Optimizing Crop Type Mapping for Fairness

ILYA GORBUNOV

October, 2024

SUPERVISORS:
Dr. C.M. Gevaert

Dr. M. Belgiu

Optimizing Crop Type Mapping for Fairness

ILYA GORBUNOV
Enschede, The Netherlands, October, 2024

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Spatial Engineering

SUPERVISORS:

Dr. C.M. Gevaert

Dr. M. Belgiu

THESIS ASSESSMENT BOARD:

Prof.dr.ir. A. Stein (Chair)

Dr. M. Rußwurm (External Examiner, Wageningen University)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

The objective of this study was to improve fairness in crop type mapping by applying fairness optimization methods to address both class and sensitive attribute imbalances using a transformer classifier. Parcel size was identified as the key sensitive attribute, with a significant disparity in performance between small and large parcels. The methods tested included Random Oversampling (RO), Weighted Cross Entropy (WCE), Focal Loss (FL), and two novel approaches that targeted both class imbalance and the performance disparity between small and large parcels: Random Oversampling with Resampling (RO-R), which increased representation of smaller parcels by redistributing random samples, and Double Objective Weighted Cross Entropy (DOWCE), which applied higher penalties to misclassification of smaller parcels. Hybrid methods, RO-DOWCE and RO-FL, were also evaluated. These methods were tested on diverse datasets subsampled from the *BreizhCrops* dataset, covering Brittany, France, to assess their generalizability under varying conditions. Results showed RO-DOWCE was most effective at addressing class imbalance across the datasets, though not significantly different from RO-R, RO, and RO-FL. Cost-sensitive methods were generally less efficient compared to sample balancing and hybrid approaches. While all methods improved performance for both small and large parcels, reductions in the disparity between the two groups were marginal. The higher prevalence of mixed pixels in smaller parcels likely introduced noise that limited improvements in their classification performance, as increasing representation alone was insufficient to overcome this limitation. These findings suggest that additional strategies are needed to fully address the performance imbalance between small and large parcels.

AKNOWLEDGEMENTS

I would like to thank the chain of causal events that brought me to this moment—from the Big Bang, and whatever might have come before, to this very instant as I write these words. Without the accumulated wealth of human knowledge, from the pre-Socratic philosophers to the modern scholars who have more directly shaped this work, none of this would have been possible. Among these current scholars, I am deeply grateful to Marc Rußwurm and his colleagues, who made publicly available the dataset used in this research. Without his and his colleagues' efforts, I would have undoubtedly spent much more time on data gathering and coding instead of focusing on research. Furthermore, my deepest gratitude goes to my supervisors, whose support and guidance have been essential. Their insights have not only shaped this thesis but have also helped me grow as a scientist. Additionally, I would like to thank the Faculty of Geo-Information Science and Earth Observation (ITC). Their amazing staff have provided me with a great wealth of knowledge and skills during my master's studies, and their computing infrastructure has made this research possible. Finally, I want to express my infinite gratitude and love to my partner, who has stood by my side through all the highs and lows. Her unwavering support, love, and encouragement have been the foundation of this journey. Just as the Big Bang propelled the universe to expand with immense energy, you have been the driving force behind my growth and success. To you, everyone, and everything else that is and was, thank you.

