Mind the Gap: Bridging Prior Shift in Realistic Few-shot Crop-type Classification

Joana Reuss[1], Ekaterina Gikalo[1], Marco Körner[1,2,3]

[1] Technical University of Munich (TUM), TUM School of Engineering and Design, Department of Aerospace and Geodesy, Chair of Remote Sensing Technology, 80333 Munich, Germany - {joana.reuss, ekaterina.gikalo, marco.koerner}@tum.de
[2] Technical University of Munich (TUM), Munich Data Science Institute (MDSI), 85748 Garching, Germany
[3] ELLIS Unit Jena, University of Jena, 07743 Jena, Germany

Keywords: few-shot learning, DirPA, prior shift, Dirichlet distribution, crop-type classification, EuroCropsML.

Abstract

Real-world agricultural distributions often suffer from severe class imbalance, typically following a long-tailed distribution. Labeled datasets for crop-type classification are inherently scarce and remain costly to obtain. When working with such limited data, training sets are frequently constructed to be artificially balanced—in particular in the case of few-shot learning—failing to reflect real-world conditions. This mismatch induces a shift between training and test label distributions, degrading real-world generalization. To address this, we propose ***Dir**ichlet **P**rior **A**ugmentation (DirPA)*, a novel method that simulates an unknown label distribution skew of the target domain proactively during model training. Specifically, we model the real-world distribution as *Dirichlet*-distributed random variables, effectively performing a prior augmentation during few-shot learning. Our experiments show that DirPA successfully shifts the decision boundary and stabilizes the training process by acting as a dynamic feature regularizer.

1. Introduction

In light of the increasing risk factors associated with food security, accurate agricultural monitoring is becoming increasingly crucial. While *machine learning (ML)* methods have achieved state-of-the-art performance on multi-spectral (time series) crop-type data (Qi et al., 2023; Saini and Ghosh, 2018), their reliability in real-world scenarios remains critically hindered by data scarcity and distributional shifts.

Real-world label distributions in crop-type classification are often highly skewed. For instance, common crops like `wheat` dominate the landscape, while rare ones like `parsley` are heavily underrepresented. Compounding this issue, the high costs and labor required to acquire accurate crop-type labels often limit the available data to only a few examples per class, making *few-shot learning (FSL)* a practical approach for this domain. Frequently, FSL is equated with the concept of *meta-learning* but it can also simply be used in order to describe the low-data constraint in isolation. We use FSL to refer to the latter, noting that recent work has shown that regular transfer learning via pretraining and fine-tuning can receive competitive results to complex meta-learning algorithms (Reuss et al., 2026; Chen et al., 2019). However, labeled training (support) datasets in FSL are often constructed with a balanced label distribution. This reflects an idealized scenario where, given data scarcity, samples are deliberately collected to stabilize learning and ensure a fair representation across all classes. In order to reflect realistic scenarios, the standard practice of using a balanced FSL test (query) set has been criticized as being unrealistic, with studies recommending the use of arbitrary and imbalanced test sets (Veilleux et al., 2021; Ochal et al., 2023; Mohammadi, Belgiu and Stein, 2024). Consequently, the training class prior $p_{\text{train}}(y)$ is not representative of the real-world test prior $p_{\text{test}}(y)$. As a result, during the testing or inference phase, the model is exposed to a distributional shift, leading it to learn a strong, incorrect bias. This typically results in poor generalization, especially on small or long-tailed dataset (Reuss et al., 2026).

While most existing methods address such label or prior shifts *post-hoc* after training by correcting the predicted class probabilities at inference time (Lipton, Wang and Smola, 2018; Kluger, Wang and Lobell, 2021, *cf.* Section 2), we propose to model the prior uncertainty proactively during the training process. Specifically, we leverage the *Dirichlet distribution* to sample a vast range of class distributions. This exposes the model to various realistic label distributions, ultimately leading to a classifier with superior robustness to prior shift during inference without any knowledge of the actual test skew. This is particularly critical when generalizing from a few labeled samples.

The main contributions of this work are:

1. **Prior-agnostic representation learning**: We introduce **Dir**ichlet **P**rior **A**ugmentation, a novel method that trains models on balanced few-shot datasets using prior augmentations in order to make the model highly robust to the class prior $p(y)$.
2. **Enhanced regularization and robustness**: We demonstrate that our proposed method acts as an effective regularizer, stabilizing the training process and improving robustness on severely imbalanced target domains, in particular in low-shot regimes.

2. Related work

2.1 Few-shot learning paradigm

Most supervised ML methods require large amounts of labeled data in order to achieve reasonable performance. few-shot learning, on the contrary, deals with the scenario where only a very limited number of labeled samples is available.

2.1.1 FSL as transfer vs. meta-learning A common framework in FSL is *meta-learning (MTL)*. In fact, both terms are often used interchangeably. The core goal of MTL is to learn

https://doi.org/10.5194/isprs-annals-XI-3-2026-787-2026

entire function spaces in order to quickly adapt to unseen, related tasks using only a few labeled samples. Therefore, it is also commonly referred to as *learning-to-learn*. One of the most prominent MTL algorithms is *Model-agnostic Meta-learning (MAML; Finn, Abbeel and Levine, 2017)* and its variants (Raghu et al., 2019; Tseng, Kerner and Rolnick, 2022).

Transfer learning, on the other hand, consists of training a model on a rather large set of labeled data before transferring it to a second, often unrelated, target task with subsequent fine-tuning. This concept has been widely used across various fields (Kurian, Jacob and Kuruvilla, 2024; Alem and Kumar, 2022; Rouba and Larabi, 2023). Chen et al. (2019) provide a comprehensive evaluation of existing FSL approaches. They find that traditional transfer learning achieves comparable or even superior performance on few-shot tasks compared to state-of-the-art meta-learning approaches.

