End-to-end learned early classification of time series for in-season crop type mapping

Marc Rußwurm, Nicolas Courty, Rémi Emonet, Sébastien Lefèvre, Devis Tuia, Romain Tavenard

End-to-End Learned Early Classification of Time Series for In-Season Crop Type Mapping

Marc Rußwurm[1], Nicolas Courty[2], Rémi Emonet[3], Sébastien Lefèvre[2], Devis Tuia[1], and Romain Tavenard[4]

[1] *Environmental Computer Science and Earth Observation Laboratory (ECEO), École Polytechnique Fédérale de Lausanne (EPFL)*
[2] *Univ. Bretagne Sud, CNRS, IRISA*
[3] *Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023*
[4] *Univ. Rennes, CNRS, LETG/IRISA*

Abstract

Remote sensing satellites capture the cyclic dynamics of our Planet in regular time intervals recorded in satellite time series data. End-to-end trained deep learning models use this time series data to make predictions at a large scale, for instance, to produce up-to-date crop cover maps. Most time series classification approaches focus on the accuracy of predictions. However, the earliness of the prediction is also of great importance since coming to an early decision can make a crucial difference in time-sensitive applications.

In this work, we present an End-to-End Learned Early Classification of Time Series (ELECTS) model that estimates a classification score and a probability of whether sufficient data has been observed to come to an early and still accurate decision. ELECTS is modular: any deep time series classification model can adopt the ELECTS conceptual idea by adding a second prediction head that outputs a probability of stopping the classification. The ELECTS loss function then optimizes the overall model on a balanced objective of earliness and accuracy. Our experiments on four crop classification datasets from Europe and Africa show that ELECTS allows reaching state-of-the-art accuracy while reducing the quantity of data massively to be downloaded, stored, and processed. The source code is available at `https://github.com/marccoru/elects`.

1. Introduction

Efficient large-scale agricultural monitoring and crop type mapping is a prime example of time series analysis in Earth observation: analyzing the temporal variation of vegetation during a growing season is crucial for efficient and accurate predictions. Models and algorithms trained from satellite time series can distinguish different crop types by observing differences in their respective phenology (life cycles). Traditionally, NDVI-based temporal profiles [1, 2] are used to extract a fixed set of hand-defined features, such as the date of the green-up, or senescence phases [2]. Remote sensing experts often manually choose the observation period in these approaches to capture the entire vegetative period of the crops in a particular region. The final classification is executed once at the end of this period to produce a crop cover map. Early time series classification has been a steady topic of interest in remote sensing but is often seen as an auxiliary objective. In crop type classification, the terms *in-season-* or *early crop type mapping* are commonly used. Several studies [3, 4, 5, 6] found that a high classification accuracy for most crop types is achievable within the growing season in a specific region. A common strategy for assessing which accuracy is possible at what day of the year is *incremental classification*, as termed by Inglada *et al.*, [5]: a supervised classifier performs a classification every time a new image becomes available. The achievable accuracy is then recorded and related to the length of the sequence. This process involves re-fitting the classifier for different sub-sequences and provides region-specific evidence across all crop types regarding the date at which an accurate classification is possible. Recent works have applied incremental classification for early crop type mapping in Germany, [7, 8] and South Africa [9]. Other approaches avoid re-fitting the classifier by choosing sequence-length invariant features [10], employing a cluster-then-labeling strategy [11], or modeling simplified two-dimensional feature space in a generative way from historical data [12]. These approaches employ increasingly sophisticated heuristics to hand-define features invariant to the sequence length. Crucially, these approaches yield a rough general knowledge of achievable accuracy given a specific day of the year for a single region across multiple crop types. Meanwhile, end-to-end deep learning architectures based on recurrence [13], self-attention [14], or convolution [15] can map a variable length series into a fixed-length representation natively. These deep neural networks learn class-discriminative features solely from a large dataset of labeled samples in an end-to-end scheme by minimizing classification error as the objective function. To our knowledge, no approach has explicitly optimized a

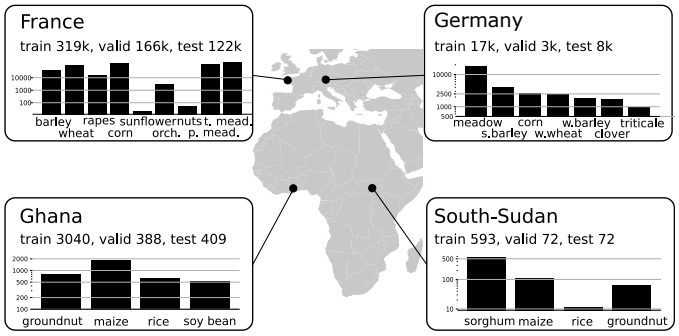


Figure 1: Overview of datasets used in this work. All datasets show a label imbalance, as dominant crops are common in the respective areas. The European datasets BreizhCrops (France) and BavarianCrops (Germany) provide large-scale data with several tens to hundreds of thousands of time series samples. In contrast, the African datasets are smaller and contain a few hundred to a few thousand crop parcels.

model for the objective of an early classification in remote sensing.

In this work, we address this research gap by End-to-end Learned Early Classification of Time Series (ELECTS). Our method provides early and accurate predictions for each field parcel. To do so, we use a neural network with a loss function optimizing for both objectives: earliness and accuracy.

ELECTS augments and is compatible with recent advances in end-to-end trainable deep time series classification models [13, 14, 15]. As these models produce a fixed-size vector from a variable-length sequence, it does not have to re-fit the classifier on shorter sub-sequences, as earlier incremental classification approaches did [3, 4, 5, 6]. Optimizing on the joint loss objective of earliness and accuracy is also conceptually more straightforward compared to the cluster-then-labeling heuristic of Konduri *et al.*, (2020) [11] or modeling transitions with two-dimensional distributions from historical data, as Lin *et al.*, (2022) [12].

2. Datasets

We evaluate ELECTS on four crop-type mapping datasets. Annotations of two datasets originate from crop type statistics collected in Europe and are available at a large scale with several tens of thousand of samples. The annotations of the two datasets in Africa originate from small-scale surveys and contain only hundreds to few thousand annotated time series samples. Figure 1 summarizes the crop-type datasets used in this work. It shows the locations and the label distribution of four crop datasets in Europe and Africa.

2.1. BreizhCrops (France)

We use the BreizhCrops dataset [16] to compare the LSTM model of Section 3.1 with several other regular classification models. BreizhCrops contains time series of 608 263 field parcels of the year 2017 in Brittany, France. The time series contains all Sentinel-2 images from January to December. Both datasets typically contain between 71 (every 5 days) and 147 (every 2.5 days) Sentinel-2 observations. The high acquisition frequency of 2.5 days and 147 observations is possible for some fields in the overlap area of two acquisition stripes. The BreizhCrops dataset [16] is split regionally into training (FRH01, FRH02; 319 258 fields), validation (FRH03; 166 391 fields), and test (FRH04; 122 614) partitions, where FRH$\{1, 2, 3, 4\}$ refers to NUTS-3 administrative boundaries. The *Nomenclature des unités territoriales statistiques* (NUTS) system delineates Europe in administrative boundaries at three levels: country, state, and province. BreizhCrops uses the division at the provincial level NUTS-3. The dataset contains nine crop classes: BARLEY, WHEAT, RAPESEED, CORN, SUNFLOWER, ORCHARDS, NUTS, PERMANENT MEADOWS, TEMPORARY MEADOWS. They are selected to contain both frequent (BARLEY, WHEAT) and rare classes (SUNFLOWER, NUTS), as well as semantically similar categories (PERMANENT- and TEMPORARY MEADOWS).

2.2. BavarianCrops (Germany)

We performed ablation studies and the comparison to one method from the early time-series community (SR2-CF2[17] in Appendix B.1) on a crop type dataset near Hollfeld in Bavaria, Germany, which is a subset of the dataset used in [18]. We chose to subset the original dataset for computational reasons in the initial development and to compare it with existing early time series classification approaches that are typically not designed for large-scale datasets. Our subset of BavarianCrops covers a 40km $\times$ 35km area and contains 27 470 fields that are split into training (16 600), validation (3057), and test (7813) partitions, each one organised in blocks of 4.5km $\times$ 4.5km with 500 meter margin between the blocks. All parcels within one block are assigned to the same train-val-test partition to avoid assigning neighboring fields to different partitions [19]. Sentinel-2 scenes with same frequency as Breizhcrops alongside associated labels are from January to December 2018 and cover the 7 common crops MEADOW, SUMMER BARLEY, CORN, WINTER WHEAT, WINTER BARLEY, CLOVER, TRITICALE.

