

Article

Field-Level Classification of Winter Catch Crops Using Sentinel-2 Time Series: Model Comparison and Transferability

Kato Vanpoucke ^{1,2,*}, Stien Heremans ^{1,2} , Emily Buls ³ and Ben Somers ^{2,4} 

¹ Research Institute for Nature and Forest (INBO), Havenlaan 88, 1000 Brussels, Belgium

² Division of Forest, Nature and Landscape (FNL), Department of Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E, 3001 Leuven, Belgium

³ Vlaamse Landmaatschappij (VLM), Koning Albert II-laan 15, 1210 Brussels, Belgium

⁴ KU Leuven Plant Institute (LPI), Kasteelpark Arenberg 31, 3001 Leuven, Belgium

* Correspondence: kato.vanpoucke@inbo.be

Abstract: Winter catch crops are promoted in the European Union under the Common Agricultural Policy to improve soil health and reduce nitrate leaching from agricultural fields. Currently, Member States often monitor farmers' adoption through on-site inspections for a limited subset of parcels. Because of its potential for region-wide coverage, this study investigates the potential of Sentinel-2 satellite time series to classify catch crops at the field level in Flanders (Belgium). The first objective was to classify catch crops and identify the optimal model and time series input for this task. The second objective was to apply these findings in a real-world scenario, aiming to provide reliable early-season predictions in a separate target year, testing early-season performance and temporal transferability. The following three models were compared: Random Forest (RF), Time Series Forest (TSF), and a One-Dimensional Convolutional Neural Network (1D-CNN). The results showed that, with a limited field-based training dataset, RF produced the most robust results across different time series inputs, achieving a median F1-score of >88% on the best dataset. Additionally, the early-season performance of the models was delayed in the target year, reaching the F1-score threshold of 85% at least one month later in the season compared to the training years, with large timing differences between the models.

Keywords: catch crop; early classification; temporal transferability; machine learning; satellite image time series; Sentinel-2



Citation: Vanpoucke, K.; Heremans, S.; Buls, E.; Somers, B. Field-Level Classification of Winter Catch Crops Using Sentinel-2 Time Series: Model Comparison and Transferability.

Remote Sens. **2024**, *16*, 4620.

<https://doi.org/10.3390/rs16244620>

Academic Editors: David M Johnson and Annamaria Castrignanò

Received: 14 October 2024

Revised: 22 November 2024

Accepted: 8 December 2024

Published: 10 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

When mineral nitrogen remains in the soil after harvest of the main crop, nitrate leaching can lead to elevated nitrate levels in surface and groundwater, adversely affecting the environment and human health [1]. For this reason, the Nitrates Directive in the European Union aims to reduce water pollution from agricultural sources (91/676/EEC) and thereby requires Member States to monitor their water quality and implement good agricultural practices. An example of such practice is the cultivation of catch crops, which are also supported by the Common Agricultural Policy (CAP) in the context of soil health and are often referred to as cover crops in this context [2]. Catch crops are grown from late summer to early spring, bridging the gap between main cropping seasons [3], with the goal of reducing nitrate leaching. Replacing fallow with catch crops reduces nitrate leaching through multiple mechanisms, including active uptake of both nitrogen and water [4,5]. Their effectiveness is influenced by factors such as (i) biomass, (ii) species, and (iii) sowing and termination dates, with a recent meta-analysis showing an average reduction in nitrate leaching by 68% [4,5].

Monitoring the adoption of catch crops by farmers currently relies on on-site field visits by authorities [3]. Field visits are, however, very labor- and time-consuming, and therefore, only a small fraction of all fields can be visited. To get a more complete overview of the

extent to which catch crops are adopted across a region, satellite remote sensing can be used to map catch crop presence and prioritize field visits to areas of higher uncertainty [3]. Satellite remote sensing offers large-scale coverage and high spatio-temporal resolution and has already been successfully used in crop monitoring and policy support for the CAP [6].

Catch crop monitoring with remote sensing has not yet been researched extensively, but the field has recently been growing due to the mainstreaming of these cropping practices combined with the recent rise in the availability of satellite data with sufficiently high spatial, spectral, and temporal resolution [7]. As outlined by [8], past studies have focused on three main aspects of catch crops: (a) their presence [3], (b) their emergence and termination date [9], and (c) their biomass [8,10,11]. These studies applied a wide range of methods, including phenology through optical satellite time series, to monitor catch crops [3,12,13]. The use of phenological profiles can help differentiate catch crops from other crops and vegetation because the spectral differences between vegetation types may be small at any single moment in time, while differences in timing may be more pronounced [14]. Especially for crop type classification, the different growth patterns and differences in the timing of the growth stages can be used as discriminating features [15,16]. Distinguishing between catch crop types can help estimate their effect on N-leaching, as different types have different nitrogen uptake capacities. Both the type of catch crop and its biomass accumulation influence the amount of nitrogen retained, making catch crop type an important variable to consider when modeling N-uptake by catch crops across a region [11].

Simultaneously, with the increased focus on time series data, data-driven machine learning models have become prominent in (catch) crop monitoring [3,17,18]. For example, for the classification of catch crops, ref. [3] used satellite time series and Random Forest (RF), a model that is also popular in main crop classification [18,19]. Some advantages of RFs are their easy-to-tune parameters and their robustness to overfitting [17,20,21]. However, RFs do not explicitly take into account the temporal dependencies between the observations, as they handle each time step as an independent variable. This drawback has been overcome by the introduction of a new branch of RFs specifically designed for time series classification and prediction. Time Series Forest (TSF) is an example of such an RF where temporal features—mean, standard deviation, and slope—are calculated over predefined (randomly chosen) temporal intervals, and these deduced features are subsequently used for classification [22]. The model, therefore, incorporates crop phenology by explicitly using the temporal dimension.

Over the last few years, deep learning algorithms have also appeared in crop monitoring research [15,23]. Different kinds of deep learning architectures have been compared for the task of crop classification, such as RNNs, MS-ResNet, and TempCNN [23]. Notably, TempCNN was introduced by [17], exploring the use of One-Dimensional CNNs (1D-CNN) for time series-based crop type classification, a lightweight model compared to other deep learning architectures. The use of deep neural networks comes with two major challenges: (1) the expertise required for optimizing both the model architecture and the hyperparameter values, shifting the difficulty from feature engineering to architecture engineering, and (2) the large number of parameters that have to be estimated, which makes deep learning methods prone to overfitting and thus reliant on a large set of labeled training data [17].

Even though deep learning models have shown promising results in main crop type mapping, catch crop monitoring has predominantly relied on machine learning models to date [3,24,25]. However, for catch crop mapping, the use of a Multilayer Perceptron network (MLP) has demonstrated improvements over the use of machine learning methods [26]. To our knowledge, this study is one of the few to compare traditional machine learning with deep learning methods for the end-of-season classification of catch crops while using optical time series input data. In contrast with traditional machine learning algorithms, deep learning models can learn feature representations from raw input (time series) without feature engineering [15,23,27], but in real-world applications with limited

training samples, optimizing the input dataset in both the time and spectral dimensions may still be necessary [14].

Moreover, deep learning models have shown promising results in early-season classification [19], which is useful when applying the models in the operational scenario of guiding in-season fieldwork to areas where the catch crop presence is most uncertain. In this application, the models also need to be generalizable in time, i.e., transferable between years. However, as most crop classification studies use training and testing data from the same year, the generalizability of models across years has not been well studied [28]. The temporal generalization of models, also named temporal transferability, is a crucial aspect of early-season classification in an operational setting when only training labels from previous years are available [29]. Specifically in the context of early-season mapping, this topic has received increasing attention recently [19,28–31].

The main objective of this paper is to test the potential of different machine learning and deep learning methods that differ in the way they handle the temporal dimension of the data for classifying catch crops from a time series of medium-resolution satellite images. We will also assess the impact of (1) the temporal resolution of the input time series and (2) the spectral resolution of the input features in order to further optimize the performance of the classification models. In the second objective, the optimized models are applied in a real-world scenario, where we aim to test the potential of these models for early-season catch crop classification. In this case, the target year is different from the training year(s), so this is an assessment of temporal model transferability. These research objectives give rise to four research questions:

- (1) What is the (relative) accuracy of different machine learning and deep learning classifiers for winter crop classification from optical satellite time series?
- (2) To what extent do the temporal and spectral resolution of the input data affect the models' classification accuracy?
- (3) How early in the season can the presence of catch crops reliably be detected to help prioritize field visits in areas of higher uncertainty?
- (4) How well do the trained models maintain their accuracy when applied to a target year not included in the training dataset?

2. Materials and Methods

This section will first describe the data that were used, then the models and their training process, and lastly, the experimental setup. An overview of the workflow is given in Figure 1.

2.1. Study Area

The study area covers the region of Flanders, which is located in the North of Belgium and is characterized by a temperate climate with warm summers and the absence of a dry season according to the Köppen–Geiger classification (Figure 2) [32]. Agriculture covers a total of 45% of the region's total land area, of which 55% is used for fodder production and 35% for arable farming [33]. In Flanders, the cultivation of catch crops in winter has been introduced as a measure to decrease nitrate leaching from agricultural fields. Reference data for modeling was collected throughout the region, with a focus on the municipalities with the highest nitrogen concentrations in their surface and groundwater. These municipalities are spread over Flanders and are located in the main agricultural clusters of the region.

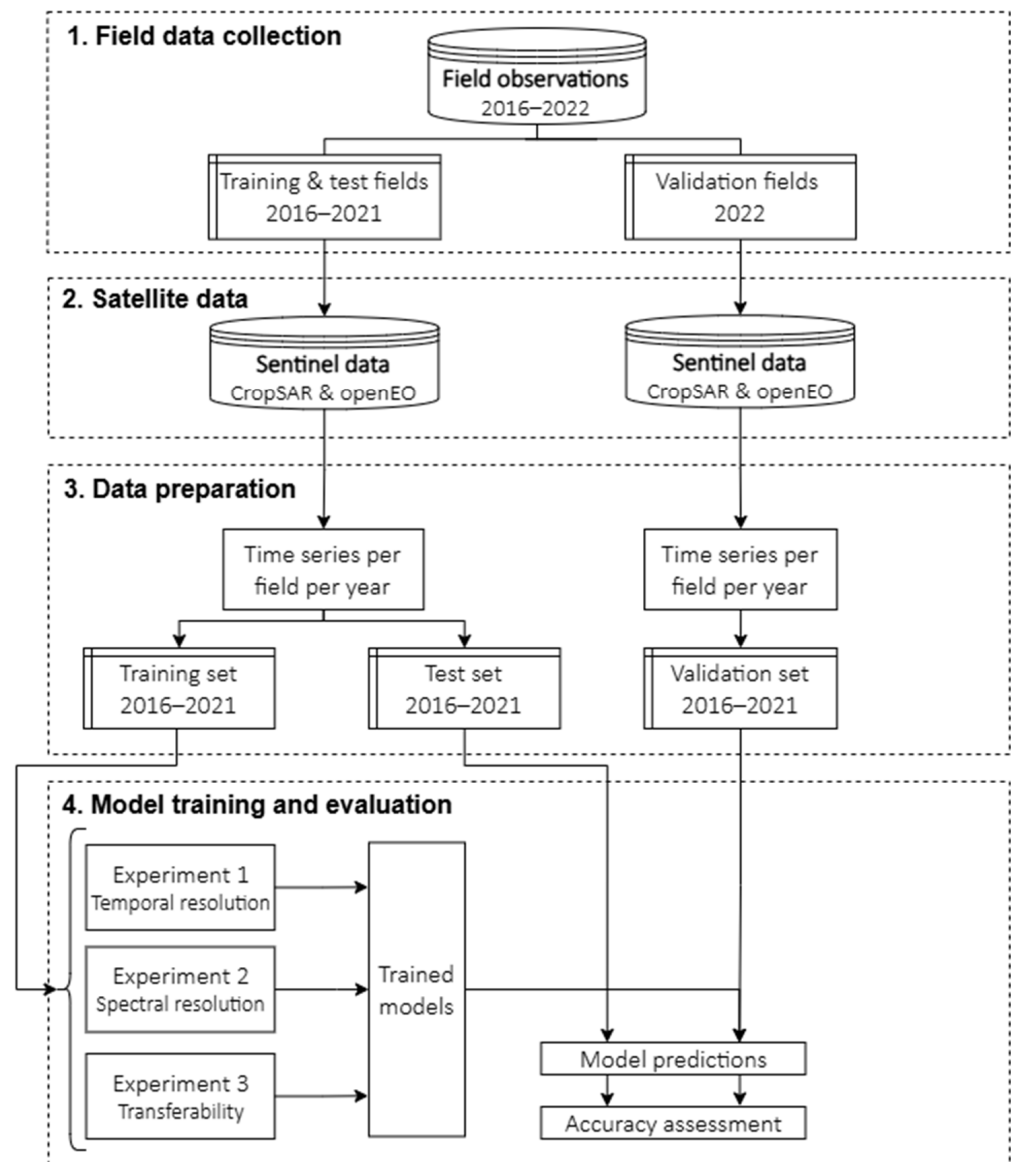


Figure 1. Highlights of the applied workflow. Three experiments were conducted: The first and second experiments aimed to study the influence of temporal resolution and input features on the classification, respectively on the classification. The third experiment evaluated the models' performance early in the winter season and their temporal transferability.

