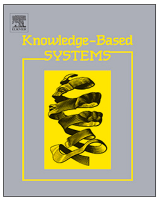
Knowledge-Based Systems 254 (2022) 109644

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



DIAFAN-TL: An instance weighting-based transfer learning algorithm with application to phenology forecasting

M.A. Molina-Cabanillas [a], M.J. Jiménez-Navarro [b], R. Arjona [a], F. Martínez-Álvarez [b], G. Asencio-Cortés [b,*]

[a] *easytosee AgTech S.L., Diego Martínez Barrio 10 (3rd floor), ES-41013 Seville, Spain*
[b] *Data Science & Big Data Lab, Universidad Pablo de Olavide, ES-41013 Seville, Spain*

ARTICLE INFO

Article history:
Received 20 December 2021
Received in revised form 26 July 2022
Accepted 5 August 2022
Available online 10 August 2022

Keywords:
Transfer learning
Phenology
Time series forecasting
Supervised learning

ABSTRACT

The agricultural sector has been, and still is, the most important economic sector in many countries. Due to advances in technology, the amount and variety of available data have been increasing over the years. However, compared to other economic sectors, there is not always enough quality data for one particular domain (crops, plantations, plots) to obtain acceptable forecasting results with machine learning algorithms. In this context, transfer learning can help extract knowledge from different but related domains with enough data to transfer it to a target domain with scarce data. This process can overcome forecasting accuracy compared to training models uniquely with data from the target domain. In this work, a novel instance weighting-based transfer learning algorithm is proposed and applied to the phenology forecasting problem. A new metric named DIAFAN is proposed to weight samples from different source domains according to their relationship with the target domain, promoting the diversity of the information and avoiding inconsistent samples. Additionally, a set of validation schemes is specifically designed to ensure fair comparisons in terms of data volume with other benchmark transfer learning algorithms. The proposed algorithm, DIAFAN-TL, is tested with a proposed dataset of 16 plots of olive groves from different places, including information fusion from satellite images, meteorological stations and human field sampling of crop phenology. DIAFAN-TL achieves a remarkable improvement with respect to 15 other well-known transfer learning algorithms and three nontransfer learning scenarios. Finally, several performance analyses according to the different phenological states, prediction horizons and source domains are also performed.

1. Introduction

The agricultural sector has been presented as one of the most important economic sectors, as it has provided the basic livelihoods of the population throughout history. In recent years, due to a growing world population, increased crop security and climate change, monitoring, as well as introducing new technologies in this sector have become essential. These new technologies make it possible to monitor the current state of crops, providing valuable information to help farmers in their decision-making.

Among all the parameters that can be monitored, we find the phenological stages of the crop to be one of the most important. Precision monitoring of phenological stages over time is essential to understand the impact of climate change on plants. Recent studies show the reaction of such changes throughout the timing of phenological stages, thus affecting crop productivity [1].

Moreover, if such monitoring is combined with the prediction or forecasting of these phenological stages through the application of artificial intelligence algorithms, it becomes a useful tool for small and large farmers for planning irrigation schedules, pesticide and disease control and fertilization management [2].

However, although all the new technologies developed are positioned as fundamental tools to help farmers make decisions, their implementations in the agricultural sector have been stagnant for several years. However, in recent years, there has been a greater interest in their applications, especially by large companies that are accessing the agricultural sector and require a greater cost and farm management control.

This incipient data collection means that a sufficiently large historical database is not yet available to be effectively used to train machine learning-based models. New techniques capable of extracting relevant knowledge from low available data are essential, especially in geographies where such data are scarce. Moreover, those techniques should produce general models able

* Corresponding author.
E-mail addresses: miguelangel.molina@ec2ce.com (M.A. Molina-Cabanillas), mjjimnav@upo.es (M.J. Jiménez-Navarro), ricardo.arjona@ec2ce.com (R. Arjona), fmaralv@upo.es (F. Martínez-Álvarez), guaasecor@upo.es (G. Asencio-Cortés).

https://doi.org/10.1016/j.knosys.2022.109644
0950-7051/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

to accurately predict the evolution of phenology in different crops. For this reason, data with diversity, in terms of both input features and label spaces, must be provided as training for such models. In this context, the transfer learning paradigm has proven to be as an effective way to capture diverse patterns from different problems (domains) and use them together to improve the prediction accuracy on a target domain. In this study of phenology forecasting it is assumed that each domain represents a crop of olive grove in a different place.

Phenology forecasting is an especially complex problem due to the relationships among the large number of different factors that can affect crop evolution. For this reason, a vast number of variables taken from different sensors and places could be needed to extract the complex patterns that can explain the phenology evolution and, consequently, produce accurate forecasting models.

In general, time series forecasting models that are based on machine learning algorithms use input training sets with lagged values of the time series as features and future ahead values as the different classes to predict the regression. Such data preprocessing is simple and effective in many cases, but some relationships among time series values cannot be captured by machine learning algorithms. For example, complex relationships between averages of values in different windows of the time series, are especially present in agronomics problems. Averaging very old values of some variables of the time series could be useful as input features along with more detailed recent values.

Usually, many transfer learning proposals for supervised learning published in the literature are validated including a comparison of those proposals with nontransfer learning strategies using the same datasets [3]. However, this type of comparison is unfair and it does not reflect whether the performance improvement of a transfer learning algorithm occurs due to the nature of the algorithm of the higher volume of data it receives as training, compared to the lower volume given to the nontransfer learners. Specifically, transfer learning algorithms receive as training one or more source domains as training along with a training part of the target domain (assuming a single-target approach). However, there are usually two nontransfer learning strategies: (1) The first strategy includes a base learner trained with a source domain and tested with a subset of the target domain. (2) The second strategy includes a base learner trained with a subset of the target domain and tested with the rest of samples. Comparing a transfer learning algorithm with both strategies is unfair, because the volume sizes are always different and such a difference could be considerably large (especially in the second nontransfer learning strategy).

To solve the problems previously described, a new transfer learning algorithm, DIAFAN-TL, is proposed and applied to phenology forecasting in several crops of olive groves. DIAFAN-TL is an instance weighting-based algorithm that weights each sample from multiple source domains in such a way that favors information diversity and penalizes false neighbors. The dataset to train and test the algorithm was specifically collected and prepared for this work, integrating an information fusion of satellite images, meteorological stations and human field sampling of phenology from olive grove crops. Moreover, vast feature engineering for the time series was performed, including variables that were not lagged but averages, maximums, minimums, sums, dispersions and value changes for different window sizes and temporary displacements.

To validate the proposal, a fair comparison with 15 other well-known transfer learning algorithms and three nontransfer learning strategies using the same volume of training data for all the algorithms was performed. The aim was to isolate the cause of the performance improvement to the nature of the algorithm, as explained before. Moreover, several performance analyses according to the different phenological states, prediction horizons and source domains were also performed.

The rest of the paper is structured as follows: Section 2 overviews recent and relevant papers in the field of transfer learning, as well as its application to phenology forecasting. Section 3 describes the proposed methodology and how it is applied to predict a phenology time series. Section 4 reports and discusses the results achieved from the different experiments that have been carried out. Finally, Section 5 summarizes the conclusions.

2. Related works

Transfer learning is becoming one of the fields of research where most of effort is being put [4]. In fact, many applications can now be found in the literature.

Within the possible classifications and configurations of transfer learning, we can find domain adaptation, unsupervised learning and even fine-tune.

Unsupervised learning is the case of abundant labeled source data and no labeled target data. In fact, [5] refers to inductive transfer learning as the case of having available labeled target domain data, transductive transfer learning as the case of having labeled source and no labeled target domain data, and unsupervised transfer learning as the case of having no labeled source and no labeled target domain data. A Unsupervised transfer learning method that mitigates nontransferable prior-knowledge by self-supervision can be seen in [6]

Fine tune transfer learning enables to start with a model, pretrained for a specific task, and then fine-tune (train) only certain layers of the neural network for a related but different target task. However, the selection of fine-tunable layers is one of the major problems of such an approach. In [7] a new method for the selection of fine-tunable layers for a target dataset under the given constraints is described.

Domain adaptation, which is particularly interesting in the context of this paper since instance weighing is used as a technique included in this area, is the process of adapting one or more source domains for the means of transferring information to improve the performance of a target learner. The domain adaptation process attempts to alter a source domain in an attempt to bring the distribution of the source closer to that of the target.

Exploring into instance weighting approaches, included in the domain adaptation configuration, many studies have been developed, most of them refers to inductive transfer learning. For example, in [8], these techniques are applied to model, analyze and detect reference points of faces. The importance of applying transfer by weighing instances during learning is emphasized over its application afterward to adjust the models.

In [9], two methods for regression based on importance weighting are presented, where, for each instance of the domain data, a weight is assigned such that the data contributes positively to the prediction of the target data.

Additionally, in [10], a new metric based on weighted instances is used to measure the similarity between two domains using common spatial patterns to reduce the amount of training data that needs to be collected at the beginning of each session to calibrate the parameters of brain–computer interfaces.

Further advances and an extended version of these techniques are used in other fields, especially in image processing. Furthermore, the technique itself has been studied in some recent articles due to the benefits it presents [4,5,11].

Other works applying these transfer learning and deep transfer learning techniques can be found in Refs. [12,13]. In [13], a domain adaptation extreme learning machine (DAELM) was developed to establish a simple soft sensor model suitable for multigrade processes with limited labeled data, inspired by the idea of transfer learning. In [12], a novel framework of an adversarial transfer learning (ATL)-based soft sensing method was designed for the quality inference of multigrade processes. By treating each grade as a domain, the concept of ATL was adopted to learn a suitable feature transformation between different domains, which reduced the data distribution discrepancy and enriched the information provided by the target domain containing limited labeled data.

Distance studies have been developed in works with deep transfer learning techniques and they have achieved remarkable results with reference to the effects of dataset similarity. Thus, in [14] some time series were classified using a distance based approach.

All the techniques applied in the works in the previous paragraphs were gradually being extended to the agricultural sector, making up most of the information generated in recent years. These techniques were not yet widespread in the agricultural sector. Only some papers introducing these techniques in this sector can be found in the literature, such as the one described in [15] where basic algorithms were used for yield prediction, disease detection or crop quality.

In this kind of study, satellite images play a key role. In recent years, satellite observations have been widely used in research work requiring difficult to obtain field data, due to their public accessibility and low cost, covering a large area with increasingly accurate resolutions [16–20].