2.3. Ghana and South Sudan

Rustowicz *et al.* [20] compiled the datasets of Ghana, and South Sudan that were incorporated in the SustainBench dataset [21]. They share a common processing history and are described together in this section. In these middle- and low-income countries, a substantial portion of the population directly depends on agriculture. An early estimate of the expected crop yield is crucial to evaluate the economic markets and uncover potential shortages. This dataset provides Sentinel-2, Sentinel-1, and PlanetScope images of size 64 by 64 pixels from the years 2016 and 2017 linearly interpolated to a time series of 365 days. We take the imagery and field boundaries and average all

pixels belonging to each field to obtain a time series. Following [20], the ten Sentinel-2 bands (10m and 20m channels) and NDVI and green chlorophyll vegetation index (GCVI) features are combined with three Sentinel-1 bands (VV, VH, and their ratio) and four PlanetScope bands (RGB+NIR), which results in a 19-dimensional feature vector for each field parcel. While the training and validation datasets were taken from 2016, the samples of the test dataset were taken from the subsequent year 2017. The crop types classified in Ghana are GROUNDNUT, MAIZE, RICE, and SOY BEAN, while information on SORGHUM, MAIZE, RICE, and GROUNDNUT are available in South Sudan.

3. Methodology

This section describes the details of the proposed method. It consists of a deep learning feature extractor with two decision heads, detailed in Section 3.1 and a loss function that optimizes for the dual objective of accuracy and earliness outlined in Section 3.2. Throughout this section, we denote vectors with bold-faced symbols, while matrices are bold-faced and capitalized. In a time series, we use t to indicate any time step, while T refers specifically to the index of the last time step in the sequence, i.e., the sequence length. Figure 2 shows a schematic view of the model and loss functions with the associated equations of this section.

3.1. Model

We use a time series classification model that consists of i) a deep feature extractor based on recursion, f_θ, that ingests time series data one observation at a time and ii) two output heads. This model can be implemented with different deep learning architectures, but we focus on recurrent neural networks (RNNs) without loss of generality. RNNs estimate a hidden representation $\mathbf{h}_t$ at a given time t from an input time series $\mathbf{X}_{\to t} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)$ of observations $\mathbf{x}$ up to the image acquisition at time t. The model can process a variable number of samples and ingest time series with different sequence lengths T. A recurrent neural network

$$\mathbf{h}_t = f_{\theta_h}(\mathbf{x}_t, \mathbf{h}_{t-1}) \tag{1}$$

updates its zero-initialized hidden representation $\mathbf{h}_{t-1}$ to $\mathbf{h}_t$ with each new observation $\mathbf{x}_t$. It is a natural choice as a feature extractor, as it projects a variable-length input sequence to a fixed-size representation. In practice, to avoid vanishing gradients [22, 23], we choose a Long Short-Term Memory (LSTM) [24] recurrent neural network $\{\mathbf{h}_t, \mathbf{c}_t\} = f_{\theta_h}(\mathbf{x}_t, \{\mathbf{h}_{t-1}, \mathbf{c}_{t-1}\})$ that updates two hidden representations where we use the cell output $\mathbf{h}_t$ for two linear decision heads: one head produces a classification probability for each class

$$\hat{\mathbf{y}}_t = \text{softmax}\left(f_{\theta_c}(\mathbf{h}_t)\right) \tag{2}$$

and another one outputs a scalar probability of stopping

$$d_t = \sigma\left(f_{\theta_d}(\mathbf{h}_t)\right) \tag{3}$$

the classification decision. The σ symbol denotes the sigmoid function that rescales the outputs of the stopping head to a probability between 0 and 1. At test time, a hard stopping decision is sampled from this stopping probability. As an example: with a stopping probability $d_t = 0.2$, the classification is stopped with a 20% probability at this time. In practice (see Fig. 3 in results), we observe that d_t raises sharply from 0 to 1 within a few time points on a trained model.

3.2. ELECTS loss function

At each time $t \leq T$, we compute the classification, earliness-rewarded loss, L_{CER}:

$$L_{\text{CER}}(\hat{\mathbf{y}}_t, \mathbf{y}) = \alpha L_c(\hat{\mathbf{y}}_t, \mathbf{y}) - (1-\alpha) R_e(\hat{\mathbf{y}}_t, \mathbf{y}, t). \tag{4}$$

We weight both terms with an $\alpha \in [0, 1]$ hyper-parameter that trades off accuracy and earliness reward. The classification loss is the negative log-likelihood or cross-entropy loss

$$L_c(\hat{\mathbf{y}}_t, \mathbf{y}) = -\sum_{c=1}^{C} y_c \log \hat{y}_{c,t}, \tag{5}$$

while the earliness reward is

$$R_e(\hat{\mathbf{y}}_t, \mathbf{y}, t) = \hat{y}_t^+ \left(\frac{T-t}{T}\right). \tag{6}$$

As such, R_e decreases linearly for later predictions when t approaches T. This term is scaled with the probability of the correct class $y_t^+ = \sum_{c=1}^{C} y_c \hat{y}_{c,t}$ with $\mathbf{y}$ as one-hot vector of C classes. This term applies the reward only if the probability of the correct class is large.

L_{ELECTS} is computed for each time t in a training sample time series of length T to minimize a joint expression of accuracy (via L_{CER}) and explicit earliness (via D_t) as:

$$L_{\text{ELECTS}}(\hat{\mathbf{d}}_{\to t}, \hat{\mathbf{y}}_t, \mathbf{y}) = D_t(\hat{\mathbf{d}}_{\to t}) L_{\text{CER}}(\hat{\mathbf{y}}_t, \mathbf{y}) \tag{7}$$

where

$$D(\hat{\mathbf{d}}_{\to t}) = \hat{d}_t \prod_{i=1}^{t-1}(1-\hat{d}_i) + \frac{\varepsilon}{T} \tag{8}$$

can be interpreted as the joint probability of making a decision $\hat{d}_t$ at time t and not having made a decision before $\prod_{i=1}^{t-1}(1-\hat{d}_i)$. At the last time step, we set $\hat{d}_T = 1$ irrespectively of the model output to make sure that the model has taken a stopping decision in the interval $[0, T]$. In practice, we add a small constant offset $\frac{\varepsilon}{T}$ to each $\hat{d}_t$, with ε as an hyper-parameter. This offset makes $D(\hat{\mathbf{d}}_{\to t})$ non-zero for all t, which encourages the model to make accurate classifications for all time steps in Eq. (7). With $\varepsilon = 0$, only

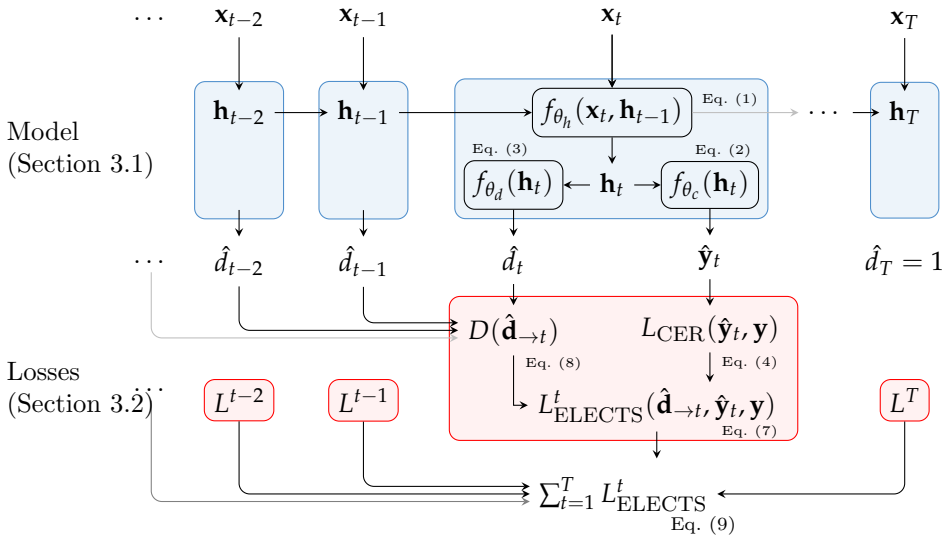


Figure 2: Schematic illustration of model (blue) and losses (red) of this Section 3. Arrows indicate function inputs and outputs. Neural network components are denoted by f_θ. At test time, with fixed weights, θ, the model (blue) can process a time series up to any time t. At training time, losses are calculated on the complete time series until the last time step T.

accurate classifications at the time steps close to the stopping time, where $D(\hat{\mathbf{d}}_{\to t})$ is large, would be encouraged. Without this offset, i.e., with $\varepsilon = 0$, we found experimentally (see Appendix A) that a randomly-initialized model tended to fall in a local minimum when optimizing Eq. (7) by predicting early at low accuracy.