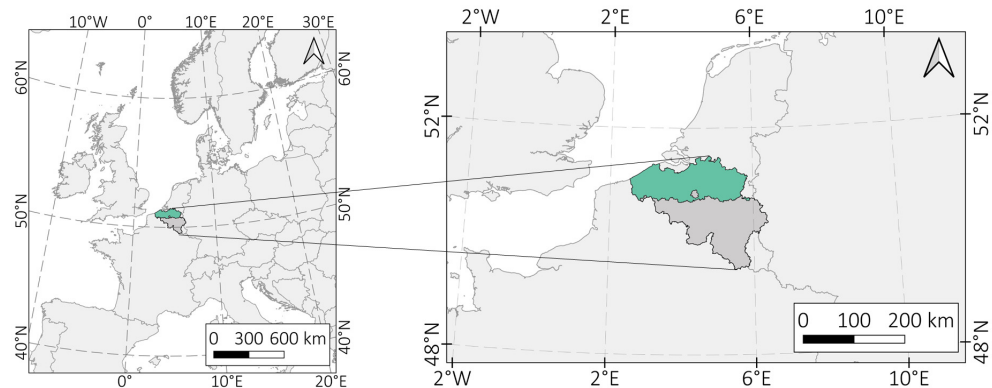


Figure 2. The location of the study region in Europe; Flanders is colored in cyan. Administrative borders: European Environment Agency (www.eea.europa.eu/datahub (accessed on 22 November 2024)).

2.2. Data

2.2.1. Reference Data

Field observations (2016–2021) were used as ground truth labels and were gathered by the Agency for Agriculture and Fisheries in the context of monitoring farmers' compliance and, more broadly, EU greening measures. Because this dataset was collected in a broader context, it contained unnecessary observations and, therefore, needed preprocessing, which is described in Appendix A. The ground truth labels obtained for this dataset consisted of the (catch) crop species present on a field during the winter period, from approximately October to January.

In addition to this dataset, extra field observations were collected during two dedicated field campaigns in collaboration with the Flemish Land Agency (VLM), the government agency responsible for monitoring farmers' compliance with the catch crop regulation. The two field campaigns were carried out in the winter seasons (October to January) of 2021 (i.e., winter 2021–2022) and 2022 (i.e., winter 2022–2023) and were focused on clusters of municipalities with the lowest water quality, which, due to regulations, had a high presence of catch crops. Data from the first field campaign were added to the training dataset, while those from the second field campaign were added to an independent validation set used for testing the temporal transferability of the trained models. Between field campaigns, the field selection process differed slightly. In the first field campaign (winter 2021–2022), fields were selected based on an annual farmer survey and a stratified random sample proportional to the declared catch crops, ensuring an even distribution across all municipalities. The second field campaign (winter 2022–2023) comprised two subsets: (i) fields selected based on label occurrence in the training dataset with priority for underrepresented labels (minority classes), and (ii) fields visited by a government agency targeting fields with previous absence of catch crops. In total, the dataset consisted of 6881 fields, of which 5921 fields were visited in observation years ranging from 2016 to 2021 (i.e., winter 2016–2017, 2017–2018, etc.). These observations comprise both the data collected by the Agency for Agriculture and Fisheries and the data collected during the first field campaign. These observations were used for training and testing the models. An additional 970 fields were observed during the second field campaign in 2022 (i.e., winter 2022–2023), which were used as the independent validation set (Figure 3).

2.2.2. Classification Groups

Because of the similarity in temporal profiles of (catch) crop species [3], species with similar spectral and temporal characteristics were grouped together. The catch crop species are listed in Table 1, and other crops cultivated during winter are referred to as winter crops. The catch crop species were subdivided into three groups based on their plant morphology and phenology (Table 1) [34,35]. The first group consists of leafy plant species; two example species in this group are white mustard and phacelia. This group is characterized by broad leaves and high aboveground biomass, properties that cause a high nitrogen uptake [34]. Plants in this group are very sensitive to frost damage and often do not survive winter. The second group consists of grass-like species with more narrow and upright leaves; two commonly used species are grass and oats (Table 1) [34,35]. As opposed to the first group, grass-like crops are winter-hardy and have a slower nitrogen uptake throughout the season. Their aboveground growth is slower, but their root system is extensive, which contributes to reduced soil slaking and an increased build-up of soil organic carbon. Mixtures of any of the allowed species can be considered as a third group, where mixtures can either consist of only leafy species, only grass-like species, or a combination of species from both groups.

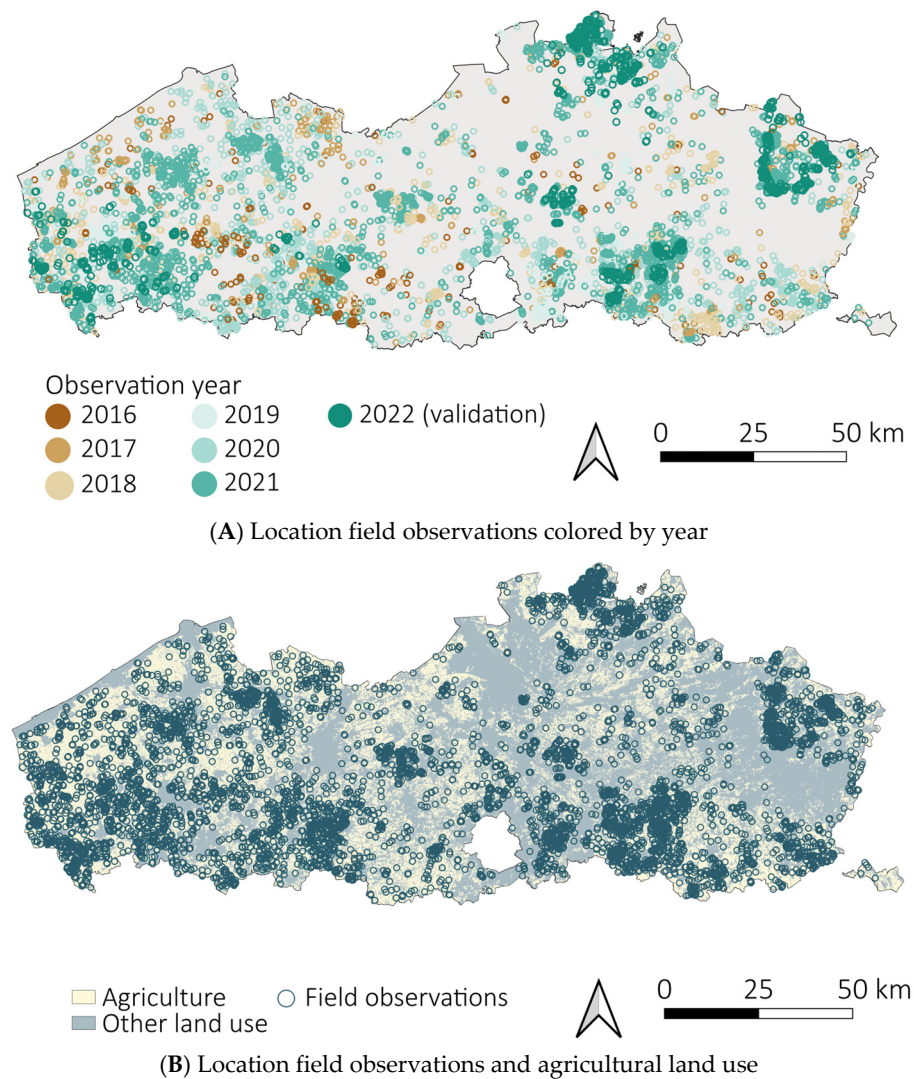


Figure 3. Locations of the 6881 field observations in the study area: (A) colored by year of observation where the color indicates the year in which the winter year started (e.g., 2022 are observations in the winter year 2022–2023) and (B) the field observations located on an agricultural land use map. Land use: Department of Environment and Spatial Development (www.vlaanderen.be/datavindplaats (accessed on 22 November 2024)).

The non-catch crop winter crop species were also grouped into three groups: (1) legumes, (2) winter cereals, and (3) other winter crops. Legumes are used as an agricultural practice to increase soil nitrogen and, therefore, are not used as a catch crop [34]. The winter cereals, winter barley (*Hordeum vulgare* L.) and winter wheat (*Triticum aestivum* L.), were put into a separate group because farmers can use these crops as an ‘equivalent measure’ instead of sowing catch crops. The difference with sowing catch crops is that the winter cereals sown in the current year remain on the field as the main crop in the following year, so they are not removed from the field after the required period, as is the case with catch crops. Examples of the third group, other winter crops, are cash crops sown during the study period, for example, leek (*Allium ampeloprasum* L.) and spinach (*Spinacia oleracea* L.).

Table 1. Overview of all allowed catch crop species with their common name, scientific name [36], and group.

Common Name	Scientific Name	Classification Group
Bristle oat	<i>Avena strigosa</i>	Grass group
Festulolium	<i>Lolium</i> × <i>Festuca</i> spp.	Grass group
Grasses	e.g.,	
Grass herbs	<i>Lolium multiflorum</i>	
Grass sods	<i>Lolium multiflorum</i> var. <i>westervoldicum</i>	Grass group
Undersowing grass		
Oat (spring)	<i>Avena sativa</i>	Grass group
Oat (winter)	<i>Avena sativa</i>	Grass group
Rye	<i>Secale cereale</i>	Grass group
Sudan grass	<i>Sorghum</i> × <i>drummondii</i>	Grass group
English ryegrass and white clover	<i>Trifolium repens</i> + <i>Lolium perenne</i>	Grass group
Arugula	<i>Eruca vesicaria</i> subsp. <i>Sativa</i>	Leafy group
Buckwheat	<i>Fagopyrum esculentum</i>	Leafy group
Cow cabbage	<i>Brassica oleracea</i> convar. <i>acephala</i>	Leafy group
Leaf mustard	<i>Brassica juncea</i>	Leafy group
Phacelia	<i>Phacelia tanacetifolia</i>	Leafy group
White mustard	<i>Sinapis alba</i>	Leafy group
Borage	<i>Borago officinalis</i>	Leafy group ¹
Nyger	<i>Guizotia abyssinica</i>	Leafy group ¹
Oil radish	<i>Raphanus sativus</i> supsp. <i>Oleiferus</i>	Leafy group ¹
Oilseed turnip	<i>Brassica rapa</i> subsp. <i>oleifera</i>	Leafy group ¹
Tagetes	<i>Tagetes</i> spp.	Leafy group ¹
Mixtures of non-legume cover species		Mixtures

¹ Interpretation based on plant morphology.

To classify these six different groups, the approach of a classification hierarchy was taken, dividing the complex classification problem into progressive classifications for improved accuracy in the case of complex classes with a lot of intra-class variability [37–39]. A limited two-level hierarchy limits process complexity and training time, as separate classifiers were trained for each level (Figure 4A). On the first level, fields were classified on winter crop presence or absence, separating bare soil and crop residues from fields occupied by a winter crop. In the second level, sown fields were further classified into six groups: three catch crop groups (leafy, grass-like, and mixtures) and three non-catch crop groups (legumes, winter cereals, and other winter crops). The resulting label distribution shows that (1) the overall majority of the dataset consists of fields with catch or winter crops present, (2) most observations were collected in the winter years of 2020–2022, and (3) there is a label imbalance within the six crop groups (Figure 4B).