TABLE OF CONTENTS

1. Introduction....1
    - 1.1. Fairness in crop type mapping....2
    - 1.2. Deep learning in crop type mapping....3
    - 1.3. Fairness optimization....4
    - 1.4. Research objectives and questions....4
2. Dataset....7
3. Methodology....9
    - 3.1. Experimental setup....10
    - 3.2. Dataset pre-processing and creation....11
    - 3.2.1. Pre-processing: Removal of thin parcels....11
    - 3.2.2. Dataset reduction....12
    - 3.2.3. Candidate dataset creation and selection of final datasets....12
    - 3.3. Sensitive attribute identification....13
    - 3.4. Fairness optimization methods....14
    - 3.4.1. Sample balancing methods....14
    - 3.4.2. Cost-sensitive methods....16
    - 3.4.3. Hybrid methods....19
    - 3.5. Identifying the parcel area threshold....20
    - 3.6. Evaluation....20
    - 3.6.1. Metrics....21
    - 3.6.2. Effects of dataset characteristics....22
    - 3.6.3. Statistical evaluation....22
    - 3.6.4. Evaluation of trade-offs....23
4. Results....24
    - 4.1. Dataset creation....24
    - 4.1.1. Pre-processing: Removal of thin parcels....24
    - 4.1.2. Dataset reduction....25
    - 4.1.3. Candidate dataset creation and selection of final datasets....26
    - 4.2. Sensitive attribute identification....27
    - 4.3. Identifying parcel area threshold....28
    - 4.4. Effects of dataset characteristics....29
    - 4.5. Evaluation results of the fairness optimization methods....31
5. Discussion....35
    - 5.1. A wicked problem case study: compliance checks in the EU....37
    - 5.1.1. CAP compliance monitoring system....38
    - 5.1.2. Stakeholders....38
    - 5.1.3. Conflicts....39
    - 5.1.4. Recommendations....41
6. Conclusion....43

References....44

Data avalaibility....53

Appendix: Auxiliary experiments and datasets....54

LIST OF FIGURES

Figure 1 Relationship between the bias and the area for 42 crop classes from US cropland data (Waldner et al., 2019) 2

Figure 2 Map of Brittany, France, partitioned into NUTS-3 regions. 7

Figure 3 Methodological framework diagram.......... 9

Figure 4 Visual representation of the RO-R method: the before and after parcel area distribution for dataset AD-5 temporary meadows class. Note, the before resampling dataset has had random oversampling applied to it.......... 16

Figure 5 Visual representation of the DOWCE weight function for the temporary meadows class of AD-5 ..19

Figure 6 Macro F1 scores across area perimeter ratio bins. 24

Figure 7 Macro F1 Scores across groups for the candidate sensitive attributes. Details of the chemical attributes are explained by Ballabio et al. (2019) 28

Figure 8 Piecewise linear regression model with the highest average R^2 value, identifying a threshold of 6,170 square meters for splitting the dataset in small and large parcels. 29

Figure 9 Relationship between performance and dataset characteristics for the baseline model results. 30

Figure 10 Baseline performance results across datasets. 31

Figure 11 Trade-offs in performance gains across metrics. (a) Gain in Macro F1 versus gain in Macro F1 Difference; (b) Gain in Macro F1 for large parcels versus gain in Macro F1 for small parcels. Each plot shows the results for all datasets as well as the average gain per method across datasets. 32

Figure 12 Box plots showing the rank distribution of methods for (a) Macro F1 and (b) Macro F1 Difference. Outliers are identified using the interquartile range rule. 33

Figure 13 Nemenyi critical distance diagrams comparing balancing methods and including the baseline, for (a) Macro F1 and (b) Macro F1 Difference. Methods connected by red horizontal lines are not significantly $(\alpha = 0.05)$ different from each other. The x-axis represents the average rank values. 34

LIST OF TABLES

Table 1 Class-wise samples before and after thin sample removal.......... 25

Table 2 Performance for full and capped datasets.......... 26

Table 3 Description of the experimental datasets. Majority class acronyms are bolded, with acronyms explained in Table 1. The subscripts indicate the number of samples in each class, where (very low) **VL** = 500, (low) **L** = 1,000, (medium) **M** = 2,500, (high) **H** = 5,000, and (very high) **VH** = 10,000.......... 27

Table 4 Stakeholder conflict matrix. The first conflict represents a conflict within the stakeholder group....... 40

Table 5 Auxiliary datasets. Note that AD-1 is the only dataset with thin parcels.......... 54

Table 6 Auxiliary experiments.......... 54

ACRONYMS

AD: Auxiliary Datasets

AE: Auxiliary Experiments

CAP: Common Agricultural Policy

CD: Critical Distance

CNN: Convolutional Neural Network

DG AGRI: Directorate-General for Agriculture and Rural Development of the European Commission

