Wu W. Generating global crop distribution maps: from census to grid. Agric Syst. 2014;127:53–60. https://doi.org/10.1016/j.agsy.2014.01.002.
12. Zhiyong F, Changdong L, Wenmin Y. Landslide susceptibility assessment through tradaboost transfer learning models using two landslide inventories. CATENA. 2023;222: 106799.
13. Fu D, Su C, Wang W, Yuan R. Deep learning based lithology classification of drill core images. PLoS ONE. 2022;17(7):0270826.
14. El Zini J, Rizk Y, Awad M. A deep transfer learning framework for seismic data analysis: a case study on bright spot detection. IEEE Trans Geosci Remote Sens. 2019;58(5):3202–12.
15. Astola H, Seitsonen L, Halme E, Molinier M, Lönnqvist A. Deep neural networks with transfer learning for forest variable estimation using sentinel-2 imagery in boreal forest. Remote Sens. 2021;13(12):2392.
16. Yandouzi M, Grari M, Indrissi I, Boukabous M, Moussaoui O, Ghoumid K, Elmiad AK. Forest fires detection using deep transfer learning. Forest. 2022;13(8):1.
17. Agarwal P, Jha G. Forest fire detection using classifiers and transfer learning. In: 2021 IEEE international conference on robotics, automation and artificial intelligence (RAAI). London: IEEE; 2021. p. 29–33.
18. Mirzaeitalarposhti R, Shafizadeh-Moghadam H, Taghizadeh-Mehrjardi R, Demyan MS. Digital soil texture mapping and spatial transferability of machine learning models using sentinel-1, sentinel-2, and terrain-derived covariates. Remote Sens. 2022;14(23):5909.
19. Padarian J, Minasny B, McBratney A. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma. 2019;340:279–88.
20. Zhao Y, Han S, Meng Y, Feng H, Li Z, Chen J, Song X, Zhu Y, Yang G. Transfer-learning-based approach for yield prediction of winter wheat from planet data and safy model. Remote Sens. 2022;14(21):5474.
21. Al Sahili Z, Awad M. The power of transfer learning in agricultural applications: Agrinet. Front Plant Sci. 2022;13: 992700.
22. Nowakowski A, Mrziglod J, Spiller D, Bonifacio R, Ferrari I, Mathieu PP, Garcia-Herranz M, Kim D-H. Crop type mapping by using transfer learning. Int J Appl Earth Obs Geoinf. 2021;98: 102313.
23. Jo H-W, Koukos A, Sitokonstantinou V, Lee W-K, Kontoes C. Towards global crop maps with transfer learning. Preprint arXiv:2211.04755; 2022.
24. Ma Y, Zhang Z, Yang HL, Yang Z. An adaptive adversarial domain adaptation approach for corn yield prediction. Comput Electron Agric. 2021;187: 106314.
25. Hao P, Di L, Zhang C, Guo L. Transfer learning for crop classification with cropland data layer data (CDL) as training samples. Sci Total Environ. 2020;733: 138869.
26. Keraani MK, Mansour K, Khlaifia B, Chehata N. Few shot crop mapping using transformers and transfer learning with sentinel-2 time series: case of Kairouan Tunisia. Int Arch Photogrammetry Remote Sens Spat Inf Sci. 2022;43:899–906.
27. Bursać P, Kovačević M, Bajat B. Instance-based transfer learning for soil organic carbon estimation. Front Environ Sci. 2022;2022:1.
28. Jones A, Fernandez-Ugalde O, Scarpa S. Lucas 2015 topsoil survey. Presentation of dataset and; 2020.
29. Janssen LLF, Middelkoop H. Knowledge-based crop classification of a landsat thematic mapper image. Int J Remote Sens. 1992;13(15):2827–37. https://doi.org/10.1080/01431169208904084.
30. Yi Z, Jia L, Chen Q. Crop classification using multi-temporal sentinel-2 data in the Shiyang river basin of china. Remote Sens. 2020;12(24):1. https://doi.org/10.3390/rs12244052.