Hence, time series of the MODIS NDVI index were used to distinguish different crop types according to their phenological evolution in [21]. Even in [22], these images were used to detect the optimal conditions for a given crop type and sowing areas, being able to distinguish each one before harvest.

Another work that also used the MODIS satellite (and the corresponding NDVI index) to perform a spatiotemporal analysis of crop phenology in a given area can be found in Ref. [23] which aimed to classify crop types. A similar study can be found in Ref. [24].

Other early works published in the agricultural sector that apply new machine learning techniques, including deep learning and transfer learning can be found in Refs. [25–28]. In [27] deep learning techniques are applied to detect phenological stages of the rice crop through images taken by air vehicles to estimate of production and harvest dates. Additionally, in [29,30], some works can be seen using deep learning for the recognition of phenological patterns or for the prediction of the phenology and incidence of pests and diseases in crops. Especially in Ref. [29] the data taken for the analysis comes from half-hourly captures from cameras mounted on agronomic stations.

Phenological patterns of crop growth are similar in different parts of the world. This makes it possible to apply models trained in some areas to others. In fact, [31] proposed a method of crop identification with LANDSAT satellite images (with a better resolution than MODIS) and the NDVI index, while the authors in [32] identified the phenological differences between different crops.

Another work that has also used satellite imagery to apply deep transfer learning techniques for yield prediction is described in Ref. [33]. It predicts the yield of soybean crops in Argentina and uses this information to predict these crops in Brazil, due to the limited availability of data in this country.

Transfer learning techniques have also been applied to the agricultural sector for crop type detection in different regions with limited access [28].

Fuzzy-based approaches can be found in the literature in the transfer learning context too. Hence, in [34] a fuzzy system with knowledge-leverage capability is proposed in order to solve problems where the available data from that scene are insufficient. In [35] Transfer learning is introduced for recognition of epileptic electroencephalogram signals. In [36] a new feature transformation method is developed with fuzzy systems. In [37] the concept of transfer learning is applied to prototype-based fuzzy clustering. Finally, in [38], the most recent advances in deep transfer learning are presented.

Finally, more advanced techniques, such as image-based deep transfer learning, have been used for disease detection in crops, as can be read in Ref. [25]. Papers explaining how transfer learning techniques work can be found in Refs. [26,39].

3. Methodology

The goal of the methodology is to create an end-to-end validation framework that ensures the benefits of using transfer learning in a machine learning task starting from some data sources as input and a set of metrics as output. Moreover, a novel method in the field of transfer learning is assessed by means of such a validation framework.

This section is structured as follows. Section 3.1 formulates the problem. Section 3.2 describes the data sources obtained in the application studied in this paper. Section 3.3 describes the preprocessing functions applied to the data obtained. Section 3.4 describes the data engineering step where expert knowledge is used to create features. Section 3.5 describes the validation framework divided into different schemes to make a fair comparison. Section 3.6 describes some background of the transfer learning area that this paper is focused on. Section 3.7 shows the training process of the proposed method.

3.1. Problem formulation

Before diving into details of mentioned techniques, let us first provide a definition of transfer learning. Let $D_S = \left\{(x_S^{(i)}, y_S^{(i)})\right\}_{i=1}^{L}$ denote a data set from a source domain $\mathcal{D}_S = \{\mathcal{X}_S, p_S(x)\}$ and source task $\mathcal{T}_S = \{\mathcal{Y}_S, p_S(y|x)\}$, where $x_S^{(z)} \in \mathcal{X}_S$, $y_S^{(z)} \in \mathcal{Y}_S$, $\mathcal{X}_S$ is the input space, $\mathcal{Y}_S$ is the output space, $p_S(x)$ is the marginal probability distribution and $p_S(y|x)$ is the posterior probability distribution. Similarly, let us define the target data, domain and task as follows: $D_T = \left\{(x_T^{(i)}, y_T^{(i)})\right\}_{i=1}^{M}$, $\mathcal{D}_T = \{\mathcal{X}_T, p_T(x)\}$ and $\mathcal{T}_T = \{\mathcal{Y}_T, p_T(y|x)\}$. If there is more than one source domains, we have a multi source problem. In this case, let us denote a multi source domain $\mathcal{D}_{S_i} = \{\mathcal{X}_S, p_S(x)\}_{i=1}^{N}$ and source task $\mathcal{T}_{S_i} = \{\mathcal{Y}_S, p_S(y|x)\}_{i=1}^{N}$, where each domain is defined as described above.

The goal of transfer learning is to use the knowledge learnt from the source/s to improve the predictive performance of a predictive model $f_T(x) : \mathcal{X} \to \mathcal{Y}$ for the target, despite the fact that the source and target tasks and domains may differ.

The case under study is collected in the category of inductive transfer learning, where the label information of the target domain instances is available. Inductive transfer learning approaches transfer knowledge between different tasks (e.g. $\mathcal{T}_S \neq \mathcal{T}_T$) while $\mathcal{D}_S = \mathcal{D}_T$ or $\mathcal{D}_S \neq \mathcal{D}_T$.

The general methodology is illustrated in Fig. 1, showing the diagram with all steps to make predictions. Subsequent sections are devoted to explain every step involved in such general flowchart.

3.2. Data acquisition

The first step is data acquisition, where different data sources are joined by time to generate the dataset. The data sources used in this paper are:

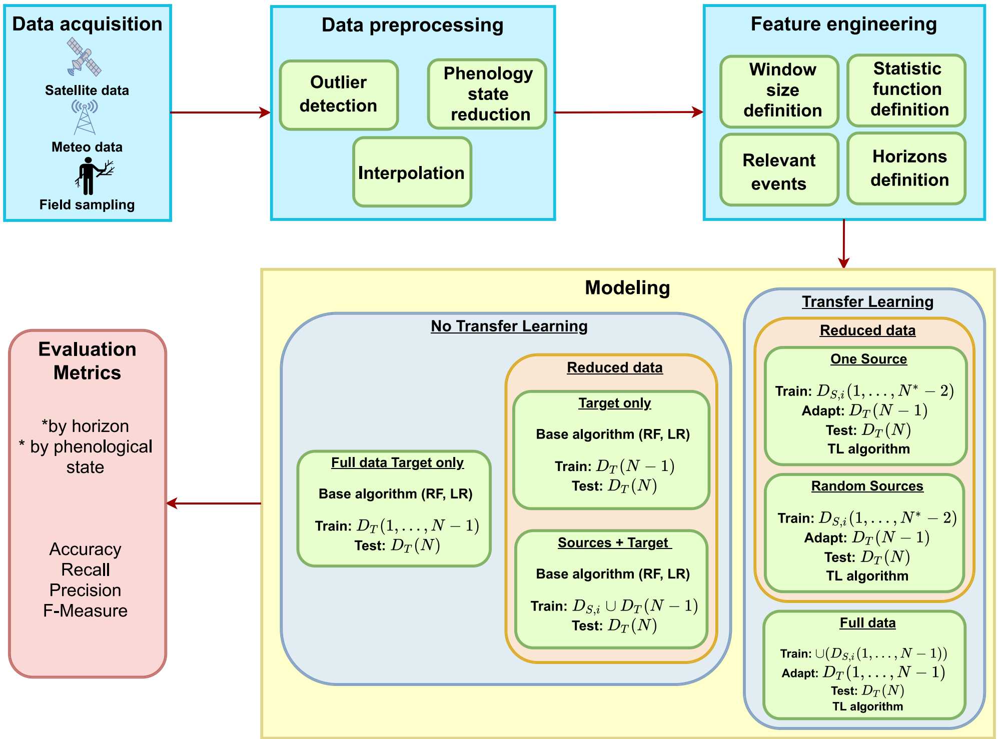


Fig. 1. Flowchart for the proposed methodology.

Table 1
Sentinel bands used from satellite images.

Band number	Band description	Wavelength range (nm)	Resolution (m)
B1	Coastal aerosol	433–453	60
B2	Blue	458–523	10
B3	Green	543–578	10
B4	Red	650–680	10
B5	Red-edge 1	698–713	20
B6	Red-edge 2	733–748	20
B7	Red-edge	773–793	20
B8	Near infrared (NIR)	785–900	10
B8A	Near infrared narrow (NIRn)	855–875	20
B9	Water vapor	935–955	60
B10	Shortwave infrared/Cirus	1360–1390	60
B11	Shortwave infrared 1 (SWIR1)	1565–1655	20
B12	Shortwave infrared 2 (SWIR 2)	2100–2280	20

1. The first dataset is the satellite data. The dataset is taken from Sentinel, with historical data since 2015, 5 day periodicity and up to 12 bands with different wavelengths (see Table 1).
2. The second dataset is the meteorological data. Different meteorological stations measure different variables, such as temperature, humidity, solar radiation and precipitation. These stations have a daily sampling periodicity.
3. The third dataset is the field sampling data. This is the target to forecast in this work. It consists of the phenology sampling obtained by experts of crops. The phenology usually consists of a set of states for which the crop evolves over time. Such data are usually collected on a weekly basis in this work.

In this work, the same data sources are obtained from different plots where phenology has been sampled from their crops. Each plot is considered a different domain as each plot demonstrates different behavior for the same task. With the data obtained, a different dataset is constructed for each domain and plot with a feature for each variable of each data source obtained. The dataset constructed is called the event table and is the starting point of the entire process.

3.3. Data preprocessing

Field sampling data for each domain is collected on a weekly basis. As the data depends on the attendance of a technician at the plot to collect the information, it is possible that there are some weeks without available data.

A simple strategy is followed to impute data. We let P_t be the phenology value observed at time t. If subsequent phenology values $P_{t+1},\dots,P_{t+n}$ $(n \geq 1)$ has an unknown value, P_{t+n+1} has the same value as P_t, and the phenologies $P_{t+1},\dots,P_{t+n}$ are imputed with the value of P_t.

3.4. Feature engineering

Feature engineering is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning algorithms. To this end, the following functions have been carried out for each feature:

1. The first function is for window size definition. This refers to the definition of the time intervals used to calculate different statistical functions.
2. The second function is the statistic function definition. This is the definition of the statistical functions to be applied to each of the variables in these time windows. The statistical functions used in this work are: means, accumulates, maximums and minimums, and they are calculated between the values of a variable in a temporal space.
3. The third function is for relevant events. These are the values of the series that exceed some thresholds are marked as relevant events. Those events are used as the beginning or end for further window calculation.
4. The fourth function id for target variables. These are the variables that are used for prediction. In this case, the phenological state is predicted in the following four weeks after the prediction date.