2.1.2 FSL in remote sensing

FSL has been widely applied to the field of remote sensing (Reuss et al., 2026; Tseng, Kerner and Rolnick, 2022; Rußwurm et al., 2020; Wang et al., 2020; Tseng et al., 2021). Tseng, Kerner and Rolnick (2022) extended the concept of the original MAML algorithm explicitly for agricultural monitoring by taking into account additional metadata such as the spatial coordinates. Reuss et al. (2026) provide a comprehensive cross-regional benchmark study using the few-shot crop-type dataset EuroCropsML (Reuss et al., 2025). Their findings show that, while meta-learning achieves superior performances compared to regular transfer learning and self-supervised learning, it comes at the expense of increased computational costs. Moreover, they highlight that none of the evaluated methods were capable of overcoming the discrepancy in distribution between the balanced train set and the imbalanced test set.

2.1.3 Class imbalance and prior shift in FSL

Ochal et al. (2023) provide a detailed evaluation and comparison of various existing few-shot learning methods under class imbalance. They found that random oversampling during balanced training significantly improves performance and outperforms rebalancing loss functions, *e.g.*, the *focal loss (FL; Lin et al., 2020)*.

Prior distribution shift correct at inference-time Recent studies address the problem of prior distribution shifts often at inference time. *Black Box Shift Estimation (BBSE; Lipton, Wang and Smola, 2018)* estimates the test distribution $p_{\text{test}}(y)$ to improve generalization for symptom-diagnose detection. Kluger, Wang and Lobell (2021) directly tackles the problem of label (prior) and feature (covariate) distribution shift in few-shot crop-type classification using crop statistics, assuming that the distribution of the test set is known. Specifically, to address the prior distribution shift, they reweigh the posterior probabilities. Sipka, Sulc and Matas (2022) present a novel prior estimation approach based on confusion matrices.

2.2 Dirichlet priors and distribution augmentation

The Dirichlet distribution is often used to model the prior in Bayesian statistics, *cf.* Section 4. Among others, previous work addressed supervised clustering (Daumé III and Marcu, 2005) and the utilization of Dirichlet priors within a Bayesian framework for regression (Rademacher and Doroslovački, 2021). The latter propose a Dirichlet prior because it possesses multiple desirable benefits:

Full support The Dirichlet distribution covers the full space of possible probability distributions. This means that the model can, technically, still learn the true data distribution, regardless of the accuracy of the initial prior.

Closed-form posterior distributions It represents the conjugate prior for multinomial data (*e.g.*, categorical counts), *cf.* Section 4. This leads to a simple, closed-form posterior distribution, which significantly simplifies mathematical derivations and provides computational efficiency.

Controllable informativeness The prior contains a so-called localization parameter α_0 which explicitly manages the bias-variance trade-off. It can be set to be highly opinionated (modeling strong prior knowledge) or non-committal (allowing the data to dominate the model's training).

2.2.1 Dirichlet for FSL evaluation

Although Dirichlet priors have been employed for few-shot learning, previous studies rely on the assumption that both the train (support) and test (query) sets are balanced. Therefore, they utilize Dirichlet sampling to generate diverse test distributions (Veilleux et al., 2021; Mohammadi, Belgiu and Stein, 2024). Thus, these approaches can be considered instantiating a few-shot evaluation method, since their sole effect is to simulate a realistic imbalanced test set.

2.3 Summary and relation to our work

While recent studies (Reuss et al., 2026) demonstrate that meta-learning methods often achieve slightly superior performance in FSL for crop-type classification, they suffer from high computational costs. Therefore, this work chooses the transfer-learning paradigm, which has been shown to achieve competitive results (Reuss et al., 2026; Chen et al., 2019). However, the underlying principle of DirPA is general and not restricted to this paradigm.

Addressing the problem of prior distribution shifts, existing methods primarily rely on correction at inference time (Kluger, Wang and Lobell, 2021; Sipka, Sulc and Matas, 2022), requiring explicit or estimated knowledge of the final test distribution. Moreover, while the Dirichlet distribution has been used in FSL in order to create diverse evaluation sets (Veilleux et al., 2021; Mohammadi, Belgiu and Stein, 2024), its utilization has been limited to evaluation only.

Our **Dir**ichlet **P**rior **A**ugmentation (DirPA) approach tackles this shift directly during training, representing a novel proactive approach. By using the sampled priors to augment the training distribution with diverse class priors, DirPA forces the model to learn a feature representation that is fundamentally *prior-agnostic*, eliminating the need for any inference-time prior estimation.

3. Dataset

In this study, we use the Estonia data from the EuroCropsML dataset (Reuss et al., 2025) for training and evaluation. EuroCropsML is a time-series dataset that combines parcel reference data and multi-class *hierarchical crop and agriculture taxonomy (HCAT; Schneider et al., 2023; Schneider, Broszeit and Körner, 2021)* labels from EuroCrops (Schneider et al., 2023) with Sentinel-2 L1C optical satellite observations captured during the year 2021. Each data point contains a time series of cloud-free multi-spectral Sentinel-2 observations for

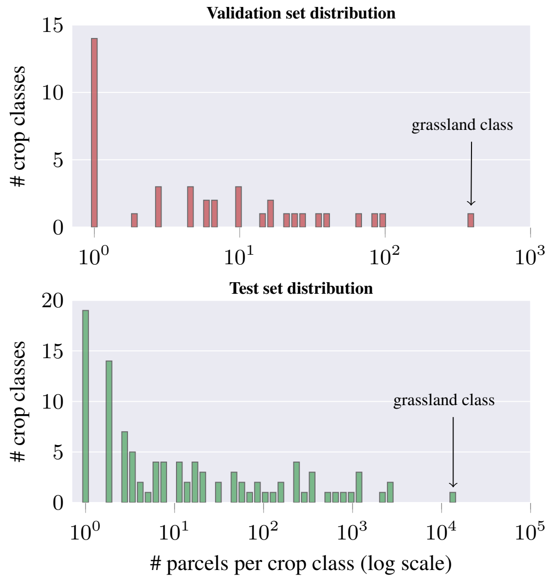


Figure 1. Abundances of crop types in Estonia. Histograms showing the binned distributions of crop-type abundances in Estonia for 1000 randomly sampled data points of the validation set and the full test set.

all 13 bands. We updated the original EUROCROPSML labels with the newest HCAT version 4 (Claverie et al., 2026) to reflect the corrected class structure.