DOWCE: Double Objective Weighted Cross Entropy

DL: Deep Learning

ED: Experimental Datasets

FL: Focal Loss

FP: False Positive

FN: False Negative

KDE: Kernel Density Estimation

ML: Machine Learning

OA: Overall Accuracy

RF: Random Forest

RO: Random Oversampling

RO-R: Random Oversampling with Resampling

SVM: Support Vector Machine

WCE: Weighted Cross Entropy

1. INTRODUCTION

Global dietary diversity has been decreasing, leading to a more homogeneous global food supply that relies primarily on a handful of staple crops (Khoury et al., 2014). This over-reliance overshadows 'minor' or underutilised crops, which are increasingly recognized for their vital role in global food security. Minor crops have been shown to positively impact agricultural income and dietary diversity for smallholder households in the developing world (Pellegrini & Tasciotti, 2014; Rajendran et al., 2017). Many minor crops possess traits that make them more resilient to hostile environments, allowing farmers to better match crops to specific conditions (Cheng et al., 2017). This resilience enhances their potential to withstand climate change-related challenges. Renard and Tilman (2019) reinforced this further by analysing half a century's worth of data on 176 crop species from various countries, finding that greater food diversity leads to more stable food systems. They noted that increased diversity could mitigate instability caused by unpredictable precipitation patterns that could arise due to climate change. Therefore, crop diversification through the use of minor crops plays an essential role in achieving food security, improving nutrition, and promoting sustainable agriculture. These goals align with the United Nations' second Sustainable Development Goal (UN, 2015) and are reflected on most objectives of the United Nations' Food and Agriculture Organization (FAO, 2019).

Policies that incentivize crop diversification have enormous potential to address the growing challenges concerning global food security. A critical component in effectively implementing such policies is the use of agricultural monitoring systems, which have seen a significant increase in demand due to the need for comprehensive agricultural information (Becker-Reshef et al., 2010). In these systems, crop classification with machine learning (ML) and remote sensing data plays a crucial role. This process creates cropland maps that can be combined with other data to predict yield, monitor growth, assess crop health, and provide other crucial information for decision-making in agriculture (Fritz et al., 2019). To monitor the implementation of crop diversification policies effectively, it is important that these monitoring systems achieve similar accuracy for minor crops as they do for major crops. However, minor crops generally have fewer samples in the training dataset, making them a minority class in ML terminology. If no measures are taken to counter this imbalance, classifications tend to be biased against minority classes, a phenomenon known as the class imbalance problem (Elrahman & Abraham, 2013).

This bias against minor crops is exemplified in Waldner et al.'s (2019) study. Their analysis of U.S. cropland data revealed a clear relationship: as the area occupied by a crop class decreases, the absolute classification bias increases. In other words, the less area a crop occupies, the more likely it is to be misclassified. This relationship is illustrated in Figure 1. In crop type mapping, this bias against minority classes leads to an inaccurate representation of the ground truth. Since these systems rely on accurate classification data (Fritz et al., 2019), this hinders the tracking and evaluation of crop diversification policies and compromises the reliability of agricultural monitoring systems.