In that way, averages of temperature or precipitation accumulations in windows to the past in the short, medium and long term are calculated. Other extreme conditions are also considered into account when defining these features, e.g., extreme temperature thresholds or extreme rainfall.

3.5. Validation schemes

To compare the improvements obtained by applying transfer learning techniques with traditional techniques, a validation framework composed of different schemes is built. This framework arises from the need to make fair comparisons in an imbalanced data context from different domains. The goal of these schemes is to ensure the benefits of using transfer learning in different scenarios balancing the data and ensuring that the comparisons are fair enough to make conclusions.

The validation schemes are applied to our method and traditional algorithms extracted from the literature as a baseline. These algorithms are shown in the comparative tables in Section 4.

The taxonomy of the different validation schemes has two main branches: no transfer learning and transfer learning. The no transfer learning includes all the techniques that do not apply any transfer learning and consider the data from different domains to be the same. The transfer learning branch uses several techniques that consider the different domains and apply adaptation from the sources and domains to the target domain. Moreover, these two branches are divided into reduced data (if we limit the source data) and full data (if we use all the data available from the sources).

The no transfer learning scheme is described below:

1. Reduced data
   (a) We described the reduced data for the target only. In this case, only data from the target domain are used for validation. This is the most basic scheme reflecting cases where very limited historical data is available. Using random forest (RF) and logistic regression (LR) as the base algorithms, the last year of the target domain is taken as the test and the penultimate year as training.
   (b) We described the reduced data for sources + targets. In this experiment, we study the influence of the source domains for predicting the test set by linking the training sets. For this purpose, we take all the years of the target domain except the last one and all the years of the source domains (domain by domain) except the last ones for the training set. The last year of the target domain is again left as the test set.
2. Full data. For this experiment, all years from all the source domains and the target domain except for their last years are used. Using RF and LR as the base algorithms, the last year of the target domain is taken as the test and the rest of the years are used for the training.

Analogously, the transfer learning scheme, based on RF and LR is detailed below:

1. The reduced data can divided into one source or random sources:
   (a) When the reduced data is considered one source, in this experiment, each source domain is taken as a separate source. For each iteration, the following steps are followed:
      i. The model was trained with all years of the source domain except the last two years.
      ii. We retrain the model with the penultimate year of the target domain.
      iii. It is tested with the last year of the target domain.
   (b) When the reduced data is divided into random sources, the same methodology is followed as in the previous experiment but each year of the training set might belong to a different source domain and this experiment is repeated n times.
2. The full data in this experiment is intended to prove that transfer learning improves the simple union of all the information provided by the source sets. For this purpose, we train with all the source domains except for those from their last year. Subsequently, it is retrained with all the years of the target set except the last one. Again, the last year is left as the test set.

3.6. Proposed method

In this paper, a novel instance weighting transfer learning method has been developed based on the use of a specific metric named DIAFAN that provides diverse information and penalizes false neighbors in training sets.

Instance weighting is a subfield in transfer learning whose objective is, during the training step, to assign a weight to the instances from the sources and domains to learn more from the most informative examples and prove the final results. In the literature, these weights are usually assigned by a distance function, such as log-likelihood or cosine. In our method, a Siamese network was trained to estimate the distance function between instances from different domains [40]. The next section describes the complete training process used in this method.

3.7. Training process

To calculate the DIAFAN metric it is necessary to have the data of all the domains: the target and all the sources. The procedure consists in weighting the instances of the source domains by their similarity in attributes and classes with the instances of the target

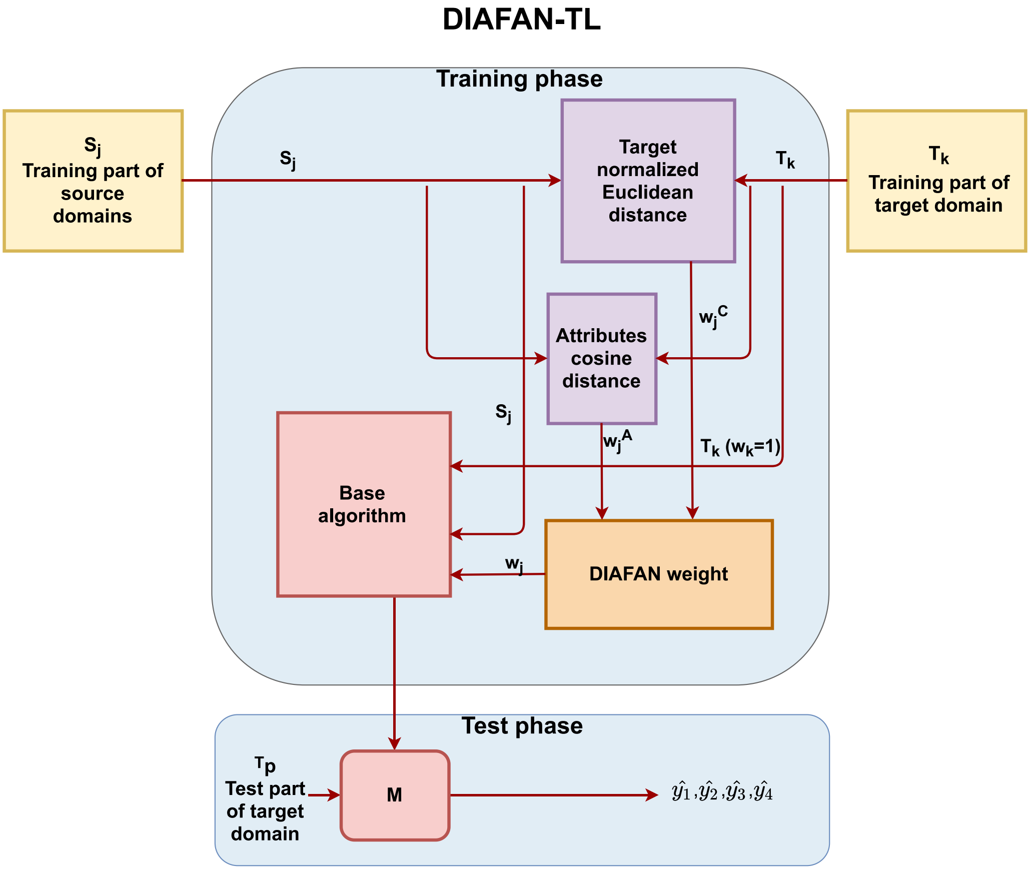

DIAFAN-TL
Training phase
S_j Training part of source domains
S_j
Target normalized Euclidean distance
T_k
T_k Training part of target domain
w_j^C
Attributes cosine distance
S_j
w_j^A
T_k (w_k=1)
Base algorithm
w_j
DIAFAN weight
Test phase
T_p Test part of target domain
M
ŷ1 ,ŷ2 ,ŷ3 ,ŷ4

Fig. 2. The training and test phases of the proposed transfer learning methodology to assign weights to instances from source domains.

domain, so that the preprocessing of assigning weights to the instances of the source domains is done jointly with the target domain. Then, the data of all the training samples of the target set are taken and the weighted instances of the source domains are added. Finally, a single model is trained with all the data: target domain and source domains together.

The DIAFAN metric can be included in contrastive algorithms. Contrastive learning, in general, learns a distance function but there are cases in which this function is already given, and this is the case in which the DIAFAN metric is included. With respect to algorithms in which a distance function is learned, the advantage of the DIAFAN metric is that it does not fit a black box, where the weights are iteratively adjusted during the learning process. The DIAFAN metric is based on prior similarities between attributes and classes and does not rely on the number of iterations to adjust it. This helps in the greater efficiency of DIAFAN in the case of problems with model re-training since there is no need to re-train to re-obtain the weights with consequent time savings. If new domains are included, it is only necessary to calculate the metric for them and not for the previous ones. In addition, with DIAFAN less weight is given to instances of false neighbors by identifying such cases. Also in our case, the weights are more interpretable than mathematically learned weights.

The training process of our method consists of calculating a weight for each sample from source domains and building a model generated by a base classifier using weighted instances from source domains along with the training part of the target domain with a fixed weight of 1.

First, a weight (w_j^C) is assigned to each instance from each source domain. This weight is assigned based on the similarity, assessed by a Euclidean distance, among the h future phenology values of such instances and each instance of the target domain.

Second, for each pair of instances, the cosine distance between attributes is also calculated, from which we will obtain another weight (w_j^A).

With both weights, a final weight is calculated. This distance is named DIAFAN, which stands for DIversity Avoiding FAlse Neighbor. It is defined in Eq. (1).

$$\mathit{DIAFAN\ Weight} = \frac{\sqrt{(S-1)^2 + A^2}}{\sqrt{2}} \tag{1}$$

In Eq. (1), S represents the weight that measures the similarity between classes and A represents the weight given by the cosine distance function between attributes. For DIAFAN ranges between [0, 1], the extreme scenarios are described as follows:

1. $S = 0$ and $A = 0$. In this case, we have the minimum distances between both classes and attributes; thus, they seem very similar to each other. It is therefore given a weight $\frac{1}{\sqrt{2}}$ in order to favor the third point.
2. $S = 1$ and $A = 0$. In this case, the classes are not very similar and the attributes are very much alike. A weight of 0 is given so as not to introduce false neighbors.
3. $S = 0$ and $A = 1$. In this case, the attributes are very similar, and the classes are very similar. This case brings variability and is given a weight of 1.
4. $S = 1$ and $A = 1$. When everything has little resemblance, it is given a weight, $\frac{1}{\sqrt{2}}$ as this is a good example to transfer knowledge.