The dataset reflects real-world agricultural complexity, including regional variations in crop types, vegetation patterns, and parcel sizes, which pose significant challenges for classification. Notably, it also exhibits a strong class imbalance with `grassland grass` being the most frequent one among the 129 crop types, representing 46 % of all samples.

3.1 Dataset split

The total dataset comprising 175 906 samples is divided into training, validation, and test sets. We allocate 60% of the samples to the training set and divide the remaining 40% equally between the validation and testing sets. This yields 105 543 samples for training purposes and 35 182 for both validation and testing. The dataset's imbalanced distribution, where some classes contain only a single sample, created partly disjoint sets. Specifically, 24 classes are unique to the training set. In total, the test set contains 95 classes of which 7 are not present in the train set. Of these 7 classes, 6 are fully unique to test. As a consequence, the model is forced to perform *zero-shot* classification when attempting to classify any of these 7 novel classes. Figure 1 shows the distributions of crop-type abundances within the validation and test set, while Figure 2 presents the spatial test class coverage.

We sample different few-shot scenarios, specifically: 1, 5, 10, 20, 100, 200, or 500 shots. Updating the crop classes to HCAT4 gave rise to alterations to the classifications of certain parcels. Therefore, the utilization of the original splits (Reuss et al., 2025) would have resulted in a violation of the few-shot setting.

4. Methodology

Our proposed method aims to simulate prior distribution shifts proactively during the training process to increase the robustness

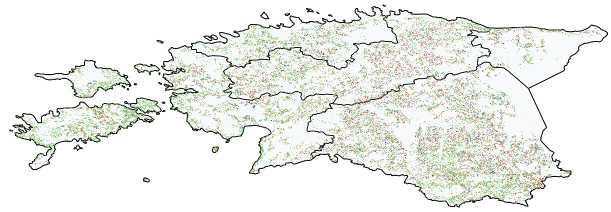


Figure 2. Spatial distribution of crop classes in the test set of Estonia. Map of Estonia showing the location and distribution of crop types within the final test set. Each data point marks the central coordinate of an agricultural parcel and is color-coded by its corresponding crop class.

of the model against various potential label distributions during inference. Instead of training with balanced priors, we inject synthetic prior shifts during training.

4.1 Dirichlet distribution

The Dirichlet distribution, denoted $\mathrm{Dir}(\boldsymbol{\alpha})$ and formally stated in Definition 1, is a family of continuous multivariate probability distributions, parametrized by $\boldsymbol{\alpha} \in \mathbb{R}_{+}^{K}$. It models the distribution of proportions or probabilities, *i.e.*, non-negative values with unit integral, and is commonly used as a prior in Bayesian statistics (Steck and Jaakkola, 2002; Daumé III and Marcu, 2005; Rademacher and Doroslovački, 2021). Figure 3 illustrates the density function of the (symmetric) Dirichlet distribution for $K = 3$ variables and different $\boldsymbol{\alpha}$.

Definition 1 *Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \in \Delta^{K-1}$ be a K-dimensional continuous random vector. The* Dirichlet distribution *is defined for $K \geq 2$ categories and parameterized by the K-dimensional concentration parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \alpha_c > 0\, \forall c \in \{1, \ldots, K\}$. The probability density function of $\boldsymbol{\pi}$ is given by*

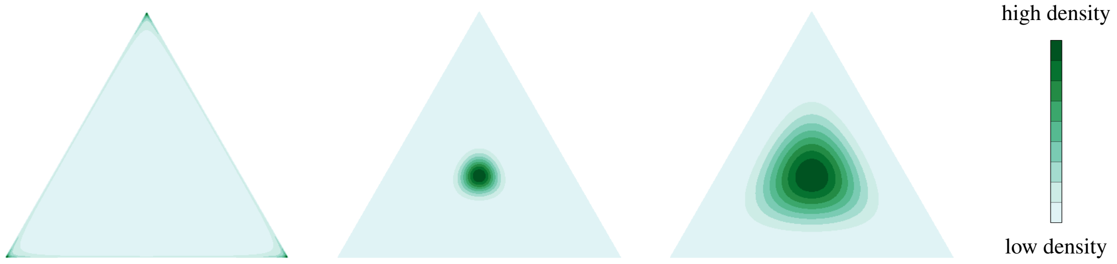
$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{c=1}^{K} \pi_c^{\alpha_c - 1} \quad ,$$

where $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{c=1}^{K} \Gamma(\alpha_c)}{\Gamma\left(\sum_{c=1}^{K} \alpha_c\right)}$ *is the* multivariate Beta function *which can be expressed using the* Gamma function Γ.
The symmetric form of the Dirichlet distribution implies equally expected class probabilities, i.e., $\mathbb{E}[\pi_c] = \frac{1}{K} \forall c \in \{1, \ldots, K\}$. *It is denoted as* $\mathrm{Dir}(\alpha \cdot \mathbf{1})$.