The learnable parameters $\theta_h, \theta_d, \theta_c$ are determined by minimizing the overall objective

$$\underset{\theta_h,\theta_d,\theta_c}{\operatorname{argmin}} \sum_{\mathbf{X},\mathbf{y}} \sum_{t=1}^{T} L_{\text{ELECTS}}(\underbrace{f(\mathbf{X};\theta_h,\theta_d,\theta_c)}_{\hat{\mathbf{d}}_{\to t},\hat{\mathbf{y}}_t}, \mathbf{y}) \tag{9}$$

for each time t over a dataset of labeled samples $\mathbf{X}, \mathbf{y}$.

3.3. Implementation Details

For all results described in the Section 4, we used a recurrent neural network with the same hyperparameters for all datasets. An initial linear layer (and layer normalization) projects the original input vector to a learned 32-dimensional feature representation at each time, followed by two mono-directional LSTM layers. We implement each decision head as a linear layer with a sigmoid activation function for the stopping decision and softmax for the classification scores, respectively. The overall model has 67 108 trainable parameters, making this implementation light weighted and trainable in any desktop machine with a GPU graphics card. As stated above, researchers can implement the ELECTS loss on any neural network for time series making it adaptable for other time series approaches. We used the Adam optimizer with a learning rate of 0.001 and a dropout of 20%. We determined these hyperparameters experimentally on the validation set of the BavarianCrops dataset (described in the next section). With a batch size of 256, we trained models in a few minutes (BavarianCrops) or a few hours (BreizhCrops) on a GeForce RTX 3090. For BavarianCrops and BreizhCrops, we randomly choose sequences of 70 observations from the originally longer complete time series to obtain sequences of equal length for training in batches. At test time, we can run inference on the complete variable-length time series. For the Ghana and South Sudan datasets, we train on the interpolated 365-day sequences, similar to [20]. We used an $\varepsilon = 10$ offset parameter throughout the experiments with a fixed sequence length of the respective training dataset.

3.4. Model Evaluation

We evaluate the model on the four different crop type mapping datasets in Europe and Africa, described in Section 2. We train, validate, and evaluate the model for each dataset on spatially disjoint training, validation, and test regions. For BavarianCrops, training and evaluation fields were separated by blocks, while different administrative boundaries were used in BreizhCrops. For Ghana and South Sudan, we followed the split of Rustowicz *et al.*, (2019) [20]. For each dataset, we re-train the ELECTS-LSTM model from scratch on the respective training dataset and evaluate the performance on the test set. We do not vary the hyper-parameters (network layers, hidden dimensions, learning rates) across the datasets in these experiments and keep the identical model architecture throughout this work.

4. Results

This section presents the results obtained with the ELECTS-trained LSTM neural network described in Section 3. We

structure this section in three parts: First, Section 4.1 shows the prediction process on individual field parcels qualitatively and quantitatively. Section 4.2 focuses on the dates of stopped decisions and relates these to phenological events. It provides interpretations of the model predictions on two crop classes (RAPESEED and BARLEY). For these experiments, we used the large BreizhCrops dataset in Sections 4.1 and 4.2. Finally, we expand the scope in Section 4.3, where we train the ELECTS recurrent neural network on multiple datasets in Europe (BreizhCrops, BavarianCrops) and Africa (Ghana, South Sudan), as outlined in Section 2. Further model comparisons developed in the time series community and ablations on the loss design on the BavarianCrops dataset are reported in Appendix A and Appendix B, respectively.

4.1. Accuracy Evaluation

Sections 4.1.1 and 4.1.2 illustrate the prediction process, while Section 4.1.3, analyzes the classification accuracy of the stopped fields quantitatively throughout the year on all field parcels in the BreizhCrops test set.

4.1.1. Single Field Prediction

Figure 3 illustrates the prediction process with the ELECTS-trained LSTM on a single time series sample from the BreizhCrops test set. The time series of this TEMPORARY MEADOW field is represented on the left as its NDVI profile. However, note that our model uses thirteen spectral bands' complete signal at each observation. This profile shows that this field parcel is photosynthetically active (high NDVI) across the year. These high-NDVI observations in this time series are interrupted by negative outliers caused by cloud cover (low NDVI). The ELECTS-trained LSTM neural network ingests this time series one-time step at a time and estimates a probability for each crop class (top right) and a probability of stopping (bottom right). The model estimates a high probability for the class WHEAT (orange) during the first hundred days of the year. If the model stopped the classification decision this early, it would incorrectly predict the class WHEAT. Further, during the first hundred days, the probability of stopping remains low, indicating that more data is necessary for a confident decision. From the day of year 100 onwards, the model assigns the highest classification probability to the correct class TEMPORARY MEADOW. The probability of stopping remains low until day-of-year 150 when a rapid increase indicates that the model is sufficiently confident to stop the classification.

4.1.2. Classification of Fields at Different Times

Figure 4 shows a crop cover map of 250 field parcels from a 2.5km $\times$ 2.5km area of interest within Brittany, France, from the BreizhCrops test set. It illustrates the prediction process in the deployment setting where the predictions of some fields are stopped at different times compared to others in the same geographic area. The top row presents RGB images from this area alongside the ground truth crop type. The other rows show the model predictions at each date (second row), the correct/incorrect predictions (third row, with blue being correct and red incorrect), and the active (white) vs. stopped (black) status of each parcel. The rightmost column shows the predictions after recombining all predictions obtained at the respective stopping time (second and third row) and a summary of the per parcel stopping date. For ELECTS, only a stopped field (black) classification decision is relevant, as active fields (shown in white) require more data. In rows two and three's prediction and correctness figures, we present parcels still active in the decision process with transparent colors. The field parcels where the model stopped the classification process are drawn with opaque colors without transparency.

On April 12th (first column in Fig. 4), most parcels were covered homogeneously with green vegetation. The model predicted most fields as TEMPORARY MEADOW and CORN, among the dataset's most frequent classes. The overall accuracy for these parcels was 54%. These early, incorrect classifications (red) are frequent, as not enough time series could be observed this early in the year. The ELECTS-LSTM model did not stop any fields at this point (no black fields in the bottom row). More time steps are required for a confident classification decision. On May 22nd (second column in Fig. 4), the overall accuracy has increased to 74%. At this date, several parcels of CORN (in red) and one of RAPESEED (green) were stopped and correctly classified. The model predictions did not vary noticeably in June 21st (third row in Fig. 4) concerning April 12th. However, the number of stopped parcels increased steadily, as shown in the bottom row. The model classified most fields within the year's first half (shown in day-of-stopping; last row and column). However, single fields were still classified later in the year, emphasizing the need for a stopping decision for each field parcel, as the ELECTS-LSTM model provides.

4.1.3. Quantitative Prediction Accuracy

Figure 5 shows the classification accuracy up to a specific day of the year for all field parcels (orange) and only the stopped field parcels (blue) in the test set of BreizhCrops in Brittany, France. The horizontal axis represents the prediction date within the year until data is accessible to the classifier. The blue shaded area shows the number of stopped fields at each respective date. The orange line shows the classification accuracy calculated on all fields and represents the performance of a regular accuracy-only time series model. The blue line shows the classification accuracy only of the stopped fields, as provided by an early classification approach, such as the ELECTS-LSTM model in this work.

Early in the year, only a low accuracy between 30% and 50% on all fields is possible before March 1st, the day-of-year (doy) 60. The model can observe no fine-grained classification-relevant features this early, as this

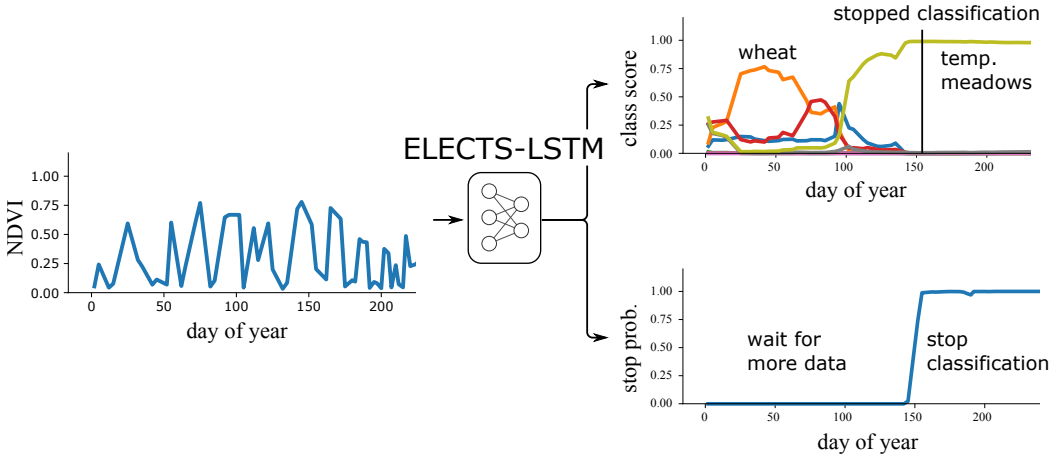


Figure 3: Prediction of the ELECTS-trained early classification model. The model ingests a time series (left) incrementally one element at a time. It estimates a probability for each crop category (top right) alongside a probability of stopping (bottom right). As long as the probability of stopping remains low, more data is necessary to obtain an accurate classification result.