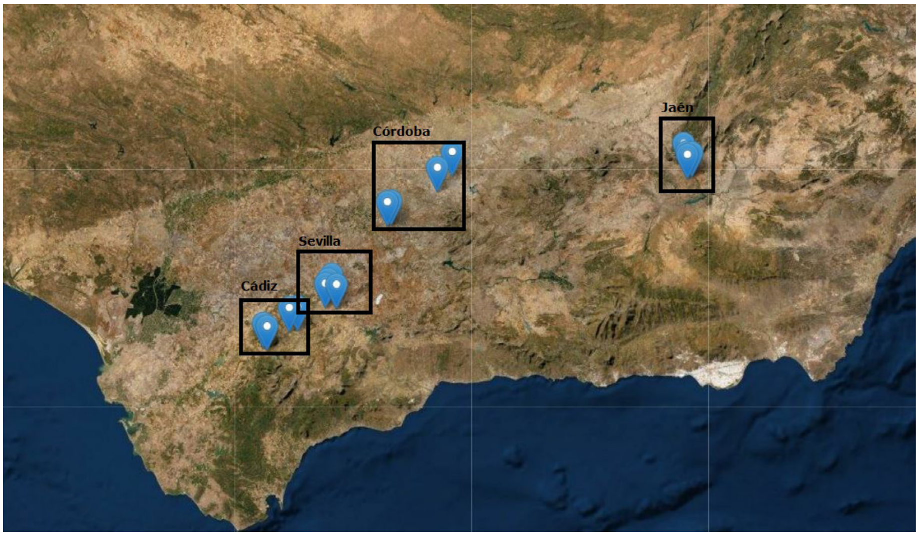
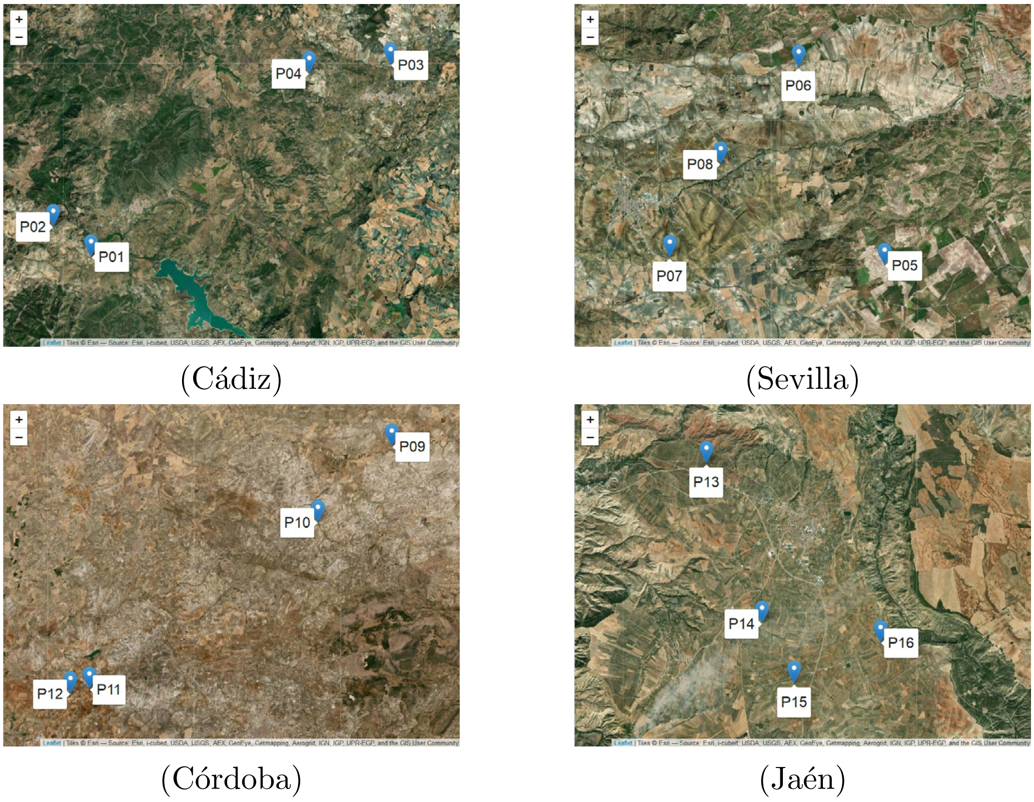


Fig. 3. Geographic location of the 16 parcels for the proposed study of transfer learning (Andalusia, Spain).


P04
P03
P02
P01
P06
P08
P07
P05
P09
P10
P12
P11
P13
P14
P16
P15
Leaflet | Tiles © Esri — Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community

(Cádiz)

(Sevilla)

(Córdoba)

(Jaén)

Fig. 4. More detailed view of plots for each region.

Finally, the pair of instances and the newly calculated weight feed the base algorithm (RF and LR). The base algorithm uses the weights calculated by the DIAFAN weight for the sources and domains training set and a value of $W_k = 1$ for the target domain training instances. The model is trained considering these weights and then, once trained, tested with the test dataset. Fig. 2 shows the complete pipeline followed to train and test the model.

Finally, in the prediction phase, the test part of the target domain is passed to the previously adjusted model, obtaining a prediction for the four weeks after the prediction date (Fig. 2).

3.8. Transfer learning benchmark algorithms

To compare the effectiveness of our proposed method, we selected a list of previously published well-known algorithms of transfer learning. These algorithms are briefly described below:

- **STRUT** [41]. This algorithm adapts a decision tree trained on the source samples to the target samples by discarding all numeric threshold values in the tree and working top-down, selecting new thresholds using the target examples.

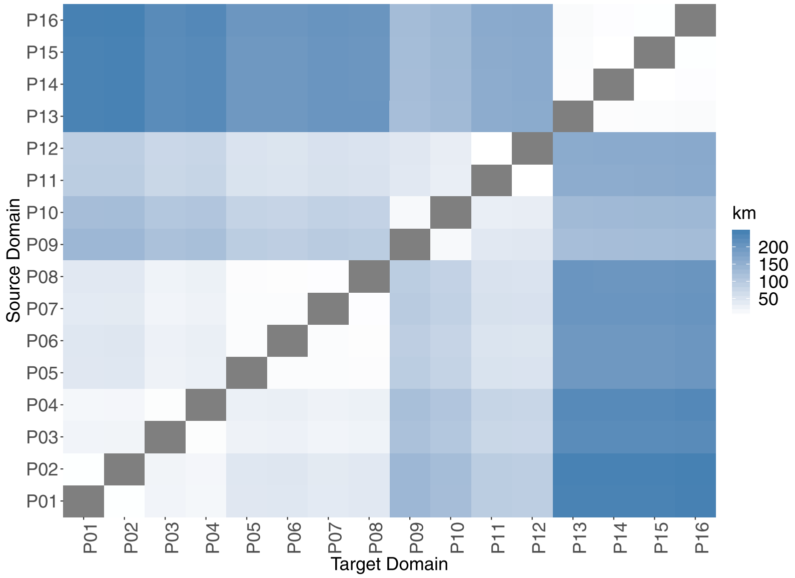


Fig. 5. Geographic distance between domains (plots) (in km).

Table 2
Characteristics of the 16 domains (plots) used in the study.

Code	Plot name	Region	Coordinates	Altitude	Surface	Slope	Dry/Irrigated	Density	Main variety
P01	C1_Algodonales1	Cádiz	36.86,−5.43	360	8,54	19	Dry	58	Lechín, Zorzaleño, Ecijano
P02	C1_Algodonales2	Cádiz	36.87,−5.45	770	2,26	19	Dry	138	Lechín, Zorzaleño, Ecijano
P03	C1_Olvera1	Cádiz	36.94,−5.25	350	1,04	5	Dry	134	Picual/Marteño
P04	C1_Olvera2	Cádiz	36.94,−5.30	425	3,1	11	Dry	74	Lechín
P05	C2_LosCorrales	Sevilla	37.05,−5.01	600	1,93	15	Dry	119	Picual/Marteño
P06	C2_ElSaucejo1	Sevilla	37.10,−5.04	460	31,59	15	Irrigated	156	Manzanillo
P07	C2_ElSaucejo2	Sevilla	37.05,−5.08	520	1,78	20	Dry	120	Hojiblanco
P08	C2_ElSaucejo3	Sevilla	37.07,−5.06	510	2,87	25	Dry	150	Hojiblanco
P09	C3_Baena	Córdoba	37.67,−4.33	300	42,53	3	Irrigated	154	Picual/Marteño
P10	C3_NuevaCarteya	Córdoba	37.60,−4.42	460	9,58	9	Dry	76	Picual/Picudo
P11	C3_Aguilar	Córdoba	37.44,−4.69	280	6,3	1	Dry	194	Picual/Marteño
P12	C3_PuenteGenil	Córdoba	37.44,−4.71	320	19,52	1	Irrigated	208	Manzanillo
P13	C4_PozoAlcon1	Jaén	37.71,−2.96	700	6,015	2	Irrigated	178	Picual
P14	C4_PozoAlcon2	Jaén	37.68,−2.94	700	2,86	5	Irrigated	200	Picual/Marteño
P15	C4_PozoAlcon3	Jaén	37.66,−2.93	700	1,09	1	Irrigated	140	Picual/Marteño
P16	C4_PozoAlcon4	Jaén	37.67,−2.91	700	0,96	3	Irrigated	92	Picual

- **SER** [41]. This algorithm pairs two local transformations of a decision tree structure. It first specializes rules induced over the source data to the target data (expansion) and then generalizes rules induced over the source data by pruning (reduction).
- **MIX** [42]. This algorithm generates two forests using both SER and STRUT, and then it defines MIX as a majority voting ensemble whose underlying model is the union of all the trees. in these forests.
- **TreesMixedEntropy** [41]. Tree construction is done via shared information gain (IG). The final IG is calculated using a weighted mixture of IGs of each training set.
- **NaiveBiasRegularizator** [41]. The weights in the original forest are changed from a uniform distribution to one, which favors trees with lower error rates on the available target training samples.
- **NaivePrunningTree** [41]. This algorithm constructs a tree-based algorithm using the source domain and then uses the target domain to perform pruning on the original forest.
- **NaiveRelabelingTree** [41]. This algorithm updates the leaves of a forest trained on the source examples using the target samples.
- **CORAL** [43]. This algorithm minimizes domain shift by aligning the second-order statistics of source and target distributions without requiring any target labels.
- **KMM** [44]. This algorithm resolves the estimation problem of the above unknown ratios by matching the means between the source-domain and the target-domain instances in a reproducing kernel Hilbert space (RKHS).
- **TCA** [45]. This algorithm learns a linear mapping from an empirical kernel feature space to a low-dimensional feature space using the maximum mean discrepancy (MMD).
- **ConsensusRegularization** [46]. This algorithm exploits the distribution differences and learns the knowledge among training data from multiple source domains to boost the learning performance in a target domain. For that purpose, a local classifier is trained at each source domain by considering both local data and the prediction consensus with the classifiers from other source domains.

Table 3
Accuracy, recall, precision and F-measure of methods obtained from the state of the art and the developed DIAFAN-TL method.