4.2 DirPA: Dirichlet prior augmentation

To illustrate the main idea, consider a labeled dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ of multivariate time series, where each $\boldsymbol{x}_i \in \mathbb{R}^{n_t \times d}$ has $n_t \leq T_{\max}$ time steps and d channels, and $y_i \in \mathcal{C} = \{1, \ldots, K\}$ is the corresponding class label. We denote $\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$, where $f_{\boldsymbol{\theta}} \colon \mathbb{R}^{T_{\max} \times d} \to \mathbb{R}^{K}$ is a model parameterized by $\boldsymbol{\theta}$ that maps observations $\boldsymbol{x}_i$ to a vector of logits $\boldsymbol{z}_i \in \mathbb{R}^{K}$. Subsequently, in general multi-class classification problems, the predictive distribution

$$\hat{\boldsymbol{p}}_i = \sigma(\boldsymbol{z}_i) = \sigma(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$$



(a) $\boldsymbol{\alpha} = (0.5, 0.5, 0.5)$ (b) $\boldsymbol{\alpha} = (30, 30, 30)$ (c) $\boldsymbol{\alpha} = (5, 5, 5)$

Figure 3. Dirichlet density for $K = 3$ (defined over the $(K-1) = 2$-simplex) and different concentration parameters $\boldsymbol{\alpha}$.

of K classes is obtained by applying the *Softmax* function σ to the logits.

In order to augment the training distribution and make the model more robust against unknown test skews, at each training step $s = 1, \ldots, S$, we introduce a class prior $\tilde{\boldsymbol{\pi}}^{(s)}$ to the model's output logits via the following steps:

1. **Sample pseudo-prior**: We sample a pseudo-prior vector

$$\tilde{\boldsymbol{\pi}}^{(s)} = (\tilde{\pi}_1^{(s)}, \ldots, \tilde{\pi}_K^{(s)}) \sim \mathrm{Dir}(\alpha \cdot \mathbf{1})$$

from the symmetric Dirichlet distribution $\mathrm{Dir}(\alpha \cdot \mathbf{1})$. The parameter $\alpha \in \mathbb{R}_+$ controls the degree of imbalance, with $\alpha < 1$ sampling highly skewed (imbalanced) distributions, and $\alpha > 1$ sampling distributions closer to uniform.

2. **Logit adjustment**: We use the sampled prior and a scaling factor $\tau \in \mathbb{R}_+$ to adjust the logits $\boldsymbol{z}_i$. The adjusted logits

$$\boldsymbol{z}'_i \leftarrow \boldsymbol{z}_i + \tau \log(\tilde{\boldsymbol{\pi}}^{(s)})$$

are computed element-wise.

Subsequently, the predictive probability distribution $\hat{\boldsymbol{p}}_i = \sigma(\boldsymbol{z}'_i)$ is computed by applying the *Softmax* function to the adjusted logits. Since we assume no prior knowledge of the actual test distribution, we sample from the symmetric Dirichlet distribution. The full pseudo-code is outlined in Algorithm 1. By applying DirPA, the model encounters many possible class-frequency scenarios and learns a representation that works under varied priors.

Algorithm 1 Dirichlet prior augmentation

```
Require: α, τ ∈ ℝ+
Require: f_θ (model parameterized by θ)
for each training step s = 1 to S do
    sample mini-batch of data points D^(s) = {(x_i, y_i)}_{i=1}^b
    sample pseudo-prior π̃^(s) ~ Dir(α · 1)
    for each data point (x_i, y_i) in D^(s) do
        compute base logits: z_i ← f_θ(x_i)
        adjust logits: z'_i ← z_i + τ log(π̃^(s))
        compute predictive distribution: p̂_i ← σ(z'_i)
    end for
    compute mini-batch loss L_batch ← (1/b) Σ_{i=1}^b L(p̂_i, y_i)
end for
```

4.3 Transformer model

All experiments, as described in Section 5, are conducted using a state-of-the-art Transformer encoder architecture with sinusoidal positional encoding (Vaswani et al., 2017; Schneider and Körner, 2021). We set the maximum sequence length T_{max} to 366 days, *i.e.*, a full year, including one leap day. This encoder serves as the model's feature extractor, which we call the *backbone*

$$f^{\text{backbone}}_{\boldsymbol{\theta}_{\text{backbone}}} \colon \mathbb{R}^{T_{\text{max}} \times d} \to \mathbb{R}^{n_e},$$

where $d \in \{1, \ldots, 13\}$ denotes the number of bands, $n_e \in \mathbb{N}$ the Transformer embedding dimension, and $\boldsymbol{\theta}_{\text{backbone}}$ all trainable model parameters of the backbone. The encoder's output is subsequently fed into a single linear layer to map the extracted features to the final class logits. We refer to this classification layer as the *head* of the model

$$f^{\text{head}}_{\boldsymbol{\theta}_{\text{head}}} \colon \mathbb{R}^{n_e} \to \mathbb{R}^{K}.$$

As before, $\boldsymbol{\theta}_{\text{head}}$ collects all of the head's trainable parameters.

The complete end-to-end model is given by the composition

$$f_{\boldsymbol{\theta}} = f^{\text{head}}_{\boldsymbol{\theta}_{\text{head}}} \circ f^{\text{backbone}}_{\boldsymbol{\theta}_{\text{backbone}}},$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\text{backbone}}, \boldsymbol{\theta}_{\text{head}}]$ represents all trainable model parameters.

5. Experiments

In all experiments, we use a single Transformer encoder block with four attention heads. Each token in the input sequence is represented by an internal embedding vector of dimension 128. This is further expanded by the fully connected network within the Transformer block, which employs a hidden dimension of $d_{\text{hidden}} = 256$. We apply additive sinusoidal temporal positional encoding (Vaswani et al., 2017) with a maximum sequence length of 366, encompassing daily samples over the span of a full year, including leap years, *cf.* Section 4.3.To predict class log-probabilities, we add a linear classification layer on top.