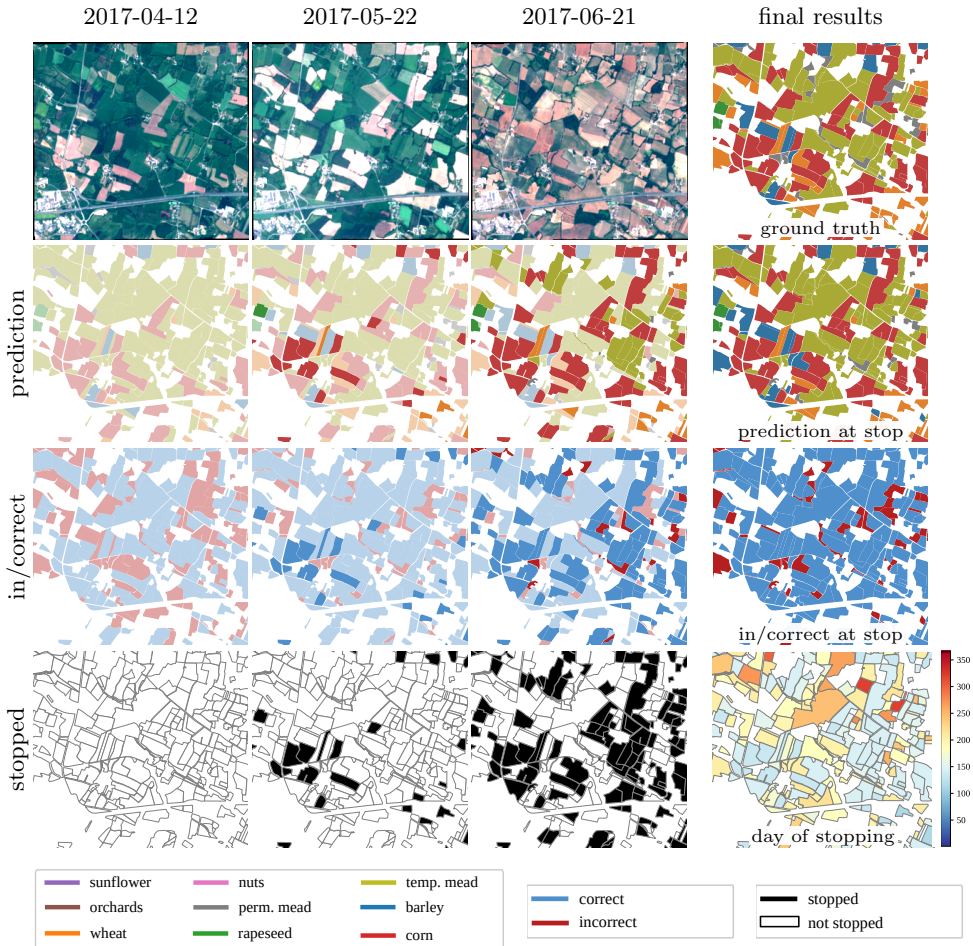


Figure 4: The ELECTS prediction process is shown in a deployment setting for one year, where the model predicts class labels and stops the predictions of individual fields. It shows early classification results in a 2.5km × 2.5km site in the Brittany test region (2.4052° West, 47.5328° North). The final results are shown in the right column. The previous columns show three dates with associated predictions (second row), and a correct/incorrect map (third row, with blue being correct predictions and red incorrect). Transparency is used for predictions in these rows to de-emphasize the fields where the prediction has stopped. The fourth row shows a binary score specifically to indicate which classifications are still active (white) or have already been stopped (black). The bottom right image depicts the stopping day for all parcels.

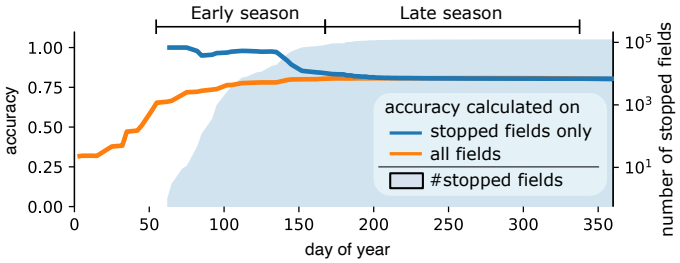


Figure 5: The accuracy of the ELECTS model during the year was calculated on all fields (orange) and only the stopped fields (blue). The ELECTS model only stops fields from doy (day-of-year) 60 onwards in the early season. The stopped fields are classified more accurately compared to all fields in the test region. A user can expect an accurate prediction of a stopped field.

period falls in the winter season. The stopping decision of the early classification model reflects this by declaring no fields as stopped before doy 60. From doy 60 onwards, an increasing number of fields are declared as stopped, as indicated by the blue shaded area. Notably, the few fields the model stopped in the early season between March and June (doys 60 and 150) are predicted at high accuracy, as shown by the blue line. Later during the year, the accuracy decreases when the classification of the majority of fields is stopped. The high accuracy of early-stopped fields reflects the intuition that the stopping decision is related to the model's confidence, as presumably easier-to-classify fields are stopped first, leading to the high accuracy in the early season. The more ambiguous and difficult fields are stopped in the late season. Wrong classifications become more common, and the accuracy drops to the same level as expected by an accuracy-only classifier.

From a practical deployment perspective, this result demonstrates that a user can be confident that the predictions are accurate for the stopped fields. This allows the user to make decisions for these individual field parcels early in the season.

4.2. Earliness Evaluation

Fig. 6 analyses the dates in which the ELECTS model stopped the classification from a phenological perspective concerning a local crop calendar of France. Figure 6a shows that stopping dates vary for individual crop types where the classification of all crops has been stopped in the agriculturally relevant period between planting and harvest. The average classification time of most crops, i.e., WHEAT, RAPESEED, CORN, SUNFLOWER, lies in the mid-season period. Notably, all crops except BARLEY (discussed later) were classified before the harvest period, which domain experts often consider the end-of-series date for accuracy-only classifiers when knowledge of local crop calendars is available.

RAPESEED parcels were classified particularly early: towards the end of April until mid-May, two months before the harvest period. We analyze these crop fields qualitatively in Fig. 6b where several RAPESEED fields are highlighted by a white outline on images from April 22nd, June 21st, and July 16th. RAPESEED fields blossom in a characteristic yellow color, as visible in the image of April 22nd. This blossoming period falls into the window where the classification of the majority of RAPESEED parcels has been stopped. Hence, we can deduce that the model uses this blossoming event as a characteristic feature to classify these parcels as RAPESEED and stop the classification confidently. In Fig. 6a, BARLEY was the only crop type where the model stopped the prediction during the harvest period end of June. We analyze this period qualitatively in the second row of Fig. 6b that shows BARLEY parcels outlined in white color. In particular, the effects of harvest are visible on June 21st, where BARLEY fields are the only field parcels that were recently harvested. After that date, bare soil is observed in the parcel, while all neighboring field parcels are covered by vegetation. From this analysis, we can deduce that these harvest operations cause the stopping decisions of the BARLEY crops, as the stopping dates of BARLEY parcels fall narrowly into this period. This analysis shows that the stopping times produced by the ELECTS-LSTM early classification model fall into a meaningful phenological period for this region. The interpretation of the crop calendar and explanation of the BARLEY and RAPESEED parcels show that the model learned to utilize meaningful features (e.g., the blossoming event of RAPESEED) to come to the stopping decision. Notably, this is learned without any direct temporal supervision as the model is optimized end-to-end solely on crop labels without any labels on time or crop cycles for this area.

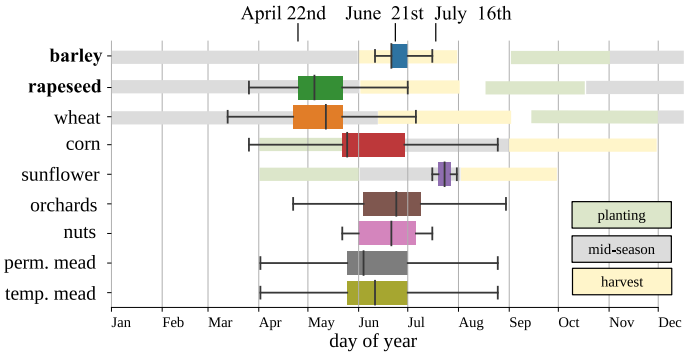
4.3. Applicability of ELECTS across datasets

The proposed model can be trained end-to-end on any time series classification problem if sufficient class-labeled data is available. This enables us to train models for different geographic areas without requiring region-specific expert knowledge aside from the labeled samples in the respective datasets. Hence, in this section, we test the applicability of the ELECTS-LSTM model with identical hyper-parameters to different datasets in Europe (France and Germany) in Fig. 7 and Africa (Ghana and South Sudan) in Fig. 8.