Base	Transfer	Acc	Recall	Prec	F1
None	STRUT	74.69%	48.65%	45.89%	47.23%
	SER	72.22%	45.90%	44.54%	45.21%
	MIX	73.53%	47.48%	44.61%	46.00%
	TreesMixedEntropy	74.54%	48.10%	46.18%	47.12%
	NaiveBiasRegularizator	59.50%	28.21%	21.51%	24.41%
	NaivePruningTree	79.69%	47.39%	46.63%	47.01%
	NaiveRelabelingTree	78.13%	45.80%	45.45%	45.62%
LR	(Only Target reduced - no TL)	68.63%	43.82%	42.68%	43.24%
	(Only Target full data - no TL)	73.50%	55.62%	47.00%	50.95%
	(Source + Target reduced data - no TL)	75.18%	58.62%	49.19%	53.49%
	CORAL	69.77%	48.22%	43.96%	45.99%
	TCA	47.12%	26.19%	26.91%	26.54%
	ConsensusRegularization	67.74%	41.37%	40.77%	41.07%
	FEDA	57.23%	34.09%	34.17%	34.13%
	TrBagg	63.41%	39.76%	38.29%	39.01%
	TrAdaBoost	59.68%	41.16%	36.46%	38.67%
	KMM	70.36%	48.31%	43.77%	45.93%
	KRR	77.81%	48.39%	47.34%	47.86%
	DIAFAN-TL	**85.73%**	**69.82%**	**61.89%**	**65.62%**
RF	(Only Target reduced - no TL)	78.44%	60.78%	46.74%	52.84%
	(Only Target full data - no TL)	87.25%	62.09%	61.10%	61.59%
	(Source + Target reduced data - no TL)	88.51%	63.47%	64.58%	64.02%
	CORAL	83.12%	54.91%	54.75%	54.83%
	TCA	52.51%	26.88%	33.48%	29.88%
	ConsensusRegularization	81.44%	47.82%	50.78%	49.26%
	FEDA	85.22%	53.56%	56.41%	54.95%
	TrBagg	85.28%	53.89%	54.37%	54.13%
	TrAdaBoost	86.40%	58.17%	58.10%	58.13%
	KMM	86.24%	57.24%	56.88%	57.06%
	KRR	79.25%	46.96%	47.78%	47.36%
	DIAFAN-TL	**90.19%**	**67.76%**	**70.75%**	**69.22%**

- **FEDA** [47]. This algorithm spans the feature space to generate a more suitable feature space for source and target domains using three representations: source, target and shared. Then, the learning algorithm adapts using extra information.
- **TrBagg** [48]. First, the algorithm generates many weak classifiers from target and source data. Then, these classifiers are filtered using target data making and the selected classifiers are used in the bagging ensemble.
- **TrAdaBoost** [48]. This algorithm extends the AdaBoost to the transfer learning scenario. A new weighting mechanism is designed to reduce the impact of the distribution difference and is used in classical Adaboost joining the source and target data.
- **KRR** [49]. This algorithm combines Ridge regression and classification (linear least squares with l2-norm regularization) with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a non-linear function in the original space.

4. Results

4.1. Datasets description

The datasets used to test the proposed methodology were retrieved from three different sources.

First, index and the band information (10 bands) was collected from the Sentinel [50] satellite images. This satellite has been capturing images since 2015, increasing the sampling rate over the years. Now, an image is captured every 4 or 5 days (from 15 days, on average, in 2015). From these images, different indices and the values of color bands are calculated for each pixel of the image. In this work, the mean value of the pixels of each band are used for each image to train the model.

Second, the olive phenology states of each studied parcel were retrieved from the open dataset, property of 'Red de Alerta e Información Fitosanitaria' [51] belonging to 'Junta de Andalucía'. Olive cultivation consists of 14 phenological stages, according to *De Andrés scale* [52]. To ensure clarity and reliability in the results, these 14 stages are reduced to 4 as follows:

1. Stage 1. The first stage consists of the first two phenological stages, corresponding to the Winter Bud and Moved Bud stages.
2. Stage 2. The second stage includes all stages of flowering (stages 3–8): Inflorescence, corolla, flowering and petal fall.
3. Stage 3. The third stage consists of all stages of fruit set and the hardening of the fruit (9–10): fruit set and stone hardening
4. Stage 4. The fourth stage includes all stages of the fruit set (11–14).

These stages are in line with the official classification based on the BBCH scale [52]. Both scales converge to the 4 phases described above and are broadly grouped in the same way.

This phenology dataset was obtained from sixteen parcels for four regions of Andalusia, Spain (four for each one), as can be observed in Fig. 3, with different characteristics among them, such as variety, altitude, type of crop (traditional, intensive or super-intensive), etc. A more detailed view of parcels for each region is shown in Fig. 4.

The characteristics of these parcels are shown in Table 2. The main characteristics that represent each domain are the altitude (in meters), surface (in hectares), the slope (in %) and the density (trees per square meter). Fig. 5 shows a heatmap of the distances (in km) between the domains.

Table 4
Bayesian test for DIAFAN in the RF group.

left	right	p(left>right)	p(undetermined)	p(right>left)
NaiveBiasRegularizator	DIAFAN-RF	4.53E−08	4.94E−08	1.0000
TCA RF	DIAFAN-RF	2.46E−06	2.94E−06	1.0000
SER	DIAFAN-RF	0.0011	0.0019	0.9970
e ConsensusRegularization-RF	DIAFAN-RF	0.0012	0.0024	0.9964
Only target reduced - no TL -RF	DIAFAN-RF	0.0014	0.0022	0.9964
MIX	DIAFAN-RF	0.0024	0.0038	0.9938
NaiveRelabelingTree	DIAFAN-RF	0.0023	0.0040	0.9937
KRR-RF	DIAFAN-RF	0.0031	0.0039	0.9931
NaivePruningTree	DIAFAN-RF	0.0025	0.0046	0.9929
TreesMixedEntropy	DIAFAN-RF	0.0044	0.0060	0.9896
STRUT	DIAFAN-RF	0.0073	0.0098	0.9828
CORAL RF	DIAFAN-RF	0.0045	0.0147	0.9808
FEDA-RF	DIAFAN-RF	0.0075	0.0179	0.9746
TrBagg-RF	DIAFAN-RF	0.0158	0.0386	0.9456
TrAdaBoost-RF	DIAFAN-RF	0.0151	0.0506	0.9343
Only target full data - no TL -RF	DIAFAN-RF	0.0231	0.0740	0.9029
KMM RF	DIAFAN-RF	0.0428	0.0675	0.8897
Source + Target reduced data - no TL -RF	DIAFAN-RF	0.0125	0.1974	0.7902

The third data source consists of meteorological variables from public sources of the *Junta de Andalucia* public organization. These variables have a daily periodicity and are the average, maximum and minimum temperatures and humidity, precipitation and solar radiation.

4.2. Evaluation metrics

To quantify the effectiveness of the methodology proposed, accuracy, recall, precision and F-measure were computed. These metrics were obtained from the 4 × 4 confusion matrix derived from the phenology classification problem with four phenological stages.

Specifically, for each phenological state, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) were considered. For example, for phenological state 1, TP described the case where phenological state 1 was predicted and correct, while FP described the case where phenological state 1 was not predicted, but it still existed. A similar reasoning was done for the negative cases.

The first metrics were computed for all four phenology classes. We started from the 4×4 confusion matrix mentioned before and the accuracy was calculated as defined in Eq. (2).

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

Subsequently, the metrics of precision, recall and F-1 were calculated for each phenological stage, and the weighted mean was also calculated.

Specifically, precision was the ratio of correctly predicted positive observations to the total predicted positive values, as defined in Eq. (3).

$$Prec = \frac{TP}{TP + FP} \tag{3}$$

Recall was the ratio of correctly predicted positive observations with respect to all actual positive instances, as defined in Eq. (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The F1 score was the weighted average of precision and recall, as defined in Eq. (5).

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{5}$$

Furthermore, all those metrics were also computed for each class (phenological state) independently. Each TP, TN, FN, FP referred to each phenological stage versus the rest.

4.3. Baseline comparative results

Table 3 compares all the state-of-the-art methods obtained with the method developed in this work. The table is divided into three main parts. The first part is composed of TL algorithms that do not need a base algorithm to run. The second part includes non-TL based algorithms and TL based algorithms that are based on a logistic regression (LR) algorithm. Finally, the last part includes those based on a random forest (RF) algorithm. In each column we observe the four evaluation metrics described in Section 4.2.

Among the methods included in the first section, the *NaivePruningTree* and the *NaiveRelabelingTree* stand out with the best results with a 79.69% and a 78.13% accuracy, respectively. However, as we have an imbalanced problem and it is preferable to consider a metric that considers the imbalance, the F1 metric has been added. Regarding the F1 metric, the best algorithms are *STRUT* and *NaivePruningTree*, with 47.23% and 47.01%, respectively.

In the group of algorithms that do not use TL, the one that uses the source domain plus the complete target yields better results using either of the two base algorithms described above with a 75.18% accuracy and an F1 metric of 53.49% in the case of logistic regression as the base. In the case of the random forest, the accuracy is 88.51%, and the F1 is 64.02%. Both logistic regression and random forest cases improve the algorithms that do not use bases in the best cases of each metric in comparing nonbase and base with a difference of 16.79% of F1 and 8.82% of accuracy.

Focusing just on the TL algorithms, *CORAL* is the best with the logistic regression base with an accuracy of 69.77% and an F1 of 46.00%. Using the random forest as a base yields better results using the TrAdaboost algorithm with an 86.40% accuracy and a 58.13% F1 score. However, this method does not surpass the results of the $Source + Targetreduceddata - noTL$ with any of the bases.

The developed DIAFAN-TL method improves all the previous methods. Compared with $Source+Targetreduceddata-noTL$, which is the best method from the baseline there is an improvement in call the metrics with an accuracy difference of 3.79%, 5.2% in F1, 3.99% in Recall and 6.17% in Precision. Moreover, in the logistic regression base, DIAFAN-TL also presents the best results with an even greater improvement, showing the robustness of this method regarding the base. The best methods in general are better with a random forest base, presumably due to the large space of attributes that hinder the convergence of the logistic regression.

In order to compare DIAFAN-TL with other methods of similar nature, Kernel Mean Matching (KMM) and Kernel Ridge Regression (KRR) metrics have been considered (TrAdaBoost also uses instance weighting but is iterative during the model training process).

For KMM, the equation to optimize is:

$$min_w \frac{1}{2} w^T K w - k^T w \quad subject\ to \quad w_i \epsilon\, [0, B]\, and \left| \sum_{i=1}^{n_s} w_i - n_s \right| \leqslant m_\epsilon \tag{6}$$

Where:

- $K_{ij} = k(x_i, x_j)\ with\ x_i, x_j \epsilon X_S\ and\ k\ a\ kernel.$
- $k_i = \frac{n_S}{n_T} \sum_{x_j \epsilon x_T} k(x_i, x_j)\ with\ x_i\ \epsilon\ X_j.$
- w_i are the source instance weights.
- X_S, X_T are the input source and target dataset, respectively.
- B, ϵ are two KMM hyperparameters.