We train a randomly initialized network from scratch and fine-tune a model pretrained on the EUROCROPSML Latvian data (Reuss et al., 2026). For the pretrained model, we reset the classification head f^{head} and reinitialize it with the 129 classes from our target training set. All models are trained end-to-end for up to 200 epochs with a batch size of 16 following a standard training paradigm with train, validation, and test sets. Hyperparameters, such as the learning rate β, the focal focusing parameter γ, and the Dirichlet parameters α and τ, are optimized on the 1000 fixed validation data points. If the validation loss does not improve for 15 epochs, we stop training. The final models are evaluated on all 35 224 test samples, containing 102 unique classes. All experiments are repeated five times, each time with a different random seed $r \in \{0, 1, 42, 123, 1234\}$ in order to evaluate the robustness of the results.

https://doi.org/10.5194/isprs-annals-XI-3-2026-787-2026

To assess the impact of the prior adjustment across different few-shot scenarios, we train the models in various few-shot settings with 1, 5, 10, 20, 100, 200, or 500 shots (*cf.* Reuss et al., 2025). It is important to note that in our specific few-shot learning regime, tasks are sampled from the originally imbalanced training set. Consequently, the actual number of samples available per class is constrained by the underlying label distribution. Thus, as the target shot count increases, the number of samples per class can often be limited. This, in turn, results in the FSL prior $p_{\text{train}}(y)$ converging toward the true empirical prior of the full training set. Hence, the few-shot settings presented in this work test the method's generalization performance regarding both the balanced and empirical underlying training label distributions. In addition, we train the models on the entire Estonian training data to establish a baseline for the task's complexity in a standard (non-few-shot) setting. We refer to this setting with *all*.

We conduct all experiments with two different loss functions—*i.e.*, *cross-entropy loss (CE)* and FL (without a class-imbalance factor)—in order to evaluate the effect of the prior augmentation across various settings. Furthermore, using FL enables analysis of the impact of DirPA in combination with a class-agnostic, difficulty-aware loss function. Both loss functions are trained with and without Dirichlet priors.

All code used for training and evaluation is openly available at https://github.com/jsreu/DirPA.

6. Results and discussion

We choose the overall classification accuracy as our core validation metric. However, when working with highly imbalanced data, accuracy is often biased toward the majority class. Therefore, we also report *Cohen's kappa* (κ) as an additional evaluation metric for the final models. We always report the test metrics of the best-performing models, measured in terms of validation accuracy.

The results for the randomly initialized model are shown in Table 1a whilst those for the pretrained one are displayed in Table 1b. A graphical visualization of the results is provided in Figure 4. Results are reported using the two aforementioned loss functions: CE and FL. The postfix DirPA is appended to the name of the loss function if the DirPA method has been utilized.

Randomly initialized model Across all few-shot settings, the randomly initialized model improved or matched the baseline with DirPA. For CE, adding prior augmentation achieved higher overall accuracy and kappa scores across all few-shot tasks from 1- to 200-shot, with the largest gains observed in the low-shot regime ($k \leq 20$). The improvement remained up to the 500 setting, where both variants converged to nearly identical results. For FL, DirPA yielded higher scores in both metrics for nearly all few-shot configurations, except for 1 and 200 samples. The CE-based models achieved better performance in the 1-shot and 200-shot scenarios, while FL-based ones overtake 5-, 10-, and 20-shot.

Pretrained model When fine-tuning the model pretrained on Latvian data, DirPA improved both metrics in all few-shot regimes. CE DirPA achieved higher accuracy and Cohen's kappa, compared to CE, for every few-shot task, with differences shrinking in the *all* setting. For FL, the DirPA setting also surpassed the baseline across all settings. FL DirPA obtained the highest accuracy among all models in most few-shot settings, while CE DirPA was slightly better at 5, 200, or 500 shots. The difference between CE DirPA and FL DirPA narrowed as shots increased, and all variants exhibited similar performance in the *all* setting.

Across both initialization regimes, the Dirichlet prior consistently improves both accuracy and Cohen's kappa, confirming DirPA benefits model robustness under label imbalance. The largest relative advantage appears in the low-shot regimes. By dynamically sampling a skewed pseudo-prior vector $\tilde{\boldsymbol{\pi}}^{(s)}$ at every step, the method serves as a feature regularizer, especially in low-shot regimes. This forces the model to stabilize predictions where data is sparse. The results also indicate that DirPA does not degrade performance when the number of available samples increases. In higher-shot or full-data conditions, all methods converge to nearly identical results. This almost identical performance is as expected since the training prior converges toward the empirical prior of the full Estonia data, dissolving the positive effect of the Dirichlet priors. The improved kappa scores indicate that the model's predictions are more robust and less likely to be due to chance. While initial experiments suggest that macro metrics show inferior performance for the DirPA method (*cf.* Figure 4), this constitutes a necessary trade-off for achieving superior overall performance and stability: The concentrated loss often occurs on stable, high-shot classes, where DirPA's strong dynamic regularization over-smoothed already established decision boundaries. However, the consistent gain in overall system reliability (κ) justifies this small, concentrated loss on single classes, as the final model is demonstrably more robust for classifying the total volume of crops.