2.2.3. Satellite Time Series

For every labeled parcel, the spectral signal was aggregated by taking the mean value over the field. These field boundaries are made publicly available by the government and updated annually, ensuring up-to-date field boundaries. They are also referred to as the Land Parcel Information System (LPIS). The object-based approach reduces noise and heterogeneity within a field and is useful in a decision-making context, as decisions are generally made at the parcel level [25]. To further reduce noise, we applied a 10 m inward buffer at each parcel, a technique that removes the spectrally mixed pixels at the edges of the parcels, thereby reducing the edge effects [10]. Input time series were downloaded for the period of August–April, chosen based on the crop calendars for catch crops and the other prominent winter crops in Flanders. This, therefore, delineates our study period.

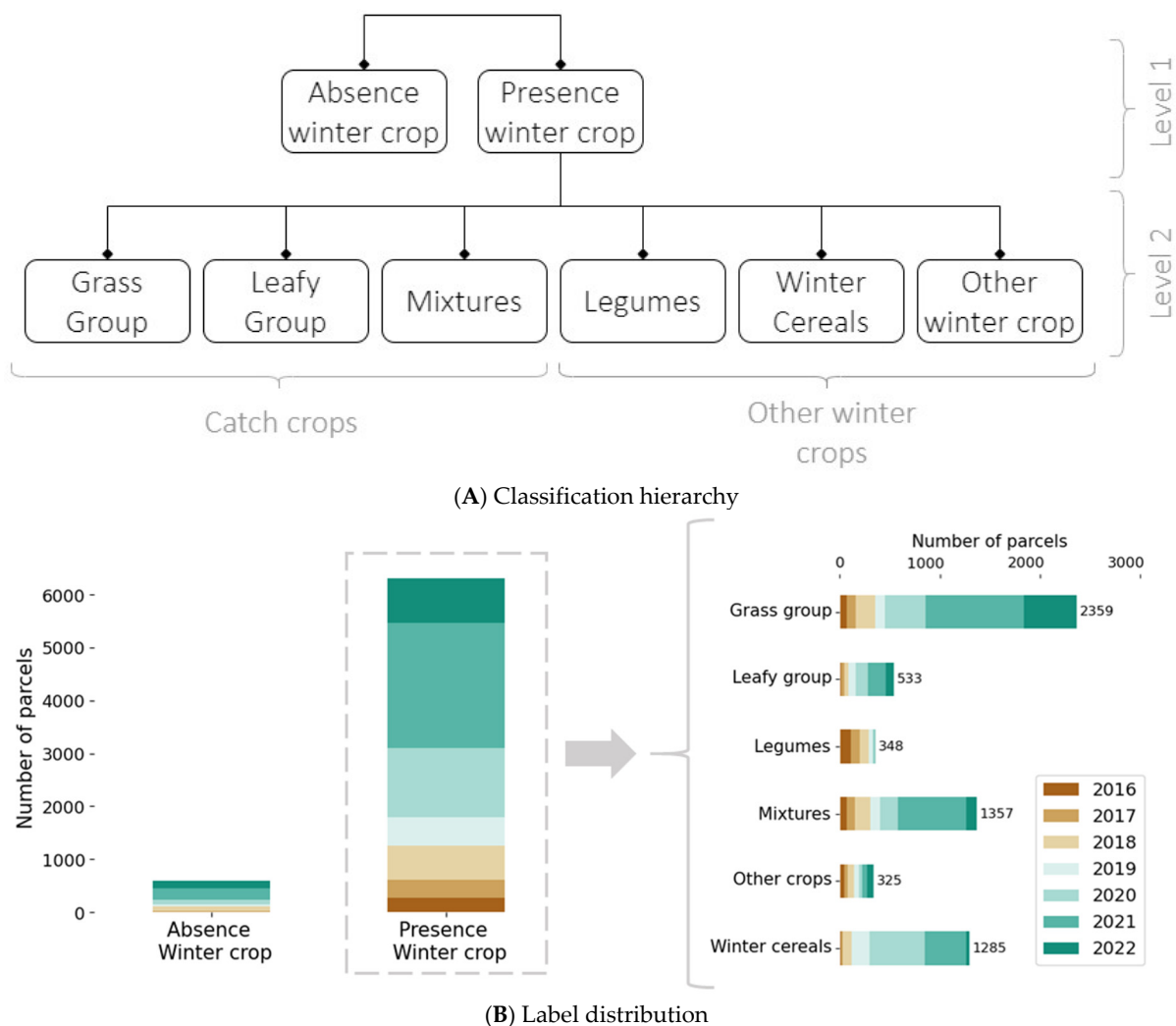


Figure 4. Overview of (A) the classification hierarchy and (B) the label distribution on both the first (left) and second level (right) of the hierarchy.

Time series data were calculated for every field using the Sentinel-2 optical satellite constellation, part of the Copernicus program. The mission consists of a two-satellite constellation in a sun-synchronous orbit with a 180° phase difference [40]. The first satellite was launched in 2015, and the second in 2017. Every 5 days, the same area is visited under equal viewing conditions, but due to overlap in swaths of different orbits, there is a revisit time of about 2–3 days over Belgium. Sentinel-2 uses optical remote sensing in the range of 400 nm to 2500 nm, which includes the Visible (VIS), Near Infrared (NIR), and Short-Wave Infrared (SWIR) regions. The multispectral sensor, MultiSpectral Instrument (MSI), measures 13 spectral bands at different spatial resolutions: four bands at 10 m, six bands at 20 m, and three bands at 60 m. To obtain data from this mission, two different services were used, CropSAR (version 1.0) and openEO (version 0.22.0).

CropSAR is a service that monitors crop growth in Belgium, and it is provided by VITO, the Flemish Institute for Technological Research [41]. The CropSAR service provides daily time series by applying a deep learning network trained on cloud-free images of both Sentinel-1 and Sentinel-2 to interpolate between the gaps of optical acquisitions, thereby combining optical and radar data [42]. CropSAR was used to download preprocessed, ready-to-use daily time series for three different indices: the Normalized Difference Vegetation Index (NDVI), the fraction of Absorbed Photosynthetically Active Radiation (fAPAR), and the Fraction of green vegetation Cover (FCover) [43]. While the NDVI is an estimation of the greenness of the land surface, fAPAR and FCover are both biophysical variables

of vegetation and can also be measured on the land surface. VITO derived fAPAR and FCOVER from satellite observations by using a neural network to estimate these biophysical variables from Sentinel-2.

OpenEO provides an API that can be used to acquire and process large Earth observation datasets from different providers in the cloud [44]. The API was used to access Terrascope Level 2A Sentinel-2 data and subsequently transform the raw series of individual images into an equidistant time series. Processing consisted of first masking clouds and cloud shadows with the Scene Classification Layer (SCL) through a kernel process provided by openEO (standard deviation 1.5, window size 11). Then, gaps in the time series were filled by linear interpolation: the time series were first divided into equidistant time intervals, using the median value over a specified period, and then interpolation was applied to fill gaps in the time series. To fill gaps at the start or end of the time series, an extra month was added at the start (July) and end (May). Where gaps remained, the value of the next valid period was duplicated, and if gaps remained, the value from the previous valid period was duplicated. Time series were downloaded with dekadal time intervals for bands B02-04 (VIS), B05-07 (vegetation red-edge), B08 (NIR), and B11-12 (SWIR) of Sentinel-2. Dekadal time intervals consist of three intervals per month, with the first two being 10 days each and the third comprising the remaining 8–11 days, depending on the length of the month. In addition, time series were also downloaded with monthly time intervals for the same bands of Sentinel-2.

2.3. Methods

We compared three machine learning models for classifying time series into the winter crop classes: Random Forest (RF), Time Series Forest (TSF), and a One-Dimensional Convolutional Neural Network (1D-CNN). The latter two models explicitly account for the temporal order of the data, while the RF model is added as a baseline method. To compare the model accuracies, each unique model–dataset combination was trained five times with different random seeds, using the same five seeds for each combination to maximize result comparability. All model runs were trained on 75% of the reference dataset and evaluated on the 25% remaining data. The train–test split was stratified by (catch) crop species to maximize the comparability of the training and test sets. Hyperparameters were tuned to maximize the Cohen kappa coefficient, in accordance with [23], and determined at the first hierarchy level, then applied to the second level to reduce selection complexity. In what follows, the models and their training process will be described in more detail. The entire overview of the selected hyperparameters per dataset is given in Appendix B.

2.3.1. Random Forest

RF is a non-parametric classifier known for its high accuracy and fast processing, using multiple decision trees created through a bagging approach [20,45]. It assigns class labels based on the majority vote from these trees. The RF model was implemented using the sklearn (v1.3.2) library in Python (v3.11.6) [46,47]. We optimized only the number of estimators to ensure maximal comparability between the two forest-based models (RF and TSF) and set the maximum number of features to zero for the same reason (Table 2). Hyperparameters were tuned using the RandomizedSearchCV function from sklearn, randomly sampling 10 hyperparameter value combinations from the corresponding search space and a 5-fold cross-validation [46]. For the RF model, the search space was too small to sample ten settings, so a maximum of four samples was used, i.e., the entire search space for this model (Table 2).

Table 2. Hyperparameter search spaces for each model.

Model	Hyperparameters	Search Space
RF	n_estimators	200, 300, 400, 500
TSF	n_estimators	200, 300, 400, 500
	min_interval	3, 5, 7, 9
CNN	batch_size	8, 16, 32, 64
	learning rate	0.001–0.1
	num_epochs	15, 20, 25, 30
	dropout	0.1, 0.2, 0.3, 0.4
	kernel_size	3, 5, 7, 9
	stride	1, 2, 3, 5

2.3.2. Time Series Forest

A Time Series Forest (TSF) model is a modified Random Forest model specifically adapted to the classification of time series [22]. TSF is a feature-based classification where the feature's mean, standard deviation, and slope are derived from randomly chosen time intervals in the time series. These features are calculated over a number of intervals and then used for classification. The model was implemented using the sktime library in Python (v0.21.0), using both the TimeSeriesForestClassifier and ColumnEnsembleClassifier for univariate time series input and multivariate input time series, respectively [48,49]. Two hyperparameters were optimized again using the RandomizedSearchCV function: the number of estimators (trees) and the minimum interval length (Table 2).

2.3.3. One-Dimensional Convolutional Neural Network

CNNs have been used in several vegetation remote sensing applications, including main crop classification and tree crown delineation [27,50]. CNNs make use of convolutions, which can be seen as sliding a filter over input data to detect local patterns, with the weights of this filter optimized during training to identify discriminative features [51,52]. A key advantage of CNNs is weight sharing across all time stamps, reducing the number of parameters to be tuned [17,51].

The 1D-CNN was implemented using the PyTorch (v2.0.1) library in Python [53]. The model architecture consists of three sequential convolutional blocks, each consisting of a one-dimensional convolutional layer, a batch normalization, and a ReLU activation function (Figure 5). These three convolutional blocks are followed by two fully connected blocks, each consisting of a linear layer, a batch normalization, and a ReLU activation function (Figure 5). The batch normalizations standardize the inputs of each input batch and thereby help regularize the network and reduce overfitting [17].

For training the network, we used the minibatch stochastic gradient descent approach and implemented it using the SGD optimizer in Pytorch [54]. For the loss function, we used binary cross-entropy at the first level of the hierarchy and cross-entropy at the second level of the hierarchy because binary and multiclass classifications require different loss functions [54]. Using the Ray Tune library in Python, we optimized the following hyperparameters for the network: window size and stride for the convolutional layers, and batch size, learning rate, number of epochs, and dropout for training the network (Table 2) [55]. For each input data set, the same 50 hyperparameter combinations were tested out, and bad trials were terminated using the median stopping rule. Instead of a 5-fold cross-validation, the training data (75% of the dataset) was split once in an 80–20 split, stratified by the classification group. This means that, in total, 60% of the entire reference dataset was used for training the models on various hyperparameter sets and 15% for evaluating. The remaining 25% was allocated to the validation set.