Furthermore, the Kernel Ridge Regression (KRR) metric, based on the same characteristics as our DIAFAN metric, has been considered for comparative purposes. This method uses the labeled data for the source and target data. First, a kernel ridge regression a on just source data. Then, the weights α are calculated using the vector a optimizing a cost function. The weight function takes the form:

$$\hat{w}^\alpha(x, y) = \sum_{l=1}^{N} \alpha_l \exp\left(- \frac{\left\| (x, y) - \left(x'_l, y'_l\right) \right\|^2}{2\eta^2} \right), \tag{7}$$

where:

- N represents the number of source instances.
- (x, y) represents the features and labels for the instance to be weighted.
- x'_l, y'_l represents the centerpoint of each source instance.
- α_l represents the values obtained from kernel ridge regression.

And the equation to optimize is:

$$\min_{\alpha \geq 0} \sum_{i=1}^{M} \left(y_i^P - a^t \hat{W}^\alpha \left(x_i^P, y_i^P\right) \underline{k} \left(x_i^P\right) \right)^2 + \gamma \left\|\alpha\right\|^2, \tag{8}$$

where:

- M represents the number of target instances.
- y_i^P represents the labels for the target instance i.
- a^t is the coefficients for the linear combination in the feature space in target data using the ridge regression dualization.
- $\underline{k}\left(x_i^P\right)$ represents the kernel applied to the target instance x_i^P.
- γ represents the regularization factor for the norm of alphas.
- $\hat{W}^\alpha$ represents the weight function previously presented.

4.4. Statistical analysis

In order to study the goodness of fit of the results obtained, the results are compared using the Bayesian analysis approach [53]. Similarly to Wilcoxon signed-rank test [54], this approach compares the results of each pair of methods with the same dataset and estimates if the methods tested belong to the same distribution. The input of the Bayesian analysis for each method uses the F1-score for each plot using the test set. The output of this method is not the usual p-value due to several issues [53], instead, the analysis obtains the probability that a method A is better than a method B, method B is better than method A and both methods belong to the same distribution. To determine when the methods are considered from the same distribution, a threshold must be established. The threshold determines the F1-Score percentage range in which the results are considered similar, in all the experiments a threshold selected determines than a difference of 1% in F1-Scores are considered similar.

Similarly to the original work in Bayesian analysis, the methods have been structured as left for all methods compared with DIAFAN in right. For the output of the Bayesian analysis, the probabilities have been divided as p(left>right), p(right>left) and p(undetermined) representing that the method in the left is better than right, the method in the right is better than left and the probability of similar distribution respectively.

The experiments and training and test sets used are the same as those of Table 3. For this purpose, two groups of methods have been made: those based on RF (see Table 4) and those based on LR (see Table 5). DIAFAN is compared with all the algorithms in its group and with those that do not depend on the base.

On the one hand, for the LR-based group or none, the range of significance or precision ranges from P> 1.0000 obtained against *TrAdaBoost*, *TCA*, *NaiveBiasRegularizator* and *FEDA* to P>0.9255 obtained for the $Source + Target\ reduced\ data\ -\ no\ TL$ method. On the other hand, for the RF-based group or none, the range is between P>0.7902 for the $Source + Target\ reduced\ data\ -\ no\ TL$ method and P> 1.000 for the *NaiveBiasRegularizator* and *TCA* methods. Hence, it is shown that DIAFAN is statistically better in all cases when evaluating the F1-Score, especially with the LR-based algorithms.

4.5. Phenological states analysis

Table 6 shows the F-measure for each of the domains studied for each of the phenological stages. Each column, represented from 1 to 4, is each of the phenological stages described in Section 4.1. Additionally, for each phenological stage, the number of samples it contains is shown.

It is clear that phenological stages 1 and 4 are much less represented. In fact, not in all domains are the samples with these stages. For that reason, we have an imbalanced problem. Because of this, there is a higher F measure in phenological stages 2 and 3, with averages of 90.30% and 94.08%, respectively. The lowest percentages correspond to those domains with low representativeness of these stages. Additionally, phenological state 2 is slightly worse than state 3 even when it has half the number of samples.

4.6. Forecasting horizon analysis

Table 7 shows the F-measure metric for each of the domains studied and the horizons to be predicted. Each column, represented from H1 to H4, is each one of the horizons. The last column indicates the weighted average of the four horizons.

Regarding the achieved average results, the best performing horizon is H2, and the worst is H4. H1 is slightly below the results of H2, possibly overfitting the model inputs from previous weeks. In general, the first two horizons perform better than the last two presumably because the nearest weeks are easier to predict than further weeks. Additionally, obtaining better results for H2 can be beneficial, as it can help to see a better future and plan with more time. By domain, $C3\_NuevaCarteya$ has the lowest average hit, followed by the $C4\_PozoAlcon4$, $C1\_Olvera2$, $C4\_PozoAlcon1$, $C1\_Olvera1$, $C4\_PozoAlcon3$ and $C4\_PozoAlcon2$ domains. The domain with the best results is $C2\_ElSaucejo1$. It is possible that the characteristics of the plots themselves influence these results and will be studied in future works.

Table 5
Bayesian test for DIAFAN LR.

left	right	p(left>right)	p(undetermined)	p(right>left)
NaiveBiasRegularizator	DIAFAN-LR	9.84E−08	1.25E−07	1.0000
TCA LR	DIAFAN-LR	1.65E−06	2.37E−06	1.0000
TrAdaBoost-LR	DIAFAN-LR	0.0000	0.0000	1.0000
FEDA-LR	DIAFAN-LR	0.0000	0.0000	1.0000
Only target reduced - no TL -LR	DIAFAN-LR	0.0001	0.0001	0.9998
TrBagg-LR	DIAFAN-LR	0.0001	0.0001	0.9998
ConsensusRegularization-LR	DIAFAN-LR	0.0003	0.0006	0.9991
KMM LR	DIAFAN-LR	0.0004	0.0012	0.9985
Only target full data - no TL -LR	DIAFAN-LR	0.0012	0.0050	0.9938
CORAL LR	DIAFAN-LR	0.0024	0.0055	0.9921
SER	DIAFAN-LR	0.0056	0.0105	0.9839
KRR LR	DIAFAN-LR	0.0060	0.0146	0.9794
MIX	DIAFAN-LR	0.0105	0.0178	0.9717
NaiveRelabelingTree	DIAFAN-LR	0.0119	0.0210	0.9671
NaivePruningTree	DIAFAN-LR	0.0120	0.0236	0.9644
TreesMixedEntropy	DIAFAN-LR	0.0145	0.0219	0.9636
STRUT	DIAFAN-LR	0.0316	0.0395	0.9289
Source + Target reduced data - no TL -LR	DIAFAN-LR	0.0192	0.0553	0.9255

Table 6
Experiment with reduced data with random Sources and Target: F-Measure by phenology state with the best base algorithm.

Target	Phenological state			
	1	2	3	4
Number of samples	36	464	1042	58
P01		91.72%	91.53%	38.24%
P02		90.88%	91.15%	32.06%
P03		92.56%	95.79%	
P04		89.98%	95.65%	
P05	75.71%	93.39%	90.11%	
P06		95.78%	92.74%	67.61%
P07		94.80%	94.39%	
P08		95.78%	94.37%	
P09		88.65%	90.11%	
P10		86.27%	95.58%	
P11	7.69%	81.93%	91.00%	
P12	46.97%	78.88%	92.96%	86.88%
P13		86.43%	97.03%	
P14		87.19%	96.63%	
P15		89.97%	96.89%	
P16	87.72%	89.30%	96.56%	
Total	**39.36%**	**90.25%**	**94.03%**	**49.68%**

Table 7
Experiment with reduced data with random Sources and Target: F-Measure by horizon with the best base algorithm.

Target	Horizon				Average
	H1	H2	H3	H4	
P01	68.90%	83.22%	82.06%	78.07%	70.23%
P02	61.28%	80.99%	80.83%	78.63%	69.12%
P03	61.78%	61.89%	78.79%	75.25%	62.84%
P04	74.54%	77.87%	75.57%	72.06%	62.20%
P05	78.92%	76.02%	64.71%	62.94%	74.71%
P06	82.30%	92.60%	87.87%	84.47%	87.50%
P07	64.08%	79.27%	75.38%	73.30%	63.47%
P08	77.59%	79.19%	75.30%	73.38%	76.33%
P09	74.04%	93.30%	89.17%	87.27%	72.42%
P10	71.74%	61.92%	73.82%	93.07%	60.87%
P11	73.25%	65.01%	71.57%	93.64%	72.82%
P12	84.26%	83.09%	74.75%	66.05%	81.32%
P13	70.07%	75.54%	63.94%	94.41%	62.23%
P14	60.50%	75.64%	75.38%	74.75%	62.29%
P15	61.59%	93.99%	74.47%	77.68%	62.63%
P16	92.82%	73.85%	75.83%	76.41%	78.55%
Total	**69.40%**	**73.79%**	**64.22%**	**57.50%**	**69.22%**

4.7. Source domains analysis

Table 8 shows the results of the reduced data experiment with a single source domain and the target domain. Each row represents the target domains and the columns represent the source domains. Domains 11 and 12 have the lowest percentages, indicating that they are not suitable for use as source domains. Additionally, there are sources more suitable for some targets than other ones, indicating that weighting could be beneficial for transferring knowledge instead of considering all the sources equally to increase the noise.

4.8. Full data analysis

Table 9 shows the F-measure of the TL experiment with full data. The structure of the table is similar to that of Table 7.

It should be noted that the average percentage per horizon and overall is better than that obtained in the reduced data case, which amplifies the benefits of the proposed methodology. Using weighting source domains is the best way to transfer the knowledge.

In addition, comparative plots are shown (Figs. 6 and 7) between the recall and precision results of the experiments with random source domains (TL reduced) and with all the complete domains (TL). Their weighted average is shown. There is a slight improvement in the full data case due to the availability of more data in the majority of plots.