7. Conclusion and future outlook

This study proposed **Dir**ichlet **P**rior **A**ugmentation (DirPA), a novel method designed to bridge the gap between training and test priors in real-world few-shot crop-type classification. DirPA augments the balanced few-shot training data with dynamically sampled pseudo-priors from the Dirichlet distribution. This process acts as a robust regularizer, improving the model's generalization and stability on imbalanced test data. It is applied directly during the training process and does not require any knowledge about the final test distribution. We evaluated the method against the challenging task of classifying 102 heterogeneous, highly imbalanced crop types in Estonia. The evaluation involved two distinct loss functions, namely CE and FL. We demonstrated that DirPA improved overall accuracy and Cohen's kappa across various few-shot regimes. Although this study focused on crop-type classification, DirPA can be applied to any few-shot learning task that suffers from a discrepancy between the training and test label distributions.

While this study is limited to a single-country analysis and focused primarily on the feasibility of DirPA in a localized setting, we have already significantly broadened the scope of our investigation in an extended study (Reuss, Gikalo and Körner, 2026). We evaluate DirPA across a comprehensive, multi-country dataset that spans eight diverse agricultural and climatic regions within the *European Union (EU)*. This expansion allows us to rigorously assess the method's robustness and adaptability in a variety of agricultural and environmental contexts, moving well beyond the constraints of our original study. Moreover, we deliver a far more in-depth analysis by systematically examining how DirPA's performance varies with the degree of class imbalance and label distribution shift inherent in each dataset. Furthermore, we introduce a detailed hierarchical evaluation to demonstrate improvements not only at highly granular class levels but also for broader land cover categories. Finally, we provide a sensitivity analysis of DirPA’s hyperparameters. These new evaluations provide a clearer understanding of the method’s strengths and its potential for practical deployment in data-scarce real-world scenarios, where natural data imbalances and distribution shifts are the norm. In addition, we investigate applying pseudo-priors sampled from an asymmetric Dirichlet distribution while still assuming an unknown, but imbalanced test prior. However, results show that sampling asymmetric priors yields no performance improvement over the symmetric version. This suggests that the original method regularizes sufficiently and that class-specific concentration adds no discriminative advantages. We believe these advances collectively make a compelling case for the utility of DirPA in large-scale agricultural monitoring.

Acknowledgments

The project is funded by the German Federal Ministry for Economic Affairs and Energy based on a decision by the German Bundestag under the funding references 50EE2007B (J. Reuss, E. Gikalo, and M. Körner) and 50EE2105 (M. Körner).

Table 1. Few-shot classification results (*classification accuracy* and *Cohen's kappa*) on the test set, reported as mean ± standard deviation over five runs, *cf.* Section 5. The overall best result per few-shot scenario is highlighted in **bold blue** and the top result of the other loss function in **bold black**. The postfix DirPA indicates the use of the DirPA method.

(a) Randomly initialized network

	Algorithm/Loss	Benchmark task (k-shot)							
		1	5	10	20	100	200	500	all
accuracy	CE	0.105 ± 0.112	0.296 ± 0.117	0.390 ± 0.041	0.436 ± 0.019	0.462 ± 0.028	0.523 ± 0.048	0.624 ± 0.018	0.783 ± 0.003
	FL	0.094 ± 0.060	0.361 ± 0.083	0.406 ± 0.023	0.448 ± 0.022	0.459 ± 0.019	0.538 ± 0.036	0.607 ± 0.023	0.777 ± 0.006
	CE DirPA	**0.163 ± 0.094**	**0.408 ± 0.050**	**0.459 ± 0.009**	**0.477 ± 0.038**	**0.520 ± 0.038**	**0.571 ± 0.033**	**0.633 ± 0.010**	**0.784 ± 0.003**
	FL DirPA	**0.152 ± 0.135**	**0.437 ± 0.017**	**0.473 ± 0.020**	**0.483 ± 0.031**	**0.537 ± 0.035**	**0.557 ± 0.014**	**0.636 ± 0.013**	**0.779 ± 0.002**
kappa	CE	0.000 ± 0.016	0.144 ± 0.043	0.199 ± 0.033	0.229 ± 0.031	0.353 ± 0.041	0.426 ± 0.037	0.535 ± 0.016	0.709 ± 0.003
	FL	**0.028 ± 0.028**	0.131 ± 0.071	0.207 ± 0.041	0.251 ± 0.046	0.337 ± 0.040	0.435 ± 0.028	0.517 ± 0.022	0.701 ± 0.006
	CE DirPA	**0.040 ± 0.027**	**0.201 ± 0.052**	**0.275 ± 0.029**	**0.289 ± 0.059**	**0.400 ± 0.042**	**0.461 ± 0.026**	**0.536 ± 0.013**	**0.710 ± 0.004**
	FL DirPA	0.004 ± 0.023	**0.172 ± 0.098**	**0.287 ± 0.026**	**0.304 ± 0.047**	**0.396 ± 0.049**	**0.439 ± 0.034**	**0.538 ± 0.015**	**0.702 ± 0.002**