31. Tatsumi K, Yamashiki Y, Canales Torres MA, Taipe CLR. Crop classification of upland fields using random forest of time-series landsat 7 etm+ data. Comput Electron Agric. 2015;115:171–9. https://doi.org/10.1016/j.compag.2015.05.001.
32. Zhong L, Hu L, Zhou H. Deep learning based multi-temporal crop classification. Remote Sens Environ. 2019;221:430–443. https://doi.org/10.1016/j.rse.2018.11.032. 428 citations (Crossref) [2023-03-28]. Accessed 2021-10-08.
33. Xu J, Yang J, Xiong X, Li H, Huang J, Ting KC, Ying Y, Lin T. Towards interpreting multi-temporal deep learning models in crop mapping. Remote Sens Environ. 2021;64:112599. https://doi.org/10.1016/j.rse.2021.112599. 39 citations (Crossref) [2023-03-31]. Accessed 2023-03-31.
34. Zhao H, Duan S, Liu J, Sun L, Reymondin L. Evaluation of five deep learning models for crop type mapping using sentinel-2 time series images with missing information. Remote Sens. 2021;3(14):2790. https://doi.org/10.3390/rs13142790.11 citations (Crossref) [2023-03-28]. Accessed 2021-11-01.
35. Rußwurm M, Pelletier C, Zollner M, Lefèvre S, Körner M. Breizhcrops: a time series dataset for crop type mapping. In: ISPRS—international archives of the photogrammetry, remote sensing and spatial information sciences XLIII-B2-2020; 2020. p. 545–1551.
36. Rußwurm M, Körner M. Self-attention for raw optical satellite time series classification. ISPRS J Photogramm Remote Sens. 2020;169:421–35.
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems; 2017. p. 6010.
38. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, vol. 27; 2014. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.
39. Wang Y, Feng L, Sun W, Zhang Z, Zhang H, Yang G, Meng X. Exploring the potential of multi-source unsupervised domain adaptation in crop mapping using Sentinel-2 images. GIScience and Remote Sens. 2022;59(1):2247–65. https://doi.org/10.1080/15481603.2022.2156123. 1 citations (Crossref) [2023-03-28]. Accessed 2023-03-20.
40. 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. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens Environ. 2012;120:25–36.
41. Hagolle O, Huc M, Desjardins C, Auer S, Richter R. MAJA algorithm theoretical basis document. Zenodo. 2017. https://doi.org/10.5281/zenodo.1209633.

42. Hagolle O, Huc M, Villa Pascual D, Dedieu G. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENμS and sentinel-2 images. Remote Sens. 2015;7(3):2668–91.
43. Maletić R, Popović B. Production of basic agricultural crops in ap vojvodina: trends and municipalities ranking. Ekonomika Poljoprivrede. 2010;57(2):275–92.
44. Reed BC, Brown JF, VanderZee D, Loveland TR, Merchant JW, Ohlen DO. Measuring phenological variability from satellite imagery. J Veg Sci. 1994;5(5):703–14.
45. Farahani A, Voghoei S, Rasheed K, Arabnia HR. A brief review of domain adaptation. In: Stahlbock R, Weiss GM, Abou-Nasr M, Yang C-Y, Arabnia HR, Deligiannidis L, editors. Advances in data science and information engineering; 2021. p. 77–894.
46. Venkateswara H, Panchanathan S. Introduction to Domain Adaptation. In: Venkateswara H, Panchanathan S, editors. Domain adaptation in computer vision with deep learning; 2020. p. 21.
47. Kouw WM, Loog M. An introduction to domain adaptation and transfer learning; 2019. arXiv:1812.11806. Accessed 2023-01-22.
48. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998;10(7):1895–923.
49. Ge S, Zhang J, Pan Y, Yang Z, Zhu S. Transferable deep learning model based on the phenological matching principle for mapping crop extent. Int J Appl Earth Observ Geoinform. 2021;102:102451. https://doi.org/10.1016/j.jag.2021.102451. 4 citations (Crossref) [2023-03-28]. Accessed 2023-03-28.
50. Boryan C, Yang Z, Mueller R, Craig M. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 2011;26(5):341–58. https://doi.org/10.1080/10106049.2011.562309.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.