5. Conclusions

The proposed method shows an important improvement compared with other transfer learning methods. The results show that not all transfer learning algorithms with insufficient available data are beneficial for adding useful information. However, the proposed model shows not only an improvement compared with other transfer learning methods but also obtains better results than the algorithms that do not use any transfer learning method, since it includes the only transfer learning tested that improves the no transfer learning algorithms. This result seems to be independent of the base algorithm used for the evaluation, as DIAFAN-TL obtains the best results in the logistic regression and random forest is the last one the best bases. The results support the idea that the effectiveness of using the weighting function (DIAFAN) that prioritizes diversity and avoids false neighbors of the data is beneficial for the algorithms in this application.

Table 8
Experiment with reduced data with one Source domain and Target: F-Measure by horizon with the best base algorithm.

TG	Source															
	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16
P01		72.01%	63.29%	62.26%	69.90%	91.49%	63.01%	76.03%	74.13%	74.37%	67.99%	87.82%	70.94%	64.39%	62.37%	92.02%
P02	68.77%		63.29%	62.26%	76.30%	92.14%	62.63%	76.24%	74.13%	74.37%	69.68%	84.99%	72.15%	75.24%	75.24%	79.32%
P03	72.01%	72.01%		62.24%	62.92%	87.78%	76.48%	75.48%	72.90%	73.15%	67.38%	76.07%	73.53%	75.17%	76.67%	81.90%
P04	70.99%	72.01%	62.15%		65.72%	92.20%	76.28%	75.48%	72.90%	73.15%	67.97%	76.28%	92.18%	94.62%	94.62%	90.48%
P05	60.09%	70.41%	62.61%	62.03%		89.79%	75.89%	75.48%	72.90%	73.15%	68.24%	80.91%	68.26%	75.59%	73.81%	77.71%
P06	72.01%	72.01%	62.84%	61.45%	81.93%		75.27%	76.45%	75.39%	74.37%	67.49%	79.22%	89.04%	72.76%	74.79%	77.93%
P07	70.41%	70.41%	62.08%	62.85%	73.41%	88.48%		75.89%	72.90%	74.37%	69.17%	80.23%	71.53%	75.59%	76.92%	83.28%
P08	68.77%	72.01%	62.84%	62.44%	82.62%	90.64%	75.69%		72.90%	74.37%	68.62%	81.43%	69.18%	76.13%	75.67%	77.02%
P09	80.10%	80.10%	62.61%	61.45%	72.17%	88.48%	74.75%	75.48%		74.37%	71.53%	85.28%	72.46%	92.75%	92.75%	72.09%
P10	70.99%	80.10%	62.78%	63.05%	67.97%	87.59%	76.09%	76.86%	74.13%		69.24%	83.47%	67.82%	92.83%	90.86%	87.62%
P11	80.10%	77.99%	63.60%	62.46%	70.21%	86.94%	76.09%	76.45%	72.90%	74.37%		71.93%	75.82%	94.62%	94.62%	90.48%
P12	72.01%	72.01%	62.61%	61.65%	63.87%	88.31%	76.28%	76.45%	72.90%	73.15%	70.11%		74.68%	94.64%	75.93%	78.87%
P13	62.14%	62.57%	83.83%	84.11%	83.98%	80.80%	93.34%	94.68%	81.48%	85.41%	79.70%	84.01%		75.00%	85.35%	85.46%
P14	57.85%	59.36%	62.84%	61.29%	75.50%	87.03%	75.18%	75.96%	71.70%	74.78%	66.04%	77.16%	60.89%		74.83%	67.93%
P15	59.94%	58.66%	64.04%	61.14%	69.54%	85.84%	75.87%	75.45%	69.36%	74.78%	67.16%	72.87%	66.95%	73.33%		67.81%
P16	58.66%	50.56%	63.22%	61.69%	70.27%	87.96%	75.45%	75.56%	68.23%	72.47%	66.60%	68.72%	68.41%	75.17%	57.76%	
Avg	**68.92%**	**69.17%**	**62.88%**	**62.03%**	**70.38%**	**88.59%**	**63.23%**	**75.99%**	**72.90%**	**73.85%**	**68.18%**	**78.62%**	**60.02%**	**62.50%**	**62.06%**	**74.40%**

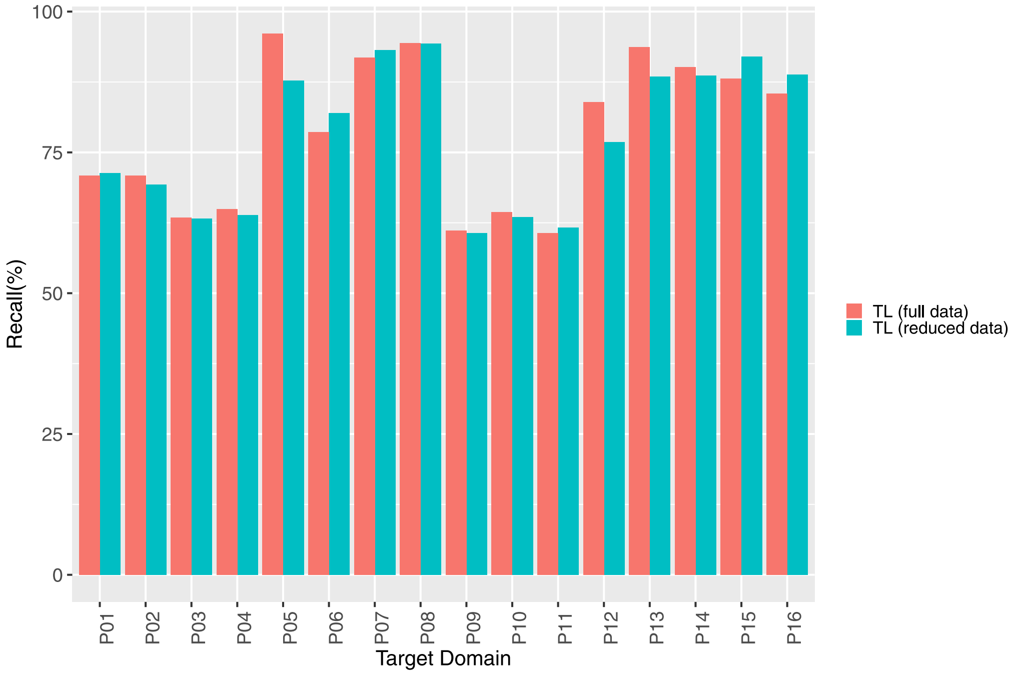


Fig. 6. Average recall values for each target domain considering full and reduced data scenarios.

These studies leave open the possibility for further research in this field. According the results in Table 8, a possible improvement of the method could be based on a specific filtering or selection of domains with certain sensitivity thresholds that improve the obtained results.

CRediT authorship contribution statement

M.A. Molina-Cabanillas: Conceptualization, Methodology, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **M.J. Jiménez-Navarro:** Conceptualization, Methodology, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing. **R. Arjona:** Investigation, Writing – review & editing. **F. Martínez-Álvarez:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision. **G. Asencio-Cortés:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Spanish Ministry of Science, Innovation and Universities for the support under projects TIN2017-8888209C2-1-R and PID2020-11795RB-C21. Funding for open access publishing: Universidad Pablo de Olavide/CBUA.

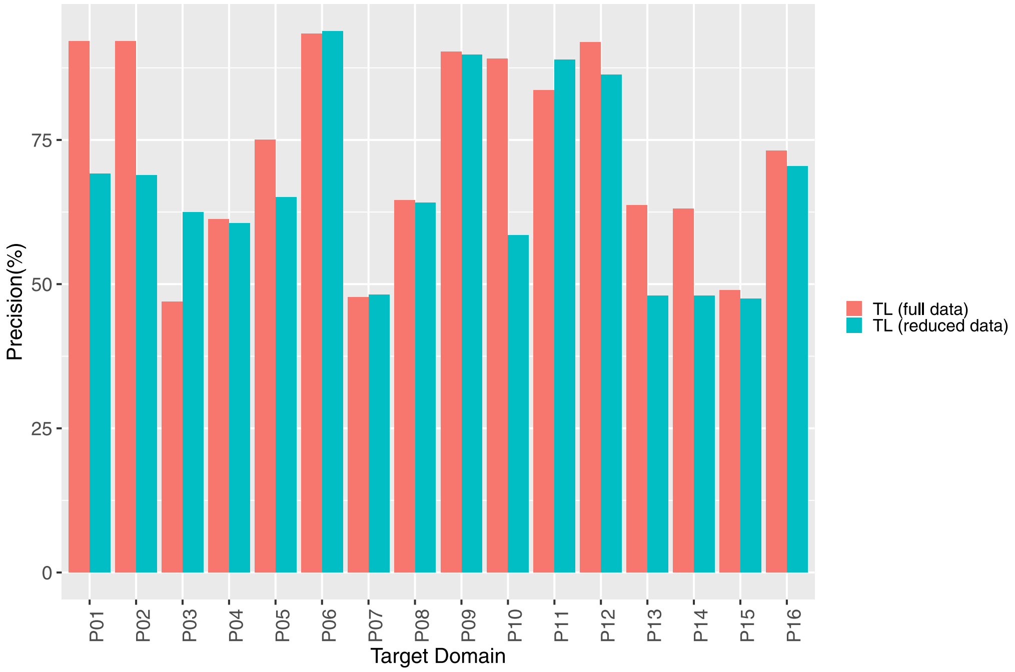


Fig. 7. Average precision values for each target domain considering full and reduced data scenarios.

Table 9
Experiment with full data: F-Measure by horizon with the best base algorithm.

Target	Horizon				Average
	H1	H2	H3	H4	
P01	78.26%	75.07%	83.32%	78.13%	80.10%
P02	78.26%	75.07%	83.32%	78.13%	80.10%
P03	76.92%	59.99%	80.00%	74.93%	53.95%
P04	76.60%	78.16%	95.95%	72.35%	63.05%
P05	85.10%	84.53%	84.00%	77.06%	84.26%
P06	75.40%	86.13%	86.78%	85.95%	85.39%
P07	62.76%	66.10%	75.30%	72.29%	62.83%
P08	96.47%	79.19%	75.30%	75.15%	76.65%
P09	72.43%	93.30%	93.30%	87.27%	72.90%
P10	76.19%	75.68%	72.92%	93.64%	74.78%
P11	69.93%	64.02%	72.92%	93.64%	70.33%
P12	87.69%	88.88%	87.04%	78.93%	87.79%
P13	65.97%	100.00%	100.00%	100.00%	75.81%
P14	94.72%	93.40%	92.10%	71.83%	74.23%
P15	77.19%	93.40%	92.10%	71.83%	62.90%
P16	100.00%	77.61%	94.36%	74.37%	78.82%
Total	**81.23%**	**82.10%**	**86.62%**	**81.36%**	**72.94%**

References

[1] S.J. Thackeray, P.A. Henrys, D. Hemming, J.R. Bell, M.S. Botham, S. Burthe, P. Helaouet, D.G. Johns, I.D. Jones, D.I. Leech, et al., Phenological sensitivity to climate across taxa and trophic levels, Nature 535 (7611) (2016) 241–245.