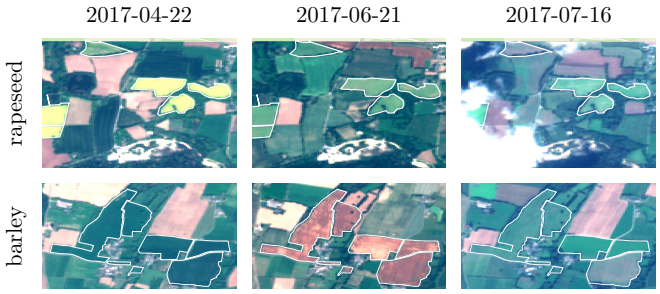
We organize this section in two parts. First, we discuss the model performance on large-scale European datasets where several ten to hundred thousands of field annotations are available. Then, we train and test the model on two African datasets where substantially less training data is available for end-to-end optimization of this deep learning model.

4.3.1. Large-scale datasets in Europe

Fig. 7 shows the accuracy as a confusion matrix and the earliness as a histogram of stopping times of field parcels in France (BreizhCrops) and Germany (BavarianCrops). All crop classes within the BavarianCrops dataset (Fig. 7a) were classified accurately with an overall accuracy of 86%.



(a) Quantitative evaluation of stopping times per crop type in Brittany, France, overlayed with planting, mid-season, and harvest dates from the crop calendar for France by the USDA Foreign Agricultural Service.



(b) Images of BARLEY and RAPESEED fields on three selected dates. The model stopped the classification of RAPESEED fields end of April when the characteristic yellow blossoms were visible, shown in the April 22nd image. BARLEY field classifications were stopped end of June during the harvest, as visible in June 21st, where bare soil is visible on the field parcels.

Figure 6: Stopping times of individual crops classes related to the local crop calendar in Brittany, France and examples of BARLEY and RAPESEED fields that reveal that specific machining and blossoming events cause early classifications of these crop categories.

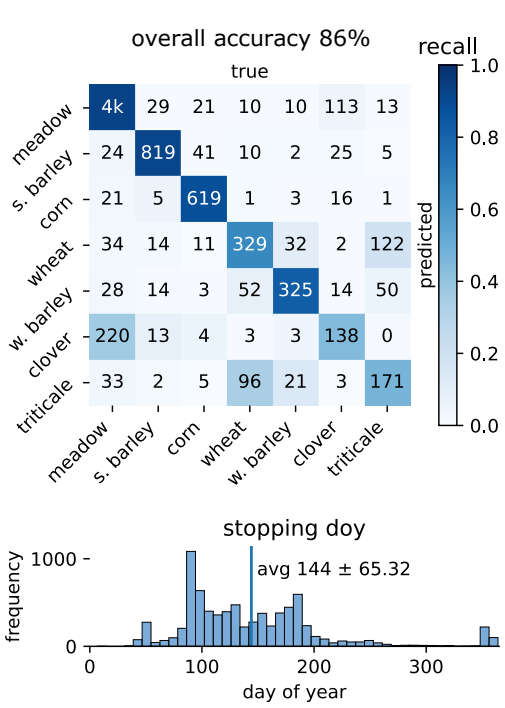
Systematic confusions were present between WHEAT, WINTER BARLEY, and TRITICALE, as these crops share biological ancestry and CLOVER and MEADOWS which are cultivated in a similar way and cut periodically throughout the year. Most notably, the model achieves this accuracy with only 40% of the entire sequence length. While the entire time series spans from January to December, most field parcels were classified within a two-month window (65 days) around May 24th. This highlights the potential of the ELECTS early classification approach to come to early and still accurate classification decisions within the year.

In Fig. 7b, we show the accuracy and earliness results on the BreizhCrops with fields of Brittany, France. Here, all crops are classified with an overall accuracy (OA) of 80% with an average stopping period of one month around June 7th. The ELECTS-LSTM model used only 32% of the overall time series on average. Most notably, regular accuracy-only models, which make predictions at the end of the entire time series, achieve comparable accuracies of 80% OA score, as shown in Fig. B.11a of Appendix B.2. In terms of classifications, systematic confusions are visible between PERMANENT MEADOWS and TEMPORARY MEADOWS. Infrequent classes, such as NUTS and SUNFLOWERS are not predicted correctly, as they have little effect on the overall loss objective. This classification of very imbalanced class distributions falls beyond the scope of this work. Overall, these results show that the ELECTS-modified LSTM model matches the accuracy of regular non-early classification models while predicting substantially earlier within the season.

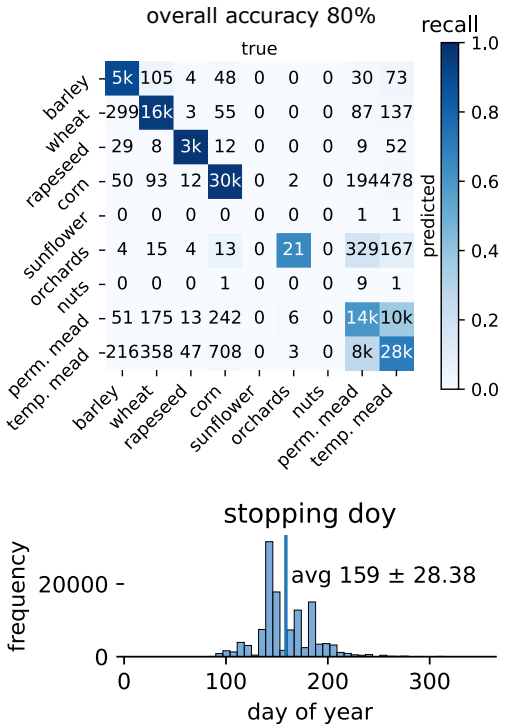
4.3.2. Small-scale datasets in Africa

Figure 8 shows confusion matrices and (class-wise) stopping times of the ELECTS-LSTM model in Ghana and South Sudan. The model finds accurate and early solutions on these datasets even though training a deep learning model on dataset sizes of 3837 and 737 individual field samples is inherently difficult. In South Sudan (Fig. 8a), an overall accuracy of 83% is achieved with average predictions on the day of year 61 (March 2nd). These very early classifications are driven mainly by RICE and SORGHUM fields that can be classified in January and February in this region. This overall accuracy is on a similar level to a convolutional LSTM model with 82.6% overall accuracy reported by Rustowicz *et al.*, (2019) [20]. Note that their underlying classification model is advantaged: the kernels in their convolutional LSTM model can make use of the pixel-neighborhood. In comparison, the LSTM implementation, which we modified for ELECTS, can only classify individual pixels separately from each other. Note, however, that ELECTS can be modified to incorporate spatio-temporal data by changing the feature extractor to the same classifier as Rustowicz *et al.*, (2019) [20]. On a dataset of this comparatively small size, both deep learning models performed 5.7% and 6.1% worse compared to a regular random forest classifier (concatenates all time points to one large feature vector) with 88.7% overall accuracy. Overall, the ELECTS-LSTM still compares well to the accuracy-only models from Rustowicz *et al.*, (2019) [20] while only requiring a fraction of the time series to come to an accurate and early decision. A similar trend is visible in the Ghana dataset shown in Fig. 8b where the ELECTS-LSTM model achieves an overall accuracy of 54% while classifying the fields on average on day of year 78 (March 19th). This accuracy is 7.1% and 5.9% worse compared to the random forest, and convolutional LSTM model from Rustowicz *et al.*, (2019) [20] that achieve 61.1%, and 59.9% accuracy, respectively. These accuracy-only models, however, can only predict after observing the entire time series, while the ELECTS-LSTM model used only 20% of the overall time series with an average stopping date of the 78th day of the year.

Overall, these results demonstrate that the ELECTS-LSTM model converges to a meaningful solution without any region-specific tuning. It produces early and still accurate predictions at a fraction of the entire time series. While the early classification model matched the accuracy



(a) BavarianCrops in Germany



(b) BreizhCrops in Brittany, France

Figure 7: Class-wise accuracy and earliness of the predictions on the two large-scale European datasets, BavarianCrops (a) and BreizhCrops (b). The ELECTS model predicts most classes accurately at a fraction of the required length of the time series in both datasets.