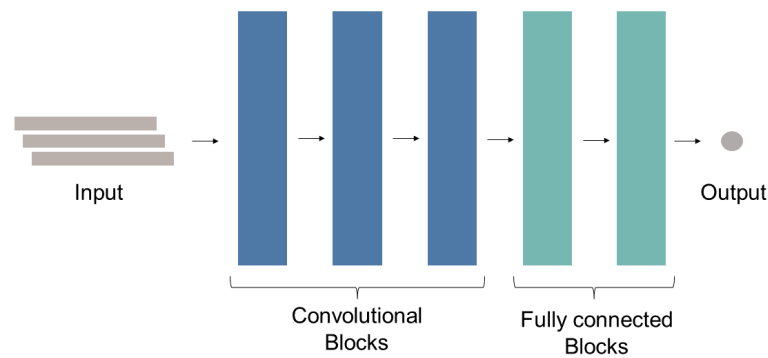


Figure 5. Network architecture of the One-Dimensional Convolutional Neural Network (1D-CNN). The convolutional blocks each comprise a convolutional layer (number of filters, 8, 16, 32), a batch normalization, and a Rectified Linear Unit (ReLU) activation function. The fully connected blocks each consist of a linear layer, a batch normalization, and ReLU activation function.

2.3.4. Experimental Setup

The first research objective aimed to compare models and inputs for catch crop classification, which was addressed through two experiments, evaluating multiple factors affecting classification accuracies for all three models. The first experiment aimed to study the influence of temporal resolution on the classification by using time series with three different temporal intervals: daily, dekadal, and monthly intervals. To achieve this, we used the daily NDVI time series downloaded from CropSAR and resampled it to dekadal and monthly intervals by taking the mean value in every interval (Figure 6). The experiment used the full time series length, ranging from August to April.

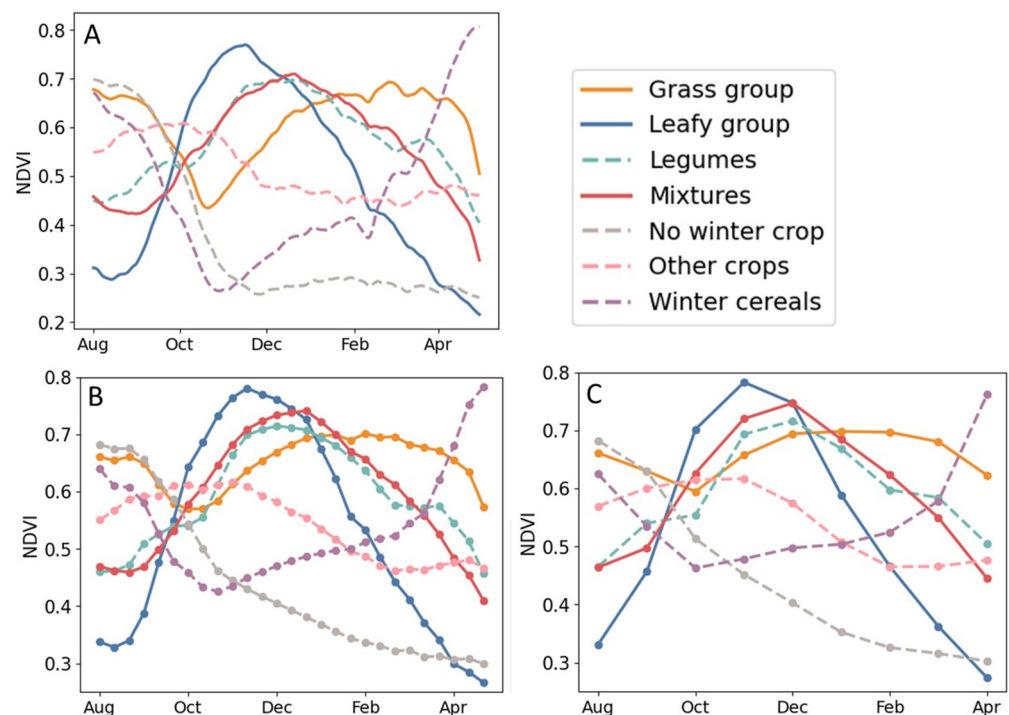


Figure 6. Average NDVI time series at (A) daily, (B) dekadal, and (C) monthly temporal resolution for each of the classes. Linear interpolation was used to create continuous time series.

The second experiment evaluated how input features affect classification accuracy, aiming to determine if additional information beyond NDVI could improve accuracy. Part A of this experiment compared daily NDVI time series with multivariate time series (NDVI, fAPAR, and FCover) to assess the added value of biophysical variables. In Part B, we

used the dekadal Sentinel-2 dataset to construct three input feature sets: (1) dekadal NDVI time series, (2) dekadal time series with the raw reflectance values from each band of Sentinel-2 and (3) dekadal time series with normalized reflectance values from each band of Sentinel-2. This tested the impact of both additional features and normalization on classification accuracy. Z-normalization was applied on each timestamp–band combination separately and was calculated as follows:

$$z = \frac{x - u}{s} \quad (1)$$

where x is the reflectance value, u is the mean, and s is the standard deviation, calculated using the StandardScaler in sklearn [46]. Again, we used the full time series length, ranging from August to April, for this experiment. These first two experiments combined are used to answer the first two research questions.

In the second research objective, we focused on applying the models in a real-world scenario for early-season classification on a year excluded from training and validation. Therefore, the third experiment evaluated model performance throughout the season by varying time series lengths, always starting in August and extending month by month to April. Testing the temporal transferability of the models was done for the years the models were trained upon and on the dataset for the winter 2022–2023, which was excluded from training. This experiment used the input feature set with the highest accuracy, which was identified in the previous two experiments, and aimed to answer the third and fourth research questions.

2.3.5. Accuracy Assessment

To evaluate the accuracy of the models, the results from the classification hierarchy were first combined into two coarse-labeled groups: catch crops and no-catch crops. The catch crop group consists of the grass group, leafy group, and mixtures, while all other classes are grouped in no catch crops. Based on the classification of these two groups, the models were compared with a macro F1-score as a measure to compare the overall classification. The F1-score was chosen over the overall accuracy because of the role of chance agreement in an unbalanced dataset, where the overall accuracy favors the dominant class [29]. The calculation of the F1-score is given by:

$$F1 = 2 * \frac{Precision * recall}{Precision + recall} \quad (2)$$

where precision is the proportion of positive (negative) predictions that are actually positive (negative), and recall is the proportion of positives (negatives) that are actually predicted as positive (negative). Because the models were run five times on each unique input dataset, each time with a different random state, all the F1-scores were calculated five times.

3. Results

3.1. End-of-Season Classification

The first experiment examined the effect of temporal resolution on the classification performance of all models. All models achieved the highest F1-scores on the daily temporal resolution for the coarse-label classification, with a median F1-score of 86.5% for the RF model, 86.5% for the TSF model, and 85.9% for the CNN model (Figure 7). Every model type reached the lowest F1-score on the monthly time resolution, and here, the median values for the F1-scores were 85.1% for RF, 81.8% for TSF, and 83.6% for CNN. Further, each model type has a different rate in the decline of the F1-scores, where the TSF model has the largest decline in performance, while RF has the lowest decline.

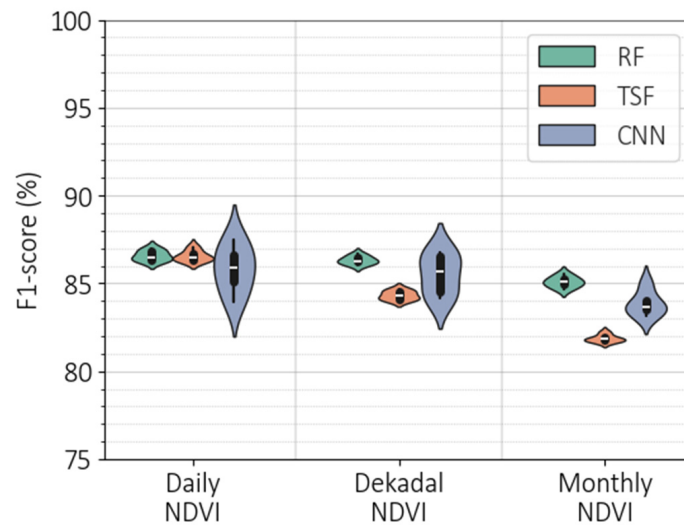


Figure 7. Violin plots for the F1-scores of the first experiment given for three temporal resolutions and three different models (Random Forest or RF, Time Series Forest or TSF, and a One-Dimensional Convolutional Neural Network or 1D-CNN). Within the violin plots, the boxplots are depicted in black.

In part A of the second experiment, we then looked at the influence of biophysical variables on the classification performance. At both daily and dekadal temporal resolutions, the dataset, including the biophysical variables (fAPAR and FCover), attained higher F1-scores for every model type compared to the NDVI-only dataset (Figure 8A).

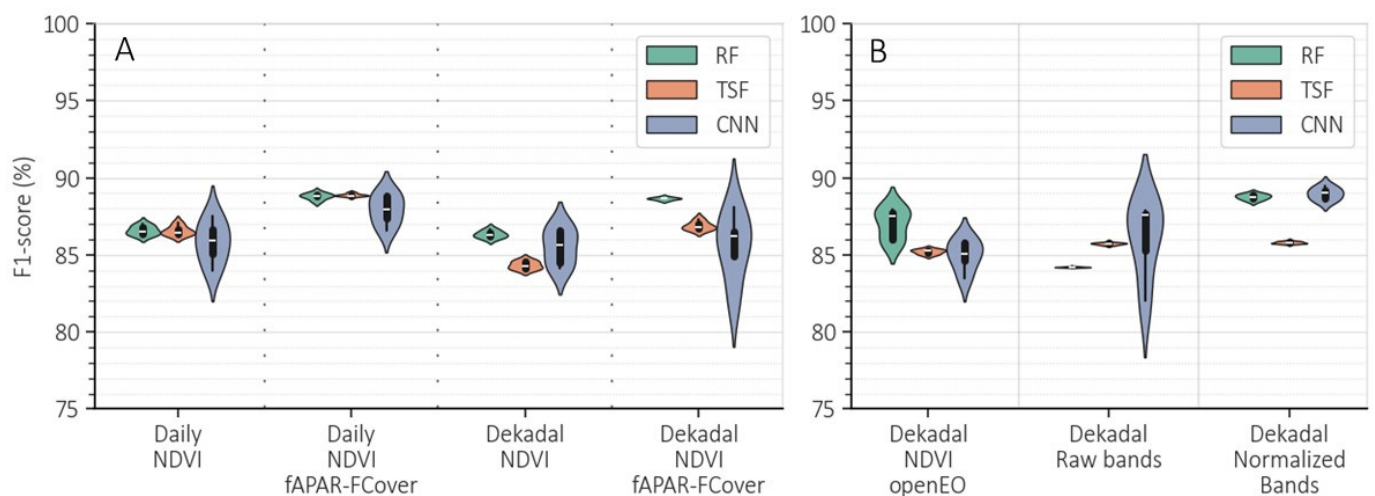


Figure 8. Violin plots for the F1-scores of the second experiment given for three temporal resolutions and three different models (Random Forest or RF, Time Series Forest or TSF, and One-Dimensional Convolutional Neural Network or 1D-CNN). (A) Part A of the second experiment and (B) part B of the second experiment. Within the violin plots, boxplots are depicted in black.

Part B of this second experiment looked at the classification performance of the models on the time series, purely derived from Sentinel-2 data. Three sets of input features were compared at a dekadal temporal resolution: NDVI, the raw bands, and normalized bands. When using NDVI as an input feature, the median F1-scores were in the range of 85.1–87.5% (Figure 8B). However, the CNN resulted in a larger range of F1-scores (83.5–86.0%) compared to the forest-based models. This is especially the case for the F1-scores with raw bands as an input dataset, where a wide range of values was obtained (F1-scores 82.0–87.9%). However, the CNN model showed less variability in model performance

when the Sentinel-2 bands were normalized (F1-scores 88.5–89.5%). In contrast, the TSF model resulted in a constant performance with an F1-score of around 85.3–85.8%. The median F1-score for the RF model showed a decrease from 87.5% to 84.2% when using the raw bands of Sentinel-2 compared to that using NDVI only. When normalizing the bands of Sentinel-2, performance increases again to a median F1-score of 88.8%.

Looking more closely at the fine-label classification of the three models, it was shown that all models confused the three catch crop groups (leafy, grass-like, and mixture) with each other (Figure 9). The leafy group was mostly confused with the mixtures, and the mixtures were mostly confused with the grass group. There were also some misclassifications between catch crops and winter cereals, mostly between the grass-like and mixture groups. Small groups like legumes, other winter crops, and the no-crop group mostly ended up in the three biggest groups: the grass group, mixtures, and winter cereals.