[2] T. Sakamoto, B.D. Wardlow, A.A. Gitelson, S.B. Verma, A.E. Suyker, T.J. Arkebauer, A two-step filtering approach for detecting maize and soybean phenology with time-series MODIS data, Remote Sens. Environ. 114 (10) (2010) 2146–2159.

[3] N. Segev, M. Harel, S. Mannor, K. Crammer, R. El-Yaniv, Learn on source, refine on target: A model transfer learning framework with random forests, IEEE Trans. Pattern Anal. Mach. Intell. 39 (9) (2017) 1811–1824, http://dx.doi.org/10.1109/TPAMI.2016.2618118.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: Proceedings of the International Conference on Artificial Neural Networks, 2018, pp. 270–279.

[5] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[6] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, New Advances in Deep-Transfer Learning, Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC., 2013.

[7] G. Vrbančič, V. Podgorelec, Transfer learning with adaptive fine-tuning, IEEE Access 8 (2020) 196197–196211, http://dx.doi.org/10.1109/ACCESS.2020.3034343.

[8] D. Haase, E. Rodner, J. Denzler, Instance-weighted transfer learning of active appearance models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[9] J. Garcke, T. Vanck, Importance weighted inductive transfer learning for regression, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 466–481.

[10] A.M. Azab, L. Mihaylova, K.K. Ang, M. Arvaneh, Weighted transfer learning for improving motor imagery-based brain–computer interface, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (7) (2019) 1352–1359.

[11] O. Day, T.M. Khoshgoftaar, A survey on heterogeneous transfer learning, J. Big Data 4 (1) (2017) 29.

[12] Y. Liu, C. Yang, M. Zhang, Y. Dai, Y. Yao, Development of adversarial transfer learning soft sensor for multigrade processes, Ind. Eng. Chem. Res. 59 (37) (2020) 16330–16345.

[13] Y. Liu, C. Yang, K. Liu, B. Chen, Y. Yao, Domain adaptation transfer learning soft sensor for product quality prediction, Chemometr. Intell. Lab. Syst. 192 (2019) 103813.

[14] A. Abanda, U. Mori, J.A. Lozano, A review on distance based time series classification, Data Min. Knowl. Discov. 33 (2) (2019) 378–412.

[15] K.G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, Machine learning in agriculture: A review, Sensors 18 (8) (2018) 2674.

[16] J. Chen, J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu, et al., Global land cover mapping at 30 m resolution: A POK-based operational approach, ISPRS J. Photogramm. Remote Sens. 103 (2015) 7–27.

[17] P. Gong, J. Wang, L. Yu, Y. Zhao, Y. Zhao, L. Liang, Z. Niu, X. Huang, H. Fu, S. Liu, et al., Finer resolution observation and monitoring of global land cover: First mapping results with landsat TM and ETM+ data, Int. J. Remote Sens. 34 (7) (2013) 2607–2654.

[18] A.J. Oliphant, P.S. Thenkabail, P. Teluguntla, J. Xiong, R.G. Congalton, K. Yadav, R. Massey, M.K. Gumma, C. Smith, NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Southeast Asia 30 m V001, NASA EOSDIS Land Processes DAAC, 2017, pp. 1–12.

[19] L. Yu, J. Wang, N. Clinton, Q. Xin, L. Zhong, Y. Chen, P. Gong, FROM-GC: 30 m global cropland extent derived through multisource data integration, Int. J. Digit. Earth 6 (6) (2013) 521–533.

[20] L. Yu, J. Wang, X. Li, C. Li, Y. Zhao, P. Gong, A multi-resolution global land cover dataset through multisource data aggregation, Sci. China Earth Sci. 57 (10) (2014) 2317–2329.

[21] S. Skakun, B. Franch, E. Vermote, J.C. Roger, I. Becker-Reshef, C. Justice, N. Kussul, Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model, Remote Sens. Environ. 195 (2017) 244–258.

[22] P. Hao, Y. Zhan, L. Wang, Z. Niu, M. Shakir, Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA, Remote Sens. 7 (5) (2015) 5347–5369.

[23] Y. Wang, Z. Xue, J. Chen, G. Chen, Spatio-temporal analysis of phenology in yangtze River Delta based on MODIS NDVI time series from 2001 to 2015, Front. Earth Sci. 13 (1) (2019) 92–110.

[24] Z. Xue, P. Du, L. Feng, Phenology-driven land cover classification and trend analysis based on long-term remote sensing image series, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (4) (2014) 1142–1156.

[25] J. Chen, J. Chen, D. Zhang, Y. Sun, Y.A. Nanehkaran, Using deep transfer learning for image-based plant disease identification, Comput. Electron. Agric. 173 (2020) 105393.

[26] Z. Wang, Z. Dai, B. Póczos, J. Carbonell, Characterizing and avoiding negative transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11293–11302.

[27] Q. Yang, L. Shi, J. Han, J. Yu, K. Huang, A near real-time deep learning approach for detecting rice phenology based on UAV images, Agricult. Forest Meteorol. 287 (2020) 107938.

[28] P. Hao, L. Di, C. Zhang, L. Guo, Transfer learning for crop classification with cropland data layer data (CDL) as training samples, Sci. Total Environ. 733 (2020) 138869.

[29] H. Yalcin, Phenology recognition using deep learning, in: Proceedings of the Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, 2018, pp. 1–5.

[30] M. Grünig, E. Razavi, P. Calanca, D. Mazzi, J.D. Wegner, L. Pellissier, Applying deep neural networks to predict incidence and phenology of plant pests and diseases, Emerg. Technol. 12 (2021) e03791.

[31] L. Zhong, P. Gong, G.S. Biging, Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery, Remote Sens. Environ. 140 (2014) 1–13.

[32] P. Hao, L. Wang, Y. Zhan, Z. Niu, Using moderate-resolution temporal NDVI profiles for high-resolution crop mapping in years of absent ground reference data: a case study of bole and manas counties in xinjiang, China, ISPRS Int. J. Geo-Inf. 5 (5) (2016) 67.

[33] A.X. Wang, C. Tran, N. Desai, D. Lobell, S. Ermon, Deep transfer learning for crop yield prediction with remote sensing data, in: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, 2018, pp. 1–5.

[34] Z. Deng, Y. Jiang, F.-L. Chung, H. Ishibuchi, S. Wang, Knowledge-leverage-based fuzzy system and its modeling, IEEE Trans. Fuzzy Syst. 21 (4) (2013) 597–609.

[35] L. Xie, Z. Deng, P. Xu, K.-S. Choi, S. Wang, Generalized hidden-mapping transductive transfer learning for recognition of epileptic electroencephalogram signals, IEEE Trans. Cybern. 49 (6) (2019) 2200–2214.

[36] P. Xu, Z. Deng, J. Wang, Q. Zhang, K.-S. Choi, S. Wang, Transfer representation learning with TSK fuzzy system, IEEE Trans. Fuzzy Syst. 29 (3) (2021) 649–663.

[37] Z. Deng, Y. Jiang, F.-L. Chung, H. Ishibuchi, K.-S. Choi, S. Wang, Transfer prototype-based fuzzy clustering, IEEE Trans. Fuzzy Syst. 24 (5) (2016) 1210–1232.

[38] Z. Deng, J. Lu, D. Wu, K.-S. Choi, S. Sun, Y. Nojima, Guest editorial: Special issue on new advances in deep-transfer learning, IEEE Trans. Emerg. Top. Comput. Intell. 3 (5) (2019) 357–359, http://dx.doi.org/10.1109/TETCI.2019.2936641.

[39] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (1) (2020) 43–76.

[40] R. Hadsell, S. Chopra, Y. Lecun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1735–1742.

[41] N. Segev, M. Harel, S. Mannor, K. Crammer, R. El-Yaniv, Learn on source, refine on target: A model transfer learning framework with random forests, IEEE Trans. Pattern Anal. Mach. Intell. 39 (9) (2017).

[42] N. Segev, R. El-Yaniv, Transfer Learning Using Decision Forests (Ph.D. thesis), Computer Science Department, Technion, 2016.

[43] B. Sun, J. Feng, K. Saenko, Correlation alignment for unsupervised domain adaptation, in: Domain Adaptation in Computer Vision Applications, Springer International Publishing, 2017, pp. 153–171.

[44] J. Huang, A.J. Smola, A. Grettonn, K.M. Borgwardt, B. Schölkopf, Correcting sample selection bias by unlabeled data, in: Proceedings of the Advances in Neural Information Processing Systems, 2007, pp. 601–608.

[45] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Netw. 22 (2) (2011) 199–210.

[46] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, Q. He, Transfer learning from multiple source domains via consensus regularization, in: Proceedings of the International Conference on Information and Knowledge Management, 2008, pp. 103–112.

[47] I. Daumé, Frustratingly easy domain adaptation, in: Proceedings of the Annual Meeting of the Association of Computational Linguistics, 2009, pp. 256–263.

[48] T. Kamishima, M. Hamasaki, S. Akaho, TrBagg: A simple transfer learning method and its application to personalization in collaborative tagging, in: Proceedings of the International Conference on Data Mining, 2009, pp. 219–228.

[49] K. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012, pp. 492–493.

[50] D. Geudtner, R. Torres, P. Snoeij, M. Davidson, B. Rommen, Sentinel-1 system capabilities and applications, in: Proceedings of the IEEE Geoscience and Remote Sensing Symposium, IEEE, 2014, pp. 1457–1460.

[51] Junta de Andalucia, RAIF website of the consejeria de agricultura, pesca y desarrollo rural, 2020, Online https://www.juntadeandalucia.es/agriculturapescaydesarrollorural/raif. (Accessed 26 March 2020).

[52] F. Sanz-Cortés, J. Martinez-Calvo, M.L. Badenes, H. Bleiholder, H. Hack, G. Llácer, U. Meier, Phenological growth stages of olive trees (*Olea europaea*), Ann. Appl. Biol. 140 (2) (2002) 151–157.

[53] A. Benavoli, G. Corani, J. Demsar, M. Zaffalon, Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis, J. Mach. Learn. Res. 18 (2017) 2653–2688.

[54] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83.