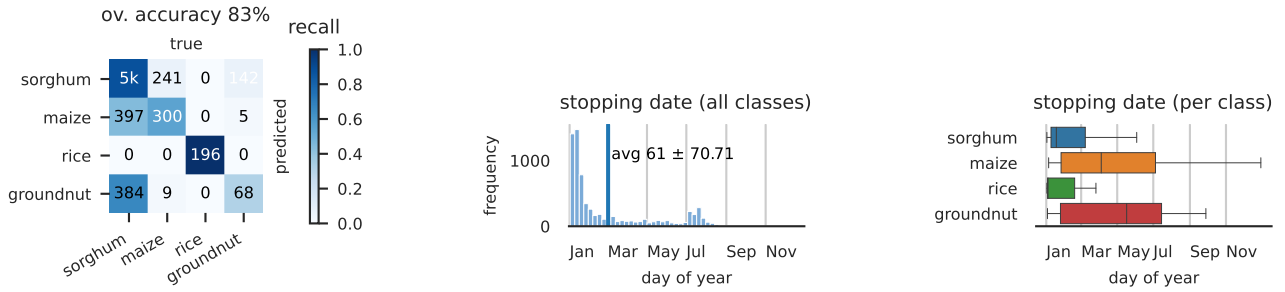
of accuracy-only models on the large BreizhCrops dataset, the model achieved a marginally lower accuracy on the substantially smaller datasets in Africa. Still, early and accurate predictions have been achieved without any region-specific parameter tuning with the ELECTS-LSTM model, which demonstrates the applicability in unexplored areas where sufficient training data is available.

5. Discussion

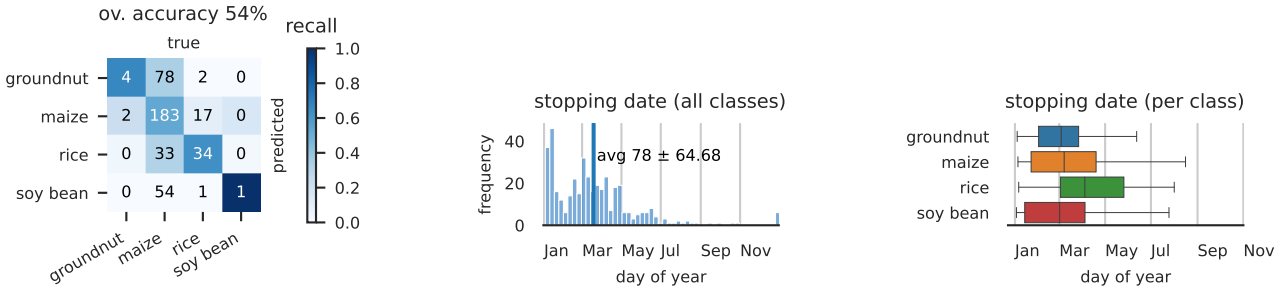
In this work, we demonstrated that a regular long short-term memory (LSTM) recurrent neural network can be effectively modified with ELECTS for early and accurate classifications on various crop-type datasets. On sufficiently large datasets, the performance is on par with that of accuracy-only models with only a fraction of the sequence length. The ELECTS-LSTM does not require refitting or predicting with different sub-sequences, conversely to related work [8**?** , 9] based on incremental classification. Other approaches [12, 11] are often developed and targeted towards one specific deployment area, often focused on the continental US, while we demonstrated the applicability of ELECTS on different continents. ELECTS inherits the limitations of deep learning: predominantly, the requirements of large annotated datasets for end-to-end optimization. This was evident in the predictions on small datasets (Ghana and South Sudan) where the accuracy of the ELECTS-LSTM was marginally worse than an accuracy-only model, while it matched the accuracy at the European large-scale datasets. Sensitivity to label imbalance is a further limitation where wrong classifications of infrequent classes are penalized less than frequent ones.

Deploying an ELECTS model on applications beyond crop type mapping would be a natural extension, as this model can be trained on any class-annotated time series dataset. Extending ELECTS for spatio-temporal data is feasible with little effort and can be done by modifying the feature extractor, for instance, by adding 2D convolutional layers.

The implications of an automated end-to-end trainable model, such as ELECTS, are manifold: acquiring predicted and accurate class labels for a subset of stopped crop parcels has direct practical implications for the control of European agricultural subsidies. In practice, sample on-site inspections often control the European subsidy after a specific pre-determined date. Field-wise, in-season predictions, as ELECTS provides, allow the start of this process weeks and months in advance. Further, the potential to save computational and storage resources is substantial: ELECTS provided accurate predictions using between 16% (South Sudan) to 40% (BavarianCrops) of the overall time series. For instance, when scaling the average earliness of predictions in BavarianCrops to the 43TB of Sentinel-2 imagery acquired in Europe each year, 26TB of downloading and processing satellite can be avoided. While Sentinel-2 data is free of charge, an increasing amount



(a) South Sudan



(b) Ghana

Figure 8: Accuracy and Earliness on datasets in Africa from [20] that was recently integrated in the SustainBench dataset [21]. The trained ELECTS-LSTM finds a generally accurate and early solution for both datasets without changing the training configuration even though the dataset sizes are substantially smaller compared to the large-scale European datasets of France and Germany.

of daily high-resolution imagery is available today [8]. This data needs to be acquired at a substantial cost and motivates the need to make confident decisions with data-efficient algorithms. Training and deploying an ELECTS-LSTM model is not expensive in terms of computational efforts. Deep learning models for 1D time series are small compared to standard 2D convolutional models for images. We trained the ELECTS-LSTM models within one hour on a single GPU on BavarianCrops. A researcher can make predictions using the trained ELECTS-LSTM model on a CPU with a regular notebook.

6. Conclusion

We presented a training framework for End-to-end Learned Early Classification of Time Series (ELECTS) that augments a regular deep time series classification model by a second decision head informing about prediction uncertainty and leading to early stopping. The core contribution is a loss function that incorporates both model outputs such that the two objectives of earliness and accuracy are balanced. Thanks to the earliness objective, ELECTS provides indirect insight into its decision process. We showed that the model linked the stopping decision to the phenological events of the plants for two crop types. Stopped classifications early in the season were also particularly accurate, highlighting that the model connects the stopping decisions to predictive confidence. ELECTS goes beyond crop types classification, as it can be applied to potentially any data where temporally coarse labels are available that are not aligned with the events, e.g., one label per year. In general, with satellites providing a constant stream of data that is necessary to monitor time-dependent processes at the surface [25], a variety of deployments are feasible, from dynamically determining cloud categories [26] to the detection of deforested areas [27] in a time-sensitive manner. The source code to the models, the ELECTS loss function, and to reproduce the experiments are available at `https://github.com/marccoru/elects`.

7. Acknowledgements

The work of Marc Rußwurm was partially funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under reference 50EE1908. Romain Tavenard was partially funded through project MATS ANR-18-CE23-0006. Nicolas Courty is partially funded through the ANR project OTTOPIA ANR-20-CHIA-0030.

References

[1] Brian D Wardlow and Stephen L Egbert. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains. *Remote Sensing of Environment*, 112(3):1096–1116, 2008. 1

[2] Per Jönsson and Lars Eklundh. Timesat—a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, 30(8):833–845, 2004. 1

[3] Heather McNairn, Angela Kross, David Lapen, Ron Caves, and Jiali Shang. Early season monitoring of corn and soybeans with TerraSAR-X and RADARSAT-2. *International Journal of Applied Earth Observation and Geoinformation*, 28:252–259, 2014. 1, 2

[4] Emmanuelle Vaudour, Paul Emile Noirot-Cosson, and Olivier Membrive. Early-season mapping of crops and cultural operations using very high spatial resolution Pléiades images. *International Journal of Applied Earth Observation and Geoinformation*, 42:128–141, 2015. 1, 2

[5] Jordi Inglada, Arthur Vincent, Marcela Arias, and Claire Marais-Sicre. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. *Remote Sensing*, 8(5):362, 2016. 1, 2

[6] Yaping Cai, Kaiyu Guan, Jian Peng, Shaowen Wang, Christopher Seifert, Brian Wardlow, and Zhan Li. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sensing of Environment*, 210:35–47, 2018. 1, 2

[7] Michael Marszalek, Maximilian Lösch, Marco Körner, and Urs Schmidhalter. Early crop-type mapping under climate anomalies. *Preprints*, 2020040316, 2020. 1

[8] Lukas Kondmann, Sebastian Boeck, Rogerio Bonifacio, and Xiao Xiang Zhu. Early crop type classification with satellite imagery-an empirical analysis. In *ICLR Practical ML for Developing Countries Workshop*, 2022. 1, 9, 10

[9] Mmamokoma Grace Maponya, Adriaan van Niekerk, and Zama Eric Mashimbye. Pre-harvest classification of crop types using a Sentinel-2 time-series and machine learning. *Computers and Electronics in Agriculture*, 169:105164, 2020. 1, 9

[10] Sergii Skakun, Belen Franch, Eric Vermote, Jean-Claude Roger, Inbal Becker-Reshef, Christopher Justice, and Nataliia Kussul. Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment*, 195:244–258, 2017. 1

[11] Venkata Shashank Konduri, Jitendra Kumar, William W Hargrove, Forrest M Hoffman, and Auroop R Ganguly. Mapping crops within the growing season across the United States. *Remote Sensing of Environment*, 251:112048, 2020. 1, 2, 9