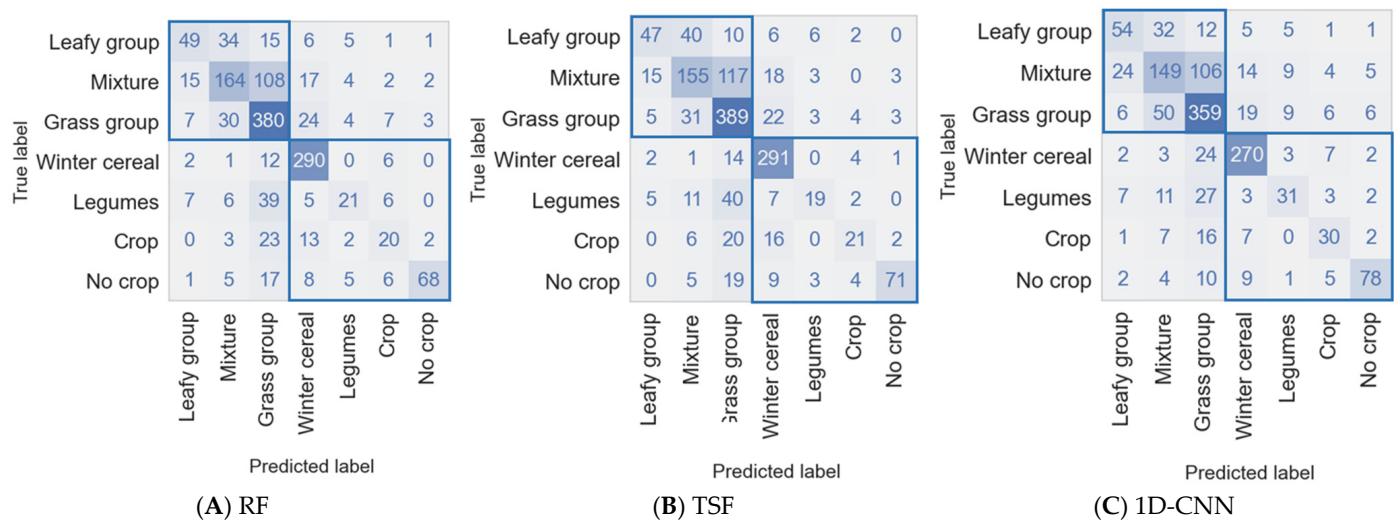


Figure 9. Confusion matrices for the fine-label, end-of-season classification for three different models: (A) Random Forest (RF), (B) Time Series Forest (TSF), and (C) One-Dimensional Convolutional Neural Network (1D-CNN). The blue frames indicate the coarse-label classification groups.

3.2. Early-Season Classification

From the first two experiments, the input feature set that achieved the highest overall F1-scores was selected, which was the daily NDVI, fAPAR, and FCover time series. The third experiment focused on assessing the early-season performance of the models and their temporal transferability using this ‘best’ input feature combination. When training and evaluating the same years, all models showed an increase in F1-scores when including extra time steps throughout the first three months (September, October, November) (Figure 10). However, adding extra time steps only marginally improves model performance from December onwards, as indicated by a stagnation in the F1-scores. The inclusion of March and April into the time series again resulted in an increase in F1-scores.

When evaluating the early-season performance of the models separately, the forest-based models (RF and TSF) attain a median F1-score surpassing 85% in November, whereas the CNN model only reaches this threshold in March (Figure 10). By April, all models reach their highest median F1-score. However, the CNN model resulted in lower and more variable F1-scores, particularly earlier in the season.

When applying the trained models to the time series of a target (winter) year to assess transferability, the increase in F1-score with time steps remained for the forest-based models (Figure 10). In contrast, the CNN model showed a very unstable performance that lacked this general upward trend and produced a larger range of F1-scores compared to the other two models. All models reached maximum model accuracy in April with median F1-scores of 88.8% for RF, 88.3% for TSF, and 81.5% for the CNN model. Therefore, the forest-based models reached very similar median F1-scores when training and evaluating in the same years in April (RF 88.7% and TSF 88.9%), while the CNN model reached a lower median F1-score in comparison (87.9%).

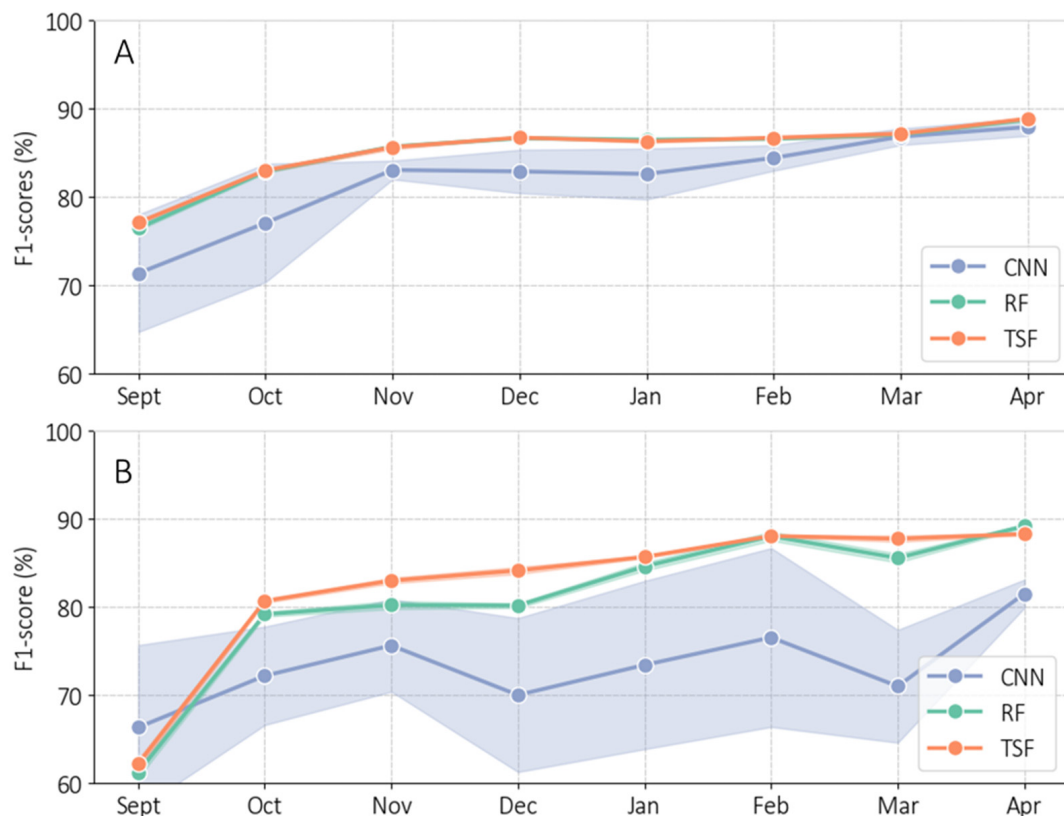


Figure 10. F1-scores for the third experiment, which was completed on daily time series with features NDVI-fAPAR-FCover for three models (RF, TSF, and 1D-CNN). (A) Results when trained and tested on the same years; (B) results when the models are applied on a target year. Month on the x-axis shows the end point of the time series: e.g., Dec represents a time series starting in August and ending in December. Shading indicates the standard deviation on the results from the different model runs. Note that the results for RF and TSF largely overlap in panel (A).

Revisiting the early-season performance of each model, TSF achieves a median F1-score of 85% in December, followed by RF in February, whereas the CNN model fails to reach a median F1-score of 85%. Compared to the same year training and testing, the models attain this threshold later in the season during the separate validation year. Maximum F1bscores were again all reached in April. The CNN model reached a score of 83%, while RF and TSF reached F1-scores of 89% and 88%, respectively. These classification results were illustrated on a map using the RF model (Figure 11).

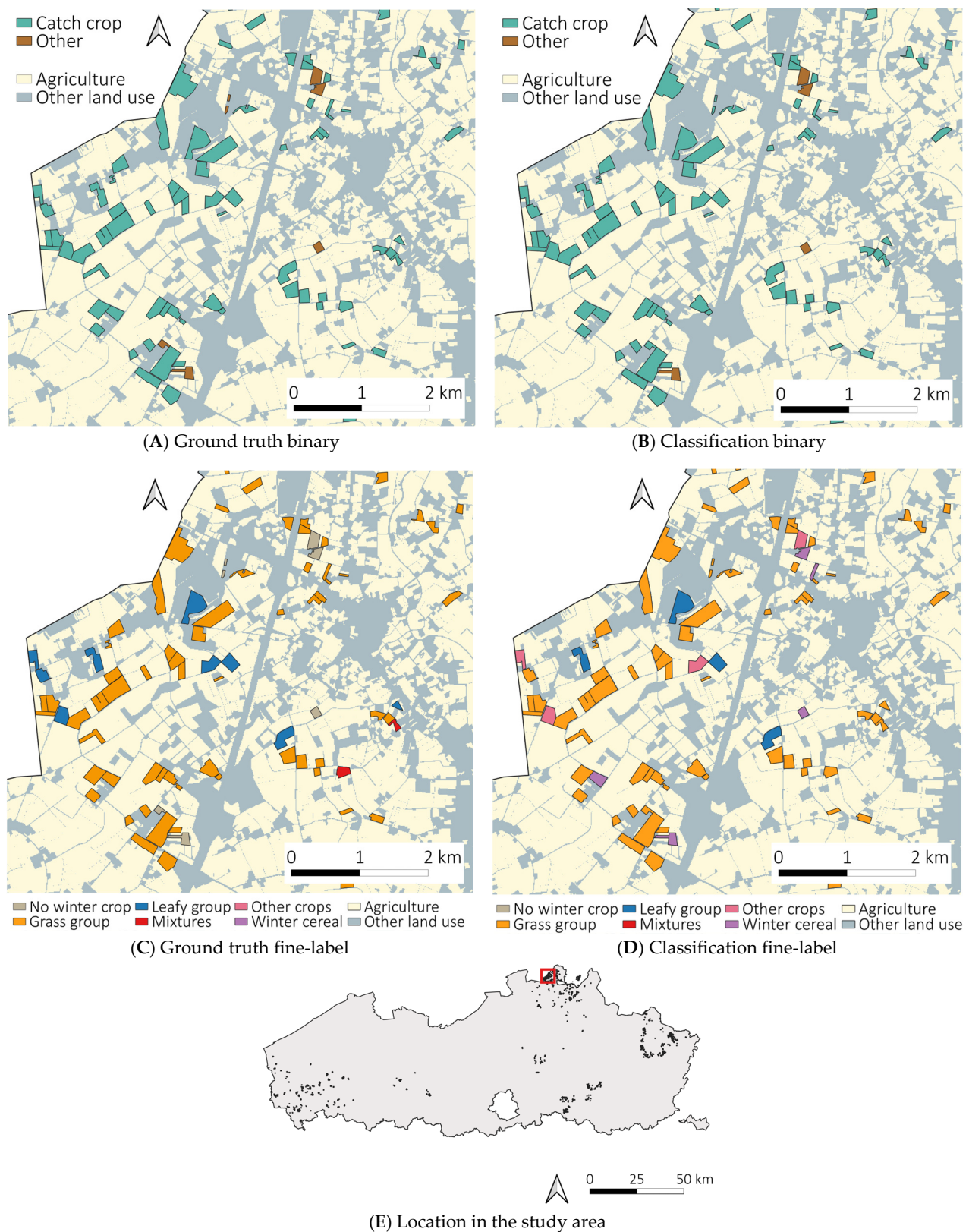


Figure 11. Maps to illustrate the results of the Random Forest (RF) model on time series until April. Binary (A) ground truth and (B) classification results, fine-label (C) ground truth and (D) classification results. (E) Location of maps (A–D) in the study area. Land use: Department of Environment and Spatial Development (www.vlaanderen.be/datavindplaats (accessed on 22 November 2024)).

4. Discussion

4.1. Model Comparison

Our results show marked differences in the performance of three tested classification models (RF, TSF, and 1D-CNN). While the RF algorithm demonstrated its versatility by delivering consistent results on different datasets, the TSF model was more dependent on the input dataset than RF, and the 1D-CNN model often resulted in lower accuracies than the other models and displayed a high degree of variability between model runs. Given that RF has been used in a wide range of remote sensing classification tasks, as reviewed by [20], our finding that it performs robustly across datasets is consistent with previous research. In contrast, the results of our time series-specific algorithms differ from previous research. For example, in their study comparing time series classification algorithms on univariate time series, TSF outperformed both baseline classifiers, with RF showing similar performance to the baseline classifiers [56]. However, this study did not include remotely sensed time series, and their datasets mostly contain time series with <1000 training instances that are longer than 50 time steps. In contrast, our datasets contain >5900 training instances with 9, 27, or 273 time steps, depending on the temporal resolution (respectively monthly, dekadal, and daily), which may also explain the different results.