(b) Pretrained network

	Algorithm/Loss	Benchmark task (k-shot)							
		1	5	10	20	100	200	500	all
Accuracy	CE	0.215 ± 0.034	0.360 ± 0.026	0.377 ± 0.033	0.470 ± 0.025	0.519 ± 0.023	0.573 ± 0.023	0.631 ± 0.009	0.790 ± 0.004
	FL	0.216 ± 0.028	0.314 ± 0.059	0.350 ± 0.038	0.408 ± 0.056	0.479 ± 0.066	0.557 ± 0.045	0.612 ± 0.038	0.785 ± 0.003
	CE DirPA	**0.308 ± 0.048**	**0.491 ± 0.056**	**0.538 ± 0.042**	**0.594 ± 0.045**	**0.614 ± 0.025**	**0.637 ± 0.018**	**0.687 ± 0.008**	**0.791 ± 0.002**
	FL DirPA	**0.298 ± 0.049**	**0.510 ± 0.025**	**0.487 ± 0.092**	**0.586 ± 0.054**	**0.616 ± 0.020**	**0.626 ± 0.018**	**0.686 ± 0.018**	**0.788 ± 0.005**
kappa	CE	0.176 ± 0.017	0.289 ± 0.017	0.301 ± 0.032	0.381 ± 0.018	0.445 ± 0.019	0.493 ± 0.020	0.557 ± 0.008	0.717 ± 0.006
	FL	0.178 ± 0.012	0.258 ± 0.039	0.292 ± 0.025	0.341 ± 0.040	0.416 ± 0.055	0.483 ± 0.037	0.541 ± 0.035	0.713 ± 0.003
	CE DirPA	**0.215 ± 0.013**	**0.343 ± 0.036**	**0.400 ± 0.043**	**0.468 ± 0.032**	**0.503 ± 0.028**	**0.536 ± 0.030**	**0.601 ± 0.009**	**0.719 ± 0.004**
	FL DirPA	**0.212 ± 0.021**	**0.367 ± 0.026**	**0.363 ± 0.066**	**0.470 ± 0.046**	**0.501 ± 0.027**	**0.534 ± 0.022**	**0.602 ± 0.015**	**0.714 ± 0.008**

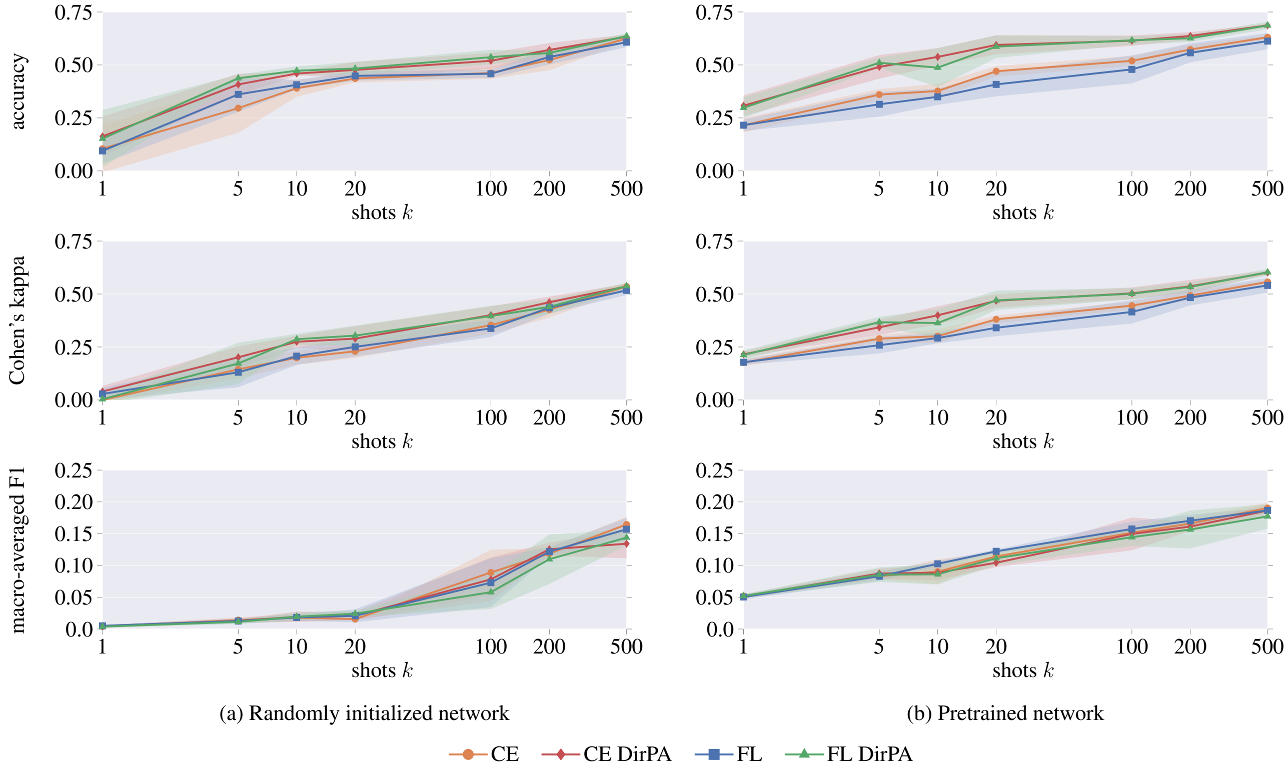


Figure 4. Visualization of test metrics (including macro-averaged F1-score) across the k-shot benchmark tasks, with the x-axis on a logarithmic scale. Metrics are shown as mean ± standard deviation over five runs, *cf.* Section 5. The postfix DirPA indicates the use of the DirPA method. Due to the highly imbalanced 102-class task, the F1-scores remain numerically low, reflecting the difficulty of achieving high performance on the many low-resource classes.

https://doi.org/10.5194/isprs-annals-XI-3-2026-787-2026

References

Alem, A., Kumar, S., 2022. Transfer learning models for land cover and land use classification in remote sensing image. In: *Applied Artificial Intelligence* 36.1, 2014192. DOI: 10.1080/08839514.2021.2014192.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Yu-Chiang, F. W., Huang, J.-B., 2019. A closer look at few-shot classification. In: *International Conference on Learning Representations (ICLR)*.

Claverie, M., Chan, A., See, L., Ramos, H., Koeble, R., Yordanov, M., Skøien, J., Urbano, F., d'Andrimont, R., Schneider, M., Körner, M., van der Velde, M., 2026. EuroCrops v2.0: Multi-annual harmonized parcel level crop type data linked to European Union-wide survey, statistical and Earth Observation products. In: *Earth System Science Data Discussions* 2026, pp. 1–31. DOI: 10.5194/essd-2025-752. under review.