[12] Chenxi Lin, Liheng Zhong, Xiao-Peng Song, Jinwei Dong, David B. Lobell, and Zhenong Jin. Early- and in-season crop type mapping without current-year ground truth: Generating labels from historical information via a topology-based approach. *Remote Sensing of Environment*, 274:112994, 2022. 1, 2, 9

[13] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018. 1, 2

[14] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[15] Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5):523, 2019. 1, 2, 14

[16] Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A time series dataset for crop type mapping. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 2020. 2

[17] Usue Mori, Alexander Mendiburu, Sanjoy Dasgupta, and Jose A Lozano. Early classification of time series by simultaneously optimizing the accuracy and earliness. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4569–4578, 2018. 2, 13

[18] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:421–435, 2020. 2

[19] Nicolas Karasiak, J-F Dejoux, Claude Monteil, and David Sheeren. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, pages 1–26, 2021. 2

[20] Rose Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2, 3, 4, 8, 10

[21] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Proceedings of the Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 10

[22] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998. 3

[23] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. 3

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3, 14

[25] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein (Editors). *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. Wiley & Sons, 2021. 10

[26] Gonzalo Mateo-García, Jose E Adsuara, Adrián Pérez-Suay, and Luis Gómez-Chova. Convolutional long short-term memory network for multitemporal cloud detection over landmarks. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 210–213, 2019. 10

[27] Johannes Reiche, Adugna Mullissa, Bart Slagter, Yaqing Gou, Nandin-Erdene Tsendbazar, Christelle Odongo-Braun, Andreas Vollrath, Mikaela J Weisse, Fred Stolle, Amy Pickens, et al. Forest disturbance alerts for the Congo Basin using Sentinel-1. *Environmental Research Letters*, 16(2):024005, 2021. 10

[28] Ashish Gupta, Hari Prabhat Gupta, Bhaskar Biswas, and Tanima Dutta. Approaches and applications of early classification of time series: A review. *IEEE Transactions on Artificial Intelligence*, 1(1):47–61, 2020. 13

[29] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The UCR time series classification archive, October 2018. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`. 13

[30] Asma Dachraoui, Alexis Bondu, and Antoine Cornuéjols. Early classification of time series as a non myopic sequential decision making problem. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 433–447, 2015. 13

[31] Romain Tavenard and Simon Malinowski. Cost-aware early classification of time series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 632–647, 2016. 13

[32] Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang. CSI-Net: Unified body characterization and action recognition. *arXiv*, 1810.03064, 2018. 14

[33] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. 14

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Infor-*

α	accuracy	κ	earliness
0.0	0.25±0.22	0.12±0.19	0.90±0.17
0.2	0.81±0.03	0.71±0.04	0.60±0.02
0.4	0.80±0.09	0.71±0.10	0.53±0.03
0.6	0.85±0.02	0.77±0.03	0.12±0.07
0.8	0.84±0.01	0.76±0.02	0.07±0.05
1.0	0.83±0.03	0.75±0.04	0.00±0.00

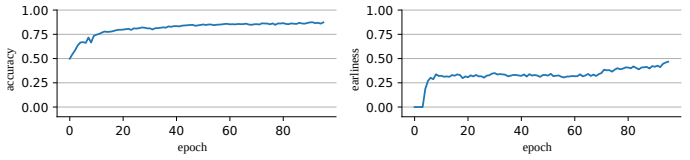
(a) BavarianCrops.

α	accuracy	κ	earliness
0.0	0.31	0.00	1.00 ±0.00
0.2	0.80	0.74	0.73 ±0.07
0.4	0.80	0.74	0.69 ±0.07
0.6	0.81	0.75	0.66 ±0.09
0.8	0.80	0.74	0.60 ±0.12
1.0	0.81	0.75	0.00 ±0.00

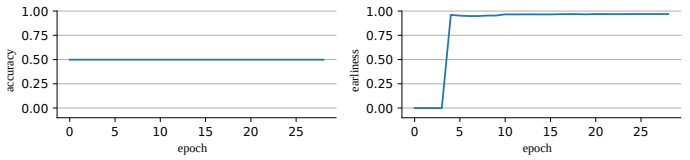
(b) BreizhCrops.

Table A.1: Varying the weighting factor α that trades-off classification loss and earliness reward. An $\alpha = 1$ corresponds to a high weight on earliness, while $\alpha = 0$ switches off the earliness reward. Results for BavarianCrops are averaged over three runs. Standard deviations of earliness refer to stopping times of single fields.

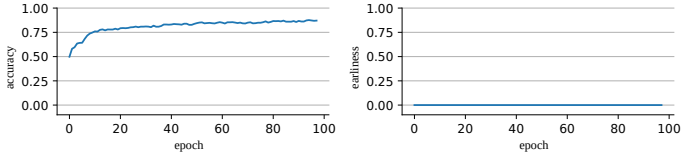
mation Processing Systems, pages 5998–6008, 2017. 14



(a) $\varepsilon = 10$. Normal training with early and accurate solutions. Observed in all 20 runs with $\varepsilon = 10$.



(b) $\varepsilon = 0$. Failure case 1. Model only optimized earliness. Observed in 2/20 runs.



(c) $\varepsilon = 0$. Failure case 2. The model only optimized for accuracy, which was observed in 1/20 runs.

Figure A.9: We show normal training behavior (a) compared to two failure cases (b,c) we observed in some training runs with $\varepsilon = 0$. In failure case 1 (b), the model only optimized the earliness, while in failure case 2 (c) only accuracy was optimized. None of the runs with $\varepsilon = 10$ experienced these failure cases.

Appendix A. Ablation Experiments

Appendix A.1. Ablations on Hyperparameters and Loss Design

In this group of experiments, we test the individual model components. In Appendix A.1.1, we vary the trade-off between accuracy and earliness while we focus in Appendix A.1.2 on the offset parameter ε.

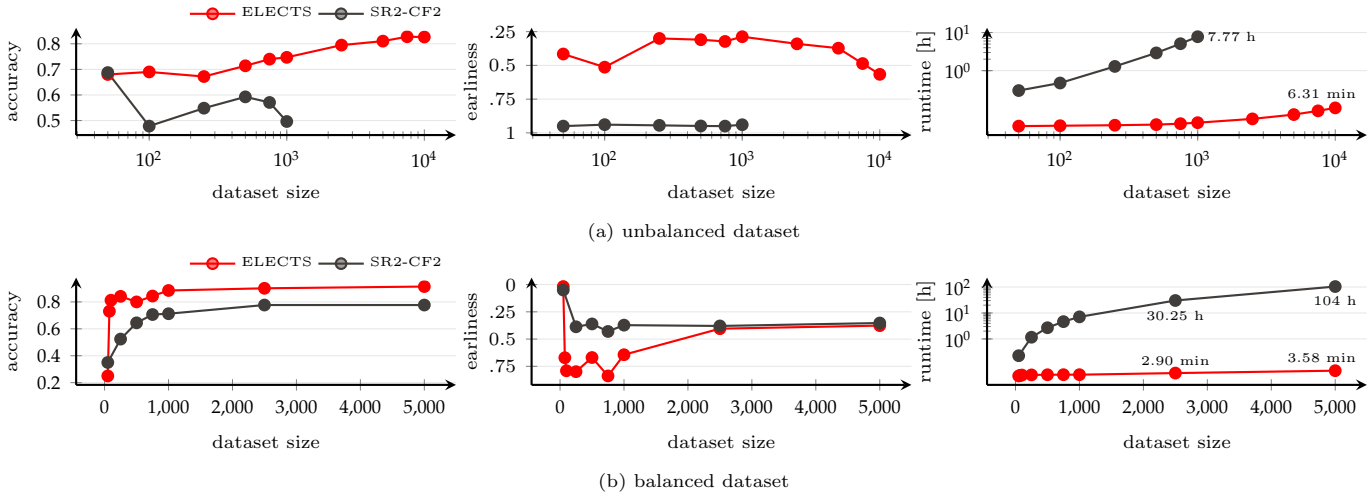
Appendix A.1.1. Controlling Earliness versus Accuracy

In this experiment, we study the effect of α of Eq. (4) on both the BavarianCrops and BreizhCrops datasets. For the BavarianCrops dataset, we observed some variance when training models from different weight initializations. Hence, we report the mean and standard deviation of 3 model runs in Table A.1a. For BreizhCrops, a similar accuracy level between 80% and 85% was achieved for a wide range of α values while the earliness decreased from 0.6 to 0.07. The accuracy-only ($\alpha = 1$) runs did not achieve the best accuracy (83%) compared to 85% with $\alpha = 0.6$. This result indicates that an earlier classification, temporally closer to the classification-relevant features, may have also been beneficial for the achieved accuracy.