Further, in time series-based crop classification, several studies have demonstrated enhanced classification results with the use of a deep learning classifier in comparison with shallow machine learning techniques. For example, refs. [15,17] reported improved crop classification results using 1D-CNNs compared to RFs. However, in our application, 1D-CNNs often did not outperform a baseline RF, with the exception of the normalized dekadal band dataset in experiment 2B. A very likely explanation for these contrasting results is the limited number of training samples in our field-visit and object-based approach, resulting in approximately 6000 training samples over multiple years and ranging from 325 to 2359 samples per class (Figure 4B). A recent study on time series-based crop classification using 1D-CNNs used 4000 training samples per class, a number chosen after an analysis of the amount of training data and the mapping accuracy [57]. These results suggest that 28,000 training samples, spread evenly across classes, would be required for deep learning to be effective in our catch crop classification. Often, many more training samples are used. References [15,17] used labels available from a single-year crop survey in combination with a pixel-based approach, resulting in approximately 1 million and 2 million training samples, respectively, with samples per class ranging from 9000 to 292,000. Due to its data limitation, the catch-crop classification could not fully benefit from the advantages that deep learning models provide. Therefore, the large number of labeled data needed for deep learning models proved to be a limitation for improved catch crop-type classification with these models. Switching to a pixel-wise classification would be a quick win for creating more training data without increasing the field efforts. Other options for handling the large data requirement of deep learning models are data augmentation and transfer learning; the latter will be discussed below.

4.2. Fine-Label Classification

Across all models, it proved difficult to make a fine-label classification of winter crops based on their temporal profile alone (Figure 9). The fine-label classification resulted in confusion between different groups, notably between the grass group and the mixtures, which was expected because mixtures can also be a mix of two grass-like catch crops. The same can be said about confusion between leafy species and mixtures, as mixtures can consist of multiple leafy species. Theoretically, mixtures between grass-like and leafy groups could result in a unique canopy structure and, thus, a distinguishing feature from the other catch crop groups. However, in reality, competition between species makes the proportion of a single catch crop in the canopy unpredictable, resulting in canopy heterogeneity [58]. Factors influencing the canopy composition are management, soil type,

and weather conditions rather than the ratio in the seed mix. The heterogeneity of the mixture group limits the separability of the catch crop groups.

Our results are similar to those found by [3], who also had difficulty distinguishing between catch crop species based on a time series-based classification with RF. Further, [59] similarly performed a time series-based classification with RF to distinguish two cover crop species: cereal rye and red clover. In particular, they noted that rye was often misclassified as fallow, which they attributed to (1) the highly variable biomass production depending on the planting date, causing fields with low biomass to resemble fallow fields with weeds, and (2) the interseeding of rye with crop residue, resulting in fields with low biomass production to appear similar to fields with maize stubble. These observations of high biomass variability and interseeding with crop residue were also noted in our fieldwork for the grass group (including rye), which also resulted in variability in spectral signals within this group for a single winter cropping year (Appendix C). The high variability within the grass group could, therefore, explain the high confusion of this group with the other classes.

4.3. Effect of Temporal Resolution and Input Variables

The first experiment revealed that a higher temporal resolution of the input time series resulted in a more accurate classification. This is consistent with the results of crop classification studies by [19,60], who also reported improved accuracy with higher temporal resolution, although with marginal improvements in [60]. The highest accuracies were achieved using multivariate daily time series consisting of NDVI, fAPAR, and FCover inputs. This suggests that incorporating a wide range of features can indeed be beneficial, contrary to the expected limitations of the Hughes phenomenon, where the addition of too many features beyond a certain point does not improve classification accuracy [61]. This added value of using more input features is observed at both daily and dekadal temporal resolution.

In addition to the use of other engineered features, the use of (normalized) reflectance bands improved the performance of most of the RF and CNN models over using NDVI alone. Notably, the combination of CNN with normalized input features resulted in the best performance for the CNN model. This finding indicates a potential loss of information when overly relying on feature engineering in the case of CNN models [50]. Normalization of input features proved crucial when using reflectance band inputs in combination with the CNN model: the normalization not only increased the F-scores but also yielded more consistent predictions between model runs. This improvement can be attributed to the normalization process ensuring that the variation in data is not overly dominated by features with a larger dynamic range, thus balancing the influence of each feature on the model's output [17]. Experiment 2B showed that both extensively preprocessed datasets for higher temporal resolution and raw input time series can achieve similar accuracies when used with the appropriate model type.

The experiments demonstrated that additional information could be extracted from the raw input by using biophysical variables in addition to the NDVI or by using a suitable model type in combination with the raw input itself. Future work could explore additional vegetation indices, similar to the approach taken by [59] in their catch-crop-type classification, especially since a number of alternative indices are aimed at addressing specific NDVI limitations. In addition, our study focused on the use of Sentinel-2, and although the daily NDVI, fAPAR, and FCover time series were derived from combining Sentinel-1 and Sentinel-2, a winter crop classification might benefit from a direct integration of radar-based (Sentinel-1) input features [8,62,63].

4.4. Early-Season Classification Performance for Separate Target Year

The third experiment showed the trends in classification accuracy throughout the winter season. The F1-scores increased until November, then stagnated, and increased again throughout March and April. Overall, the models improved as the time series lengthened, which is an expected result and consistent with previous research [14]. The stagnation could potentially be explained by the limited number of cloud-free images that are available in these parts of the season, which could restrict the early-season classification for winter crops compared to that of main crops. Peak F1-scores were reached by the end of April, which is likely due to the distinct temporal profile of winter cereals when April is included (Figure 6). Because winter cereals, unlike catch crops, remain on the field after winter, their time series show a continuous increase in the vegetation index values around March and April, which contrasts with the downward trend of catch crops over the same period (Figure 6). This discriminating characteristic is probably also the reason that the fine-label classification of winter cereals performed relatively well compared to other classes (Figure 9).

When comparing the models for their early-season classification capabilities on the validation year, we see that the forest-based models reach the threshold F1-score of 85% later in the season than for the years included in the training. A slight drop in performance on a separate target year was expected, as temporal (and spatial) transferability is generally a challenge in crop classification, even when multiple years have been used for training [28]. This is due to different weather conditions between years, which affect crop development and result in different temporal profiles for the same crop [28,29]. Reference [19] even reports that the variation in spectral profiles between years is the main factor in reducing the accuracy of the early-season mapping of main crops. Interannual differences could, therefore, offer an explanation for the slightly lower accuracies of the forest-based models earlier in the season (Appendix C). The differing weather conditions during the summer and autumn months every year impact the timing of growth stages of the catch crops, resulting in variations in crop development earlier in the season. For example, a large part of the training data come from the year 2021 (winter year 2021–2022), with an exceptionally wet summer and a rather dry autumn, while 2022 (winter year 2022–2023) had an exceptionally dry summer followed by a warm autumn, which is likely to have affected the growth stages of the catch crops differently compared to the previous year [64,65].

However, interannual differences do not entirely explain the lack of generalization to the target year for the 1D-CNN model; its results most likely point to an overfitting problem. Although various regularization methods, such as dropout and batch normalization, were used to prevent overfitting, the model type itself also plays an important role: given the large number of parameters that need to be fitted for a deep learning model such as a CNN, the limited number of samples is again the most likely explanation. Even though the potential of deep learning approaches, and more specifically, 1D-CNN, have been reported for early-season mapping [19], our application does not seem to benefit from deep learning models over machine learning models. RF is the better option for early-season classification, given our limited dataset.

This issue of temporal model generalizability, particularly in the context of early-season mapping, has gained attention in recent crop mapping research [28–30]. These studies have focused on early-season time series classification with transformer-based models, which represent the current state-of-the-art in deep learning. Transformers can exploit temporal dependencies and handle variable time series lengths as inputs [66], making them highly suitable for early-season mapping as a single model can be used for progressive mapping throughout the season [28,29]. Ref. [29] demonstrated the effectiveness of a transformer-based deep learning architecture for early-season crop mapping, highlighting its temporal (and spatial) generalizability. However, their approach requires a substantial amount of labeled data to train this transformer-based model, an architecture that often requires larger datasets in comparison to CNNs to achieve better performance [67]. Our

study already faced the constraint of limited sample availability with CNNs, so it can be questioned how useful this specific model would be for our catch crop classification.

Fortunately, recent research also addresses this aspect of limited labeled data availability. Here, transfer learning, a machine learning technique that transfers knowledge between related tasks, emerged as a strategy to alleviate the need for labeled data in environmental remote sensing [67]. For example, ref. [28] demonstrated pre-trained models, providing a strong starting point for early-season crop mapping in a separate target year for which no data are yet available, even when the vegetation trend differs between years. Reference [30] similarly demonstrated the effectiveness of transfer learning for the transferability of crop mapping models on Sentinel-1 data. Therefore, there is a need for the development of transferable models or even foundation models in remote sensing, which are large pre-trained and often transformer-based models that can be applied to a variety of tasks downstream through transfer learning [67]. As such, future work can explore transfer learning for early-season catch crop classification—for example, with a recently developed pre-trained transformer model for remote sensing time series by [68]—as this approach seems to generalize well even with limited training data.

5. Conclusions

This study aimed to assess the feasibility of end-of-season catch crop classification with Sentinel-2 time series, with specific attention given to the optimization of the model type and the input features. The best model–input combination was then tested in a real-world scenario for early-season catch crop classification in a separate target year, thereby testing its temporal transferability. In summary, RF displayed the highest adaptability to input features and often outperformed time series-specific classification methods, TSF, and 1D-CNN. This study demonstrated the benefits of using high temporal resolution time series, as well as a diverse range of input features for the catch crop classification task. When the models were applied for early-season classification in a separate target year, the models reached the F1-score threshold of 85% at least one month later than for the training years. Overall, the advantages of deep learning approaches (1D-CNN) could not be demonstrated, probably due to the limited amount of reference data available to train the models. However, very recent research into early-season crop classification and temporal transferability shows opportunities for the real-world application of deep learning classification models for catch crop classification.

Author Contributions: Conceptualization, K.V., S.H., E.B., and B.S.; methodology, K.V., S.H., and E.B.; software, K.V. and E.B.; validation, K.V.; formal analysis, K.V.; investigation, K.V.; data curation, K.V. and E.B.; writing—original draft preparation, K.V.; writing—review and editing, K.V., S.H., E.B., and B.S.; visualization, K.V.; supervision, S.H. and B.S.; project administration, S.H. and B.S.; funding acquisition, S.H. and B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research Foundation Flanders (FWO), grant number S006421N.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Acknowledgments: A special thanks to the Vlaamse Landmaatschappij (Flemish Land Agency) and the Agency for Agriculture and Fisheries for their help regarding data collection. Also, thanks to Zifu Wang for providing useful feedback regarding deep learning.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

For preprocessing the dataset from the department of Agriculture and Fisheries, we took a two-step approach. First, the useful observations were selected based on the label and the date of the observation. Second, the selected labels were checked with their corresponding time series to filter out any discrepancies. This was done because the field observations are taken at a single point in time, while fields are subject to change. For example, it could be that a field visit in October results in the label 'bare soil', while visiting the same field a few weeks later results in the presence of vegetation.

Two factors were taken into account for the selection of observations: (1) the observation date and (2) the relative occurrences of the land cover types in the reference dataset. First, most observations occurred in the period of October–January, which is also the period that most catch crops are required as a cover, and thus, the period that is under study here. Observations outside this period were compared with the crop calendar of the observed crop to determine if this crop can be on the field at the time of the field visit. Second, the observation could not be used if a winter crop type was only observed once. For model training and validation, reference data are split into separate training and validation sets stratified by crop species. This split requires a minimum of two observations per crop species. Also, one winter crop type, 'mixture of catch crops', was about ten times more abundant than the second most represented type. However, as a heavily unbalanced dataset is not beneficial for model training, only a subset of the parcels of this type was retained in the final reference dataset. Parcel selection within the most prevalent crop type depended on two principles: (1) parcels with a larger area were favored because it facilitates the next step of downloading the time series for these parcels, as it results in more usable pixels per field, and (2) only parcels with an agreement between the declared crop and observed crop (both were 'mixture of catch crops') were retained to maximize the reliability of the remaining training data. Observations of permanent land cover types, like pastures and greenhouses, were removed as no winter crops were found on these fields.