Daumé III, H., Marcu, D., 2005. A Bayesian model for supervised clustering with the Dirichlet process prior. In: *Journal of Machine Learning Research* 6.53, pp. 1551–1577.

Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In: *International Conference on Machine Learning (ICML)*. Proceedings of Machine Learning Research, pp. 1126–1135.

Kluger, D. M., Wang, S., Lobell, D. B., 2021. Two shifts for crop mapping: Leveraging aggregate crop statistics to improve satellite-based maps in new regions. In: *Remote Sensing of Environment* 262, 112488. DOI: 10.1016/j.rse.2021.112488.

Kurian, V., Jacob, V., Kuruvilla, J., 2024. Approach of transfer learning in remote sensing image classification. In: *International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, pp. 1–3. DOI: 10.1109/ICTEST60614.2024.10576178.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2, pp. 318–327. DOI: 10.1109/TPAMI.2018.2858826.

Lipton, Z., Wang, Y.-X., Smola, A., 2018. Detecting and correcting for label shift with black box predictors. In: *International Conference on Machine Learning (ICML)*. Vol. 80. Proceedings of Machine Learning Research, pp. 3122–3130.

Mohammadi, S., Belgiu, M., Stein, A., 2024. Few-shot learning for crop mapping from satellite image time series. In: *Remote Sensing* 16.6, 1026. DOI: 10.3390/rs16061026.

Ochal, M., Patacchiola, M., Vazquez, J., Storkey, A., Wang, S., 2023. Few-shot learning with class imbalance. In: *IEEE Transactions on Artificial Intelligence* 4.5, pp. 1348–1358. DOI: 10.1109/TAI.2023.3298303.

Qi, Y., Bitelli, G., Mandanici, E., Trevisiol, F., 2023. Application of deep learning crop classification model based on multispectral and SAR satellite imagery. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-1/W2-2023, pp. 1515–1521. DOI: 10.5194/isprs-archives-XLVIII-1-W2-2023-1515-2023.

Rademacher, P., Doroslovački, M., 2021. Bayesian learning for regression using Dirichlet prior distributions of varying localization. In: *IEEE Statistical Signal Processing Workshop (SSP)*, pp. 236–240. DOI: 10.1109/SSP49050.2021.9513745.

Raghu, A., Raghu, M., Bengio, S., Vinyals, O., 2019. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In: *International Conference on Learning Representations (ICLR)*.

Reuss, J., Gikalo, E., Körner, M., 2026. DirPA: Addressing prior shift in imbalanced few-shot crop-type classification. In: *ISPRS Journal of Photogrammetry and Remote Sensing*. arXiv: 2603.12905. Submitted.

Reuss, J., Macdonald, J., Becker, S., Gikalo, E., Schultka, K., Richter, L., Körner, M., 2026. Benchmarking for practice: Few-shot time-series crop-type classification on the EuroCropsML dataset. In: *ISPRS Open Journal of Photogrammetry and Remote Sensing* 19, 100117. DOI: 10.1016/j.ophoto.2026.100117.

Reuss, J., Macdonald, J., Becker, S., Richter, L., Körner, M., 2025. The EuroCropsML time series benchmark dataset for few-shot crop type classification in Europe. In: *Scientific Data* 12.1, 664. DOI: 10.1038/s41597-025-04952-7.

Rouba, M., Larabi, M. E. A., 2023. Improving remote sensing classification with transfer learning: Exploring the impact of heterogenous transfer learning. In: *Engineering Proceedings* 56.1, 316. DOI: 10.3390/ASEC2023-15505.

Rußwurm, M., Wang, S., Körner, M., Lobell, D., 2020. Meta-learning for few-shot land cover classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 788–796. DOI: 10.1109/cvprw50498.2020.00108.

Saini, R., Ghosh, S. K., 2018. Crop classification on single date Sentinel-2 imagery using random forest and support vector machine. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-5, pp. 683–688. DOI: 10.5194/isprs-archives-XLII-5-683-2018.

Schneider, M., Broszeit, A., Körner, M., 2021. EuroCrops: A pan-European dataset for time series crop type classification. In: *Conference on Big Data from Space (BiDS)*. From Insights to Foresight. Publications Office of the European Union. DOI: 10.2760/125905.

Schneider, M., Körner, M., 2021. [Re] Satellite image time series classification with pixel-set encoders and temporal self-attention. In: *ReScience C* 7.19 (2): *ML Reproducibility Challenge*. DOI: 10.5281/zenodo.4835356.

Schneider, M., Schelte, T., Schmitz, F., Körner, M., 2023. EuroCrops: The largest harmonized open crop dataset across the European Union. In: *Scientific Data* 10.1, 612. DOI: 10.1038/s41597-023-02517-0.

Sipka, T., Sulc, M., Matas, J., 2022. The hitchhiker's guide to prior-shift adaptation. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2031–2039. DOI: 10.1109/WACV51458.2022.00209.

Steck, H., Jaakkola, T., 2002. On the Dirichlet prior and Bayesian regularization. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 15.

Tseng, G., Kerner, H., Rolnick, D., 2022. TIML: Task-informed meta-learning for agriculture. arXiv: 2202.02124.

Tseng, G., Zvonkov, I., Nakalembe, C., Kerner, H., 2021. CropHarvest: A global dataset for crop-type classification. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Datasets and Benchmarks Track. Vol. 1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30, 15.

Veilleux, O., Boudiaf, M., Piantanida, P., Ben Ayed, I., 2021. Realistic evaluation of transductive few-shot learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34, pp. 9290–9302.

Wang, S., Rußwurm, M., Körner, M., Lobell, D. B., 2020. Meta-learning for few-shot time series classification. In: *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 7041–7044. DOI: 10.1109/IGARSS39084.2020.9441016.