These observations are mirrored in the BreizhCrops dataset in Table A.1b. Consistent accuracy of 80-81% is achieved for all $\alpha > 0.2$. Similarly, larger weights on earliness reward with larger α values lead to slightly earlier classifications of 13% of the overall sequence length. Comparing BreizhCrops and BavarianCrops, we observe that the accuracies in BreizhCrops were more consistent throughout the entire α-range, which we associate with the 20-times larger training set size. This larger quantity in labeled samples helps the model to find the optimum classification-relevant features in a certain time regardless of the model initialization and α-weights.

Appendix A.1.2. Effect of the Offset Parameter ε

We trained ELECTS-LSTM 40 times for 100 epochs on the BavarianCrops dataset in this experiment. In 20 training runs, we set $\varepsilon = 0$ leading to no offset in Eq. (8). In the second set of 20 runs, we set $\varepsilon = 10$. In 37 of 40 runs, we observed a normal training behavior, as shown in Fig. A.9a where classification accuracy and earliness increased throughout the training. All 20 training runs with $\varepsilon = 10$ showed this normal training behavior, while in 20 runs with $\varepsilon = 0$, we experienced two rare failure cases. Two of twenty runs experienced failure case 1, as shown in Fig. A.9b. In these runs, the earliness increased to 1 early in the training leading to classification at the beginning of the sequence. The accuracy did not improve upon the initial epoch at 50%, which lies between the accuracy of a random predictor of 16% and predicting only the most frequent class (MEADOW) at an accuracy of 57%. Here, the model fell in a local optimum where it solely minimized the earliness reward. In Fig. A.9c, we show a second failure case that appeared one of twenty times on the runs with $\varepsilon = 0$. The accuracy increased steadily, but the predictions remained at the end of the sequence with an earliness of 0. In this case, the model minimized the classification loss



(a) unbalanced dataset

(b) balanced dataset

Figure A.10: This figure shows the evaluation of ELECTS (ours) and SR2-CF2 performance on class-balanced subsets of the BavarianCrops dataset. The x-axis refers to the size of the training data to train both models. We can observe that SR2-CF2 does not converge to a meaningful solution on imbalanced data (a) where it predicts at the beginning of the sequence (earliness = 0) at a low accuracy (as it could not observe any classification-relevant features that early). We needed to artificially balance the dataset in (b) for SR2-CF2 to predict early and accurately. In comparison, the ELECTS-LSTM converges to a meaningful solution in both cases and is computationally more efficient. With 5000 training time series in the balanced case, it required four minutes to train, while SR2-CF2 required 104 hours.

but did not improve upon the earliness objective. None of these failure cases were observed in the runs of $\varepsilon = 10$ leading to a more stable convergence to an early and accurate solution with this offset parameter.

Appendix B. Model Comparison

In parallel, early classification has been discussed in the time series classification community, as summarized in the review of Gupta *et al.* [28]. Here, early time series classification approaches are tested on a set of benchmark datasets in the UCR Archive [29]. While these datasets cover a diverse range of applications, their small size of at most a few thousand examples favors shallow learning solutions. In this application space, several approaches introduced the idea to explicitly model the maximization of earliness in the optimization objective function [30, 31]. In particular, Mori *et al.* [17] also consider explicitly optimizing the trade-off between earliness and accuracy. Their SR2-CF2 model variant first independently trains a Gaussian Process Classifier for each sub-sequence length. It then uses a genetic algorithm to find the parameter for a stopping rule that takes prediction confidence for each class into account.

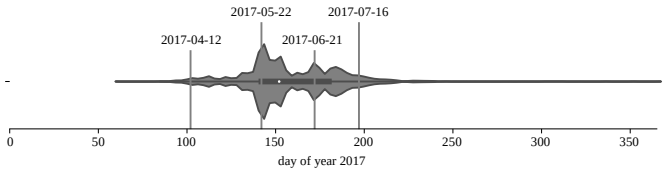
Appendix B.1. Comparison to SR2-CF2

In this first comparison, we use the BavarianCrops dataset. The source code of SR2-CF2 was explicitly designed for uni-variate and class-balanced data in the UCR Time Series archive [29]. We extended it to multi-variate time series in the modified source code[1]. We could not successfully run SR2-CF2 on the complete BavarianCrops dataset and sampled sub-datasets of 50, 100, 250, 500, 750, 1000, 2500, 5000, 7500, and 10 000 samples where we successfully ran SR2-CF2 on subsets up to 1000 samples. Results are presented in Fig. A.10a where the SR2-CF2 method struggled to converge to a good solution. It predicted the most common class very late (small earliness) throughout differently-sized subsets of the crop type mapping dataset. We connect this to the class imbalance present in the dataset. To alleviate this class imbalance, we sampled a second class-balanced dataset by undersampling frequent classes (e.g., MEADOW) and oversampling rare ones (e.g., TRITICALE). We created differently sized subsets of this balanced variant with 50, 75, 100, 250, 500, 750, 1000, 2500, 5000 samples and show the results in Fig. A.10b. Here, the SR2-CF2 model achieved accurate and early classifications consistent with results reported on the UCR archives [29]. The ELECTS-trained model provided accurate but late (small earliness) classifications for datasets smaller than 2500 samples. For datasets with 2500 training series or more, ELECTS and SR2-CF2 achieve comparable earliness, whereas the ELECTS-trained LSTM model predicted the classes more accurately. While accuracy and earliness were generally comparable for datasets with more than 2500 samples, the difference in runtime, as shown in the last column, became a substantial factor. The computational complexity of SR2-CF2 is $\mathcal{O}(N^2T^2)$ where N refers to the number of samples in the dataset and T to the sequence length. The ELECTS-trained LSTM model relies on vanilla gradient descent that can utilize modern automatic differentiation libraries with a complexity of $\mathcal{O}(n_{\text{epochs}}NT)$. In total, with a 5000 sample-sized dataset, ELECTS required 4 minutes while SR2-CF2 104 hours.

[1] the extended source code to multi-variate time series of Mori *et al.*, (2017) [17] is available as fork `https://github.com/marccoru/earlyclassification`

model	accuracy	kappa	earliness	average date of classification
Random Forest	0.78	0.69	0 (fixed)	Dec 28th (fixed)
TempCNN [15]	0.79	0.73	0 (fixed)	Dec 28th (fixed)
MS-ResNet [32]	0.77	0.70	0 (fixed)	Dec 28th (fixed)
Inceptiontime [33]	0.77	0.73	0 (fixed)	Dec 28th (fixed)
LSTM [24]	0.80	0.74	0 (fixed)	Dec 28th (fixed)
Transformer [34]	0.80	0.74	0 (fixed)	Dec 28th (fixed)
ELECTS-LSTM	0.80	0.74	0.68 ± 0.07	**June 7th ± 28 days**

(a) Comparison with regular classification models on the BreizhCrops dataset.



(b) Frequency of dates where the classification has been stopped.

Figure B.11: Comparison of the ELECTS-LSTM model with other accuracy-only methods. ELECTS-LSTM matched the accuracy of the other models (a) while predicting at an earliness of 0.68 ± 0.07, meaning that only $32\% \pm 7\%$ of the time series was necessary. The average time of stopping is June 7th $\pm$ 28 days (earliness 0.68 ± 0.07) which lies in the greenup period in Brittany (b), as indicated by the four highlighted dates which correspond to the images (and analysis) in Fig. 4.

Appendix B.2. Comparison to non-early classification models on BreizhCrops

In this section, we compare the ELECTS-trained LSTM model with several models from the literature, optimizing for accuracy (Random Forest, TempCNN [15], MS-ResNet [32], Inceptiontime [33], accuracy-only LSTM [24], Transformer [34]). For such a comparison, we focus on the BreizhCrops benchmark. The results are shown in Fig. B.11a. The accuracy and kappa score measure the classification performance. In contrast, earliness $1 - \frac{t}{T}$ measures how much data from the original T-length sequence was not needed to come to a prediction. The accuracy-only comparison models are always classified at the end of the sequence ($t = T$), which corresponds to a hard-coded earliness of 0. From Fig. B.11a, we see that the ELECTS-trained LSTM model matches the accuracies of the comparison models while predicting before the end of the sequence. But additionally to matching accuracy, ELECTS allows for earlier predictions (and the related savings in data download, storage, and processing time): in BreizhCrops, ELECTS achieves an earliness of 0.68 ± 0.07, meaning that only $32\% \pm 7\%$ of the time series was necessary for the classification. This also means that the classification was stable (and stopped on June 7th $\pm$28 days rather than on December 28 for the other methods, which need the entire time series.

Given that the evaluated earliness is early, we investigated the nature of the stopping period in greater detail in Fig. B.11b. Here, we show the frequency of stopped dates of the ELECTS-LSTM model. We see that no classifications have been stopped before March (the 60th day of the year), which lies in the non-informative winter period. Most classifications have been made between the end of May and early June.