In addition to this, the field observation labels were compared with the corresponding daily NDVI time series. Because field observations are only a snapshot at a certain point in time and we worked with time series as input to the model, there was sometimes a disparity between the label and the time series. For example, if a field observation in October indicated that there was no catch crop present, the time series could still indicate a green-up of the field from November onwards. Therefore, we decided to trust the ground truth label unless the daily NDVI time series clearly disproved it. A crop observation was not used if, for the entire period of October–December, the NDVI value surpassed 0.35. Similarly, a bare soil observation was not used if, for the entire period of October–December, the NDVI value was higher than 0.35. We chose December as the endpoint for this because, in our application, there should still be time to visit the field after making a model prediction. Therefore, we trained the model on the examples that we wanted to detect, i.e., the field that contained a catch crop before the end of December, so that field visits can still be made in January.

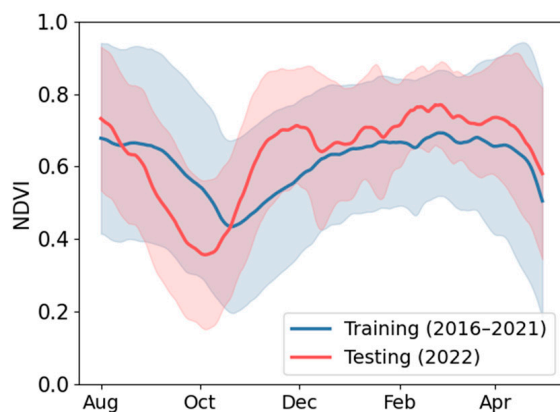
Appendix B

Table A1. Selected hyperparameters for all models per dataset.

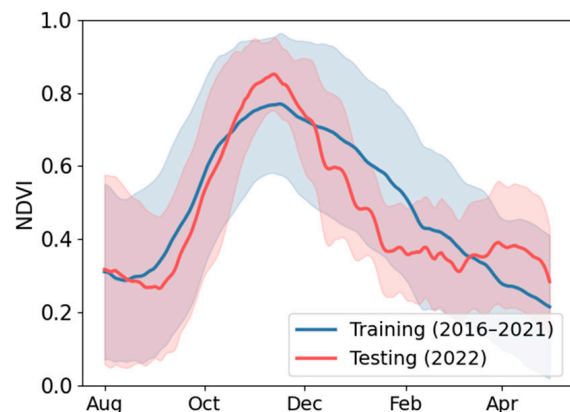
Model	Hyperparameters	Search Space	Selected Values (per Dataset)			
			S1 + 2 fAPAR + FCover + NDVI Daily	S1 + 2 NDVI Daily	S1 + 2 fAPAR + FCover + NDVI Dekadal	S1 + 2 NDVI Dekadal
RF	n_estimators	200, 300, 400, 500	500	400	400	500
TSF	n_estimators min_interval	200, 300, 400, 500 3, 5, 7, 9	300-400-200 3-3-3	300 3	400-200-300 9-3-5	500 5
CNN	batch_size	8, 16, 32, 64	32	32	32	64
	learning rate	0.001–0.1	0.042	0.042	0.062	0.022
	num_epochs	15, 20, 25, 30	20	20	25	15
	dropout	0.1, 0.2, 0.3,	0.1	0.1	0.1	0.1
	kernel_size	0.4	5	5	9	7
	stride	3, 5, 7, 9 1, 2, 3, 5	2	2	3	1
			S1 + 2 NDVI Monthly	S2 NDVI Dekadal	S2 9 Bands Dekadal	S2 9 Norm Dekadal
RF	n_estimators	200, 300, 400, 500	200	400	300	500
TSF	n_estimators min_interval	200, 300, 400, 500 3, 5, 7, 9	500 3 *	400 3	200-300-500- 200-400-200- 200-500-200 3-3-3 * 3-3-3 3-3-3	400-500-200- 200-500-300- 500-400-400 3-3-3 * 3-3-3 3-3-3
CNN	batch_size	8, 16, 32, 64	64	64	64	64
	learning rate	0.001–0.1	0.022	0.022	0.022	0.022
	num_epochs	15, 20, 25, 30	15	15	15	15
	dropout	0.1, 0.2, 0.3,	0.1	0.1	0.1	0.1
	kernel_size	0.4	7	7	7	7
	stride	3, 5, 7, 9 1, 2, 3, 5	1	1	1	1

* In the case of the dataset with nine bands from Sentinel-2, only the number of estimators was optimized because the search space was too big when the minimum interval length was included (monthly min interval was also not optimized—too short).

Appendix C



(A) Grass group



(B) Leafy group

Figure A1. Cont.

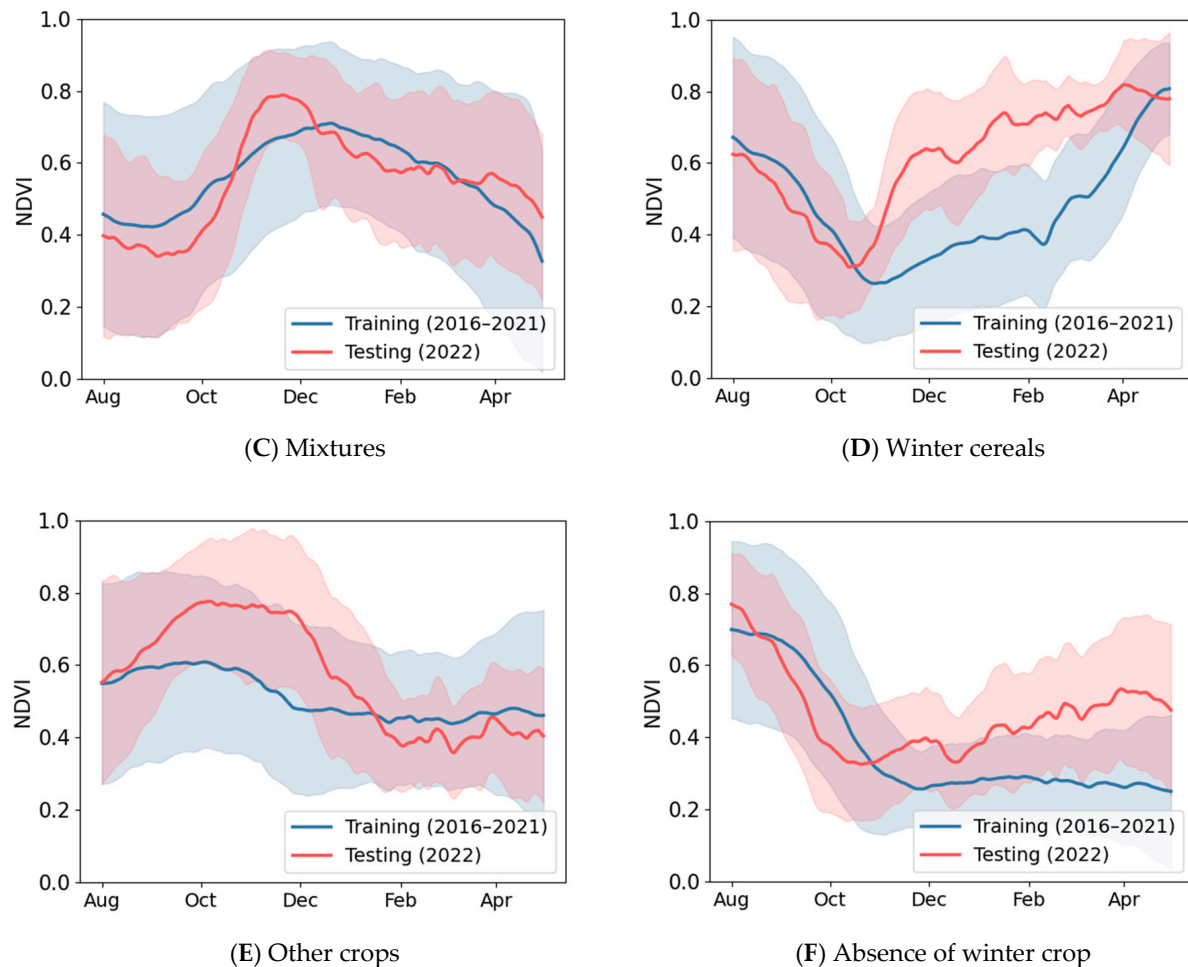


Figure A1. Comparison of the temporal profiles (NDVI, 1-day interval) of the different classification groups between the training data (2016–2021) and the validation data (2022). Shadowing indicates the standard deviation on each averaged profile. The group ‘Legumes’ is not visualized because of its absence in the validation dataset (2022).

References

1. Cameron, K.C.; Di, H.J.; Moir, J.L. Nitrogen Losses from the Soil/Plant System: A Review. *Ann. Appl. Biol.* **2013**, *162*, 145–173. [\[CrossRef\]](#)
2. Fendrich, A.N.; Matthews, F.; Van Eynde, E.; Carozzi, M.; Li, Z.; d’Andrimont, R.; Lugato, E.; Martin, P.; Ciais, P.; Panagos, P. From Regional to Parcel Scale: A High-Resolution Map of Cover Crops across Europe Combining Satellite Data with Statistical Surveys. *Sci. Total Environ.* **2023**, *873*, 162300. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Schulz, C.; Holtgrave, A.-K.; Kleinschmit, B. Large-Scale Winter Catch Crop Monitoring with Sentinel-2 Time Series and Machine Learning—An Alternative to on-Site Controls? *Comput. Electron. Agric.* **2021**, *186*, 106173. [\[CrossRef\]](#)
4. Nouri, A.; Lukas, S.; Singh, S.; Singh, S.; Machado, S. When Do Cover Crops Reduce Nitrate Leaching? A Global Meta-Analysis. *Glob. Chang. Biol.* **2022**, *28*, 4736–4749. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Blanco-Canqui, H. Cover Crops and Water Quality. *Agron. J.* **2018**, *110*, 1633–1647. [\[CrossRef\]](#)
6. Sitokonstantinou, V.; Papoutsis, I.; Kontoes, C.; Lafarga Arnal, A.; Armesto Andrés, A.P.; Garraza Zurbano, J.A. Scalable Parcel-Based Crop Identification Scheme Using Sentinel-2 Data Time-Series for the Monitoring of the Common Agricultural Policy. *Remote Sens.* **2018**, *10*, 911. [\[CrossRef\]](#)
7. Barnes, M.L.; Yoder, L.; Khodaei, M. Detecting Winter Cover Crops and Crop Residues in the Midwest US Using Machine Learning Classification of Thermal and Optical Imagery. *Remote Sens.* **2021**, *13*, 1998. [\[CrossRef\]](#)
8. Jennewein, J.S.; Lamb, B.T.; Hively, W.D.; Thieme, A.; Thapa, R.; Goldsmith, A.; Mirsky, S.B. Integration of Satellite-Based Optical and Synthetic Aperture Radar Imagery to Estimate Winter Cover Crop Performance in Cereal Grasses. *Remote Sens.* **2022**, *14*, 2077. [\[CrossRef\]](#)

9. Fan, X.; Vrieling, A.; Muller, B.; Nelson, A. Winter Cover Crops in Dutch Maize Fields: Variability in Quality and Its Drivers Assessed from Multi-Temporal Sentinel-2 Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *91*, 102139. [\[CrossRef\]](#)
10. Thieme, A.; Yadav, S.; Oddo, P.C.; Fitz, J.M.; McCartney, S.; King, L.; Keppler, J.; McCarty, G.W.; Hively, W.D. Using NASA Earth Observations and Google Earth Engine to Map Winter Cover Crop Conservation Performance in the Chesapeake Bay Watershed. *Remote Sens. Environ.* **2020**, *248*, 111943. [\[CrossRef\]](#)
11. Goffart, D.; Curnel, Y.; Planchon, V.; Goffart, J.-P.; Defourny, P. Field-Scale Assessment of Belgian Winter Cover Crops Biomass Based on Sentinel-2 Data. *Eur. J. Agron.* **2021**, *126*, 126278. [\[CrossRef\]](#)
12. Kc, K.; Zhao, K.; Romanko, M.; Khanal, S. Assessment of the Spatial and Temporal Patterns of Cover Crops Using Remote Sensing. *Remote Sens.* **2021**, *13*, 2689. [\[CrossRef\]](#)
13. Pan, L.; Xia, H.; Zhao, X.; Guo, Y.; Qin, Y. Mapping Winter Crops Using a Phenology Algorithm, Time-Series Sentinel-2 and Landsat-7/8 Images, and Google Earth Engine. *Remote Sens.* **2021**, *13*, 2510. [\[CrossRef\]](#)
14. Cai, Y.; Guan, K.; Peng, J.; Wang, S.; Seifert, C.; Wardlow, B.; Li, Z. A High-Performance and in-Season Classification System of Field-Level Crop Types Using Time-Series Landsat Data and a Machine Learning Approach. *Remote Sens. Environ.* **2018**, *210*, 35–47. [\[CrossRef\]](#)
15. Zhong, L.; Hu, L.; Zhou, H. Deep Learning Based Multi-Temporal Crop Classification. *Remote Sens. Environ.* **2019**, *221*, 430–443. [\[CrossRef\]](#)
16. Rußwurm, M.; Körner, M. Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 129. [\[CrossRef\]](#)
17. Pelletier, C.; Webb, G.I.; Petitjean, F. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sens.* **2019**, *11*, 523. [\[CrossRef\]](#)
18. Van Tricht, K.; Gobin, A.; Gilliams, S.; Piccard, I. Synergistic Use of Radar Sentinel-1 and Optical Sentinel-2 Imagery for Crop Mapping: A Case Study for Belgium. *Remote Sens.* **2018**, *10*, 1642. [\[CrossRef\]](#)
19. Yi, Z.; Jia, L.; Chen, Q.; Jiang, M.; Zhou, D.; Zeng, Y. Early-Season Crop Identification in the Shiyang River Basin Using a Deep Learning Algorithm and Time-Series Sentinel-2 Data. *Remote Sens.* **2022**, *14*, 5625. [\[CrossRef\]](#)
20. Belgiu, M.; Drăguţ, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [\[CrossRef\]](#)
21. Mellor, A.; Boukir, S.; Haywood, A.; Jones, S. Exploring Issues of Training Data Imbalance and Mislabelling on Random Forest Performance for Large Area Land Cover Classification Using the Ensemble Margin. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 155–168. [\[CrossRef\]](#)
22. Deng, H.; Runger, G.; Tuv, E.; Vladimir, M. A Time Series Forest for Classification and Feature Extraction. *Inf. Sci.* **2013**, *239*, 142–153. [\[CrossRef\]](#)
23. Rußwurm, M.; Körner, M. Self-Attention for Raw Optical Satellite Time Series Classification. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 421–435. [\[CrossRef\]](#)
24. Ahmed, Z.; Nalley, L.; Brye, K.; Steven Green, V.; Popp, M.; Shew, A.M.; Connor, L. Winter-Time Cover Crop Identification: A Remote Sensing-Based Methodological Framework for New and Rapid Data Generation. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *125*, 103564. [\[CrossRef\]](#)
25. Henits, L.; Szerletics, Á.; Szokol, D.; Szlovák, G.; Gojdár, E.; Zlinszky, A. Sentinel-2 Enables Nationwide Monitoring of Single Area Payment Scheme and Greening Agricultural Subsidies in Hungary. *Remote Sens.* **2022**, *14*, 3917. [\[CrossRef\]](#)
26. Tao, Y.; You, F. Prediction of Cover Crop Adoption through Machine Learning Models Using Satellite-Derived Data. *IFAC-PapersOnLine* **2019**, *52*, 137–142. [\[CrossRef\]](#)
27. Simon Sanchez, A.-M.; Gonzalez-Piqueras, J.; de la Ossa, L.; Calera, A. Convolutional Neural Networks for Agricultural Land Use Classification from Sentinel-2 Image Time Series. *Remote Sens.* **2022**, *14*, 5373. [\[CrossRef\]](#)
28. Račić, M.; Oštir, K.; Zupanc, A.; Čehovin Zajc, L. Multi-Year Time Series Transfer Learning: Application of Early Crop Classification. *Remote Sens.* **2024**, *16*, 270. [\[CrossRef\]](#)
29. Weilandt, F.; Behling, R.; Goncalves, R.; Madadi, A.; Richter, L.; Sanona, T.; Spengler, D.; Welsch, J. Early Crop Classification via Multi-Modal Satellite Data Fusion and Temporal Attention. *Remote Sens.* **2023**, *15*, 799. [\[CrossRef\]](#)
30. Pandžić, M.; Pavlović, D.; Matavulj, P.; Brdar, S.; Marko, O.; Crnojević, V.; Kilibarda, M. Interseasonal Transfer Learning for Crop Mapping Using Sentinel-1 Data. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103718. [\[CrossRef\]](#)
31. Lin, C.; Zhong, L.; Song, X.-P.; Dong, J.; Lobell, D.B.; Jin, Z. Early- and in-Season Crop Type Mapping without Current-Year Ground Truth: Generating Labels from Historical Information via a Topology-Based Approach. *Remote Sens. Environ.* **2022**, *274*, 112994. [\[CrossRef\]](#)
32. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution. *Sci. Data* **2018**, *5*, 180214. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Statistics Flanders Landbouwareaal. Available online: <https://www.vlaanderen.be/statistiek-vlaanderen/landbouw-en-visserij/landbouwareaal> (accessed on 18 July 2024).
34. Hermans, I.; Elsen, A.; Bries, J. *Groenbemesters En Nitraatresidu*; Bodemkundige Dienst van België vzw: Heverlee, Belgium, 2010; p. 44.

35. Vandendriessche, H.; Vanongeval, L.; Smeets, E.; Geypens, M. Monitoring of N-Uptake by Green Manures and of the Influence of N-Release on N-Availability, Production and Quality of Sugar Beet. In *Progress in Nitrogen Cycling Studies: Proceedings of the 8th Nitrogen Workshop Held at the University of Ghent, 5–8 September 1994*; Van Cleemput, O., Hofman, G., Vermoesen, A., Eds.; Developments in Plant and Soil Sciences; Springer: Dordrecht, The Netherlands, 1996; pp. 147–151; ISBN 978-94-011-5450-5.
36. EPPO. EPPO Global Database. 2024. Available online: <https://gd.eppo.int/> (accessed on 18 July 2024).
37. Waśniewski, A.; Hościło, A.; Chmielewska, M. Can a Hierarchical Classification of Sentinel-2 Data Improve Land Cover Mapping? *Remote Sens.* **2022**, *14*, 989. [CrossRef]
38. Xie, G.; Niculescu, S. Mapping Crop Types Using Sentinel-2 Data Machine Learning and Monitoring Crop Phenology with Sentinel-1 Backscatter Time Series in Pays de Brest, Brittany, France. *Remote Sens.* **2022**, *14*, 4437. [CrossRef]
39. Turkoglu, M.O.; D’Aronco, S.; Perich, G.; Liebisch, F.; Streit, C.; Schindler, K.; Wegner, J.D. Crop Mapping from Image Time Series: Deep Learning with Multi-Scale Label Hierarchies. *Remote Sens. Environ.* **2021**, *264*, 112603. [CrossRef]
40. European Space Agency. S2 Mission. Available online: <https://sentiwiki.copernicus.eu/web/s2-mission> (accessed on 18 July 2024).
41. Biesemans, J. Supporting Land and Infrastructure Management with CropSAR. Available online: <https://vito.be/en/news/supporting-land-and-infrastructure-management-cropsar> (accessed on 18 July 2024).
42. Van Tricht, K. Looking Through the Clouds to Improve Crop Monitoring! Available online: <https://blog.vito.be/remotesensing/cropsar2019> (accessed on 18 July 2024).
43. Piccard, I.; Swinnen, E.; Keukelaere, L.D.; Kerchove, R.V.D.; Eerens, H. Terrascope Sentinel-2 Algorithm Theoretical Base Document S2—NDVI & BIOPAR—V102. 2020. Available online: <https://tinyurl.com/4mru98y5> (accessed on 12 April 2023).
44. Schramm, M.; Pebesma, E.; Milenković, M.; Foresta, L.; Dries, J.; Jacob, A.; Wagner, W.; Mohr, M.; Neteler, M.; Kadunc, M.; et al. The openEO API—Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. *Remote Sens.* **2021**, *13*, 1125. [CrossRef]
45. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
46. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
47. van Rossum, G. *Python Reference Manual*; Centrum voor Wiskunde en Informatica: Amsterdam, The Netherlands, 1995.
48. Király, F.; Löning, M.; Bagnall, T.; Middlehurst, M.; Ray, A.; Ganesh, S.; Walter, M.; Oastler, G.; Lines, J.; ViktorKaz; et al. sktime/sktime. 2020. Available online: <https://zenodo.org/records/14337818> (accessed on 18 July 2024).
49. Löning, M.; Bagnall, A.; Ganesh, S.; Kazakov, V. Sktime: A Unified Interface for Machine Learning with Time Series. *arXiv* **2019**, arXiv:1909.07872.
50. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
51. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep Learning for Time Series Classification: A Review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
52. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature*. **2015**, *521*, 436–444. [CrossRef] [PubMed]
53. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
54. Chollet, F. *Deep Learning with Python*; Manning Publications Co.: Shelter Island, NY, USA, 2018; ISBN 978-1-61729-443-3.
55. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv* **2018**, arXiv:1807.05118.
56. Bagnall, A.; Lines, J.; Bostrom, A.; Large, J.; Keogh, E. The Great Time Series Classification Bake off: A Review and Experimental Evaluation of Recent Algorithmic Advances. *Data Min. Knowl. Discov.* **2017**, *31*, 606–660. [CrossRef]
57. Pham, V.-D.; Tetteh, G.; Thiel, F.; Erasmi, S.; Schwieder, M.; Frantz, D.; Van Der Linden, S. Temporally Transferable Crop Mapping with Temporal Encoding and Deep Learning Augmentations. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103867. [CrossRef]
58. Holzhauser, K.; Rübiger, T.; Rose, T.; Kage, H.; Kühling, I. Estimation of Biomass and N Uptake in Different Winter Cover Crops from UAV-Based Multispectral Canopy Reflectance Data. *Remote Sens.* **2022**, *14*, 4525. [CrossRef]
59. Wang, X.; Blesh, J.; Rao, P.; Paliwal, A.; Umashaanker, M.; Jain, M. Mapping Cover Crop Species in Southeastern Michigan Using Sentinel-2 Satellite Data and Google Earth Engine. *Front. Artif. Intell.* **2023**, *6*, 1035502. [CrossRef]
60. Huang, X.; Huang, J.; Li, X.; Shen, Q.; Chen, Z. Early Mapping of Winter Wheat in Henan Province of China Using Time Series of Sentinel-2 Data. *GISci. Remote Sens.* **2022**, *59*, 1534–1549. [CrossRef]
61. Hughes, G. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Trans. Inf. Theory* **1968**, *14*, 55–63. [CrossRef]
62. Orynbaikyzy, A.; Gessner, U.; Conrad, C. Crop Type Classification Using a Combination of Optical and Radar Remote Sensing Data: A Review. *Int. J. Remote Sens.* **2019**, *40*, 6553–6595. [CrossRef]
63. Selvaraj, S.; Bargiel, D.; Htitiou, A.; Gerighausen, H. Analyzing Temporal Characteristics of Winter Catch Crops Using Sentinel-1 Time Series. *Remote Sens.* **2024**, *16*, 3737. [CrossRef]
64. Royal Meteorological Institute. Klimatologisch Jaaroverzicht 2021. Available online: https://www.meteo.be/resources/climatology/pdf/klimatologisch_jaaroverzicht_2021.pdf (accessed on 17 July 2024).
65. Royal Meteorological Institute. Klimatologisch Jaaroverzicht 2022. Available online: https://www.meteo.be/resources/climatology/pdf/klimatologisch_jaaroverzicht_2022.pdf (accessed on 17 July 2024).

66. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.U.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Vancouver, BC, Canada, 2017; Volume 30.
67. Ma, Y.; Chen, S.; Ermon, S.; Lobell, D.B. Transfer Learning in Environmental Remote Sensing. *Remote Sens. Environ.* **2024**, *301*, 113924. [[CrossRef](#)]
68. Tseng, G.; Cartuyvels, R.; Zvonkov, I.; Purohit, M.; Rolnick, D.; Kerner, H. Lightweight, Pre-Trained Transformers for Remote Sensing Timeseries. *arXiv* **2024**, arXiv:2304.14065.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.