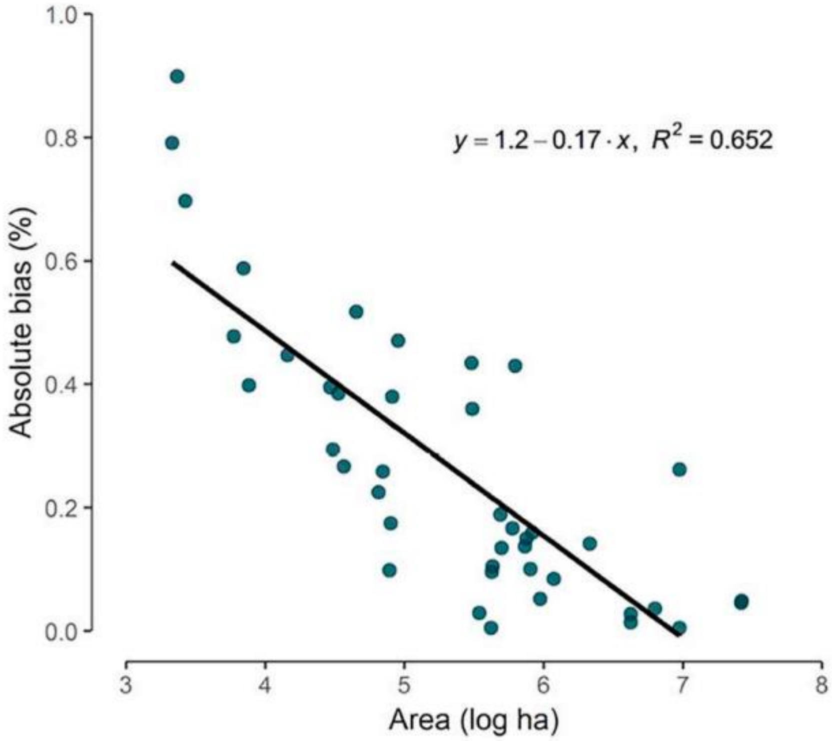

Figure 1 Relationship between the bias and the area for 42 crop classes from US cropland data (Waldner et al., 2019)

1.1. Fairness in crop type mapping

ML fairness is a field focused on ensuring that biases in data and algorithms do not lead to unfavourable treatment of individuals or groups based on sensitive attributes such as gender or race (Oneto & Chiappa, 2020). This objective is also emphasized in various responsible AI guidelines and recommendations (Akbarighatar et al., 2023; Tahaei et al., 2023; Tomašev et al., 2020; UNESCO, 2022).

Applying fairness to crop type mapping requires identifying how unfavourable treatment based on sensitive attributes can occur. Since the primary function of crop type mapping is accurate crop classification, the root of unfavourable outcomes could be boiled down to misclassification errors. For example, in the EU, member states may use remote sensing to monitor compliance with the Common Agricultural Policy (CAP), ensuring that declared crops match actual crops (Devos et al., 2018). False-positive errors—where a compliant farmer is incorrectly flagged as non-compliant—can disproportionately affect minor crop farmers. This leads to outcome disparities where these farmers are flagged as non-compliant more often than those growing major crops, resulting in potential unfavourable consequences, such as delays in subsidy payments.

In ML fairness, such outcome disparities are particularly relevant when they occur based on sensitive or protected attributes (Oneto & Chiappa, 2020). Farmers cultivating minor crops can be considered a sensitive group. For example, in Nepal, finger millet—a minor crop—is experiencing a decline in sales as it is replaced by more common staples like maize, posing risks to the livelihoods of farmers growing finger millet (Pallante et al., 2016). Moreover, the higher misclassification rate of minor crops can directly lead to negative consequences, as the potential delayed subsidy payments in the CAP compliance monitoring example.

I argue that, in crop type mapping, fairness can be considered along two key dimensions. The first dimension relates to class imbalance. The second dimension, then concerns sensitive attributes that are class independent. The goal of this research is to identify an appropriate sensitive group and implement methods to address their higher misclassification rate. Small parcels are a likely candidate for a sensitive group, as they have been found to be more prone to misclassification in both Sentinel-1 and Sentinel-2 data in Germany (Orynbaikyzy et al., 2020). This group is societally significant because, for example, in the EU, farm size is a key factor in farm income distribution (Loughrey et al., 2016).

Given the focus of fairness in crop type mapping on reducing misclassification errors between groups—whether defined by class or sensitive attributes—this aligns with the statistical fairness criterion known as separation (Barocas et al., 2023). This criterion, also referred to as equality of odds posits that the model's error rates should be balanced across all groups to ensure fairness.

1.2. Deep learning in crop type mapping

The application of deep learning (DL) in satellite image classification has garnered significant attention, achieving impressive accuracies in tasks such as satellite Image time series classification (Moskolaï et al., 2021) and Hyperspectral Image classification (S. Li et al., 2019). In hyperspectral image classification, deep neural networks have outperformed traditional methods like random forest (RF), with the performance gap widening with an increased number of training samples (Shadman Roodposhti et al., 2019). A meta-analysis of DL applications in remote sensing reported that DL methods for land use and land cover classification have a median overall accuracy of approximately 91%, significantly higher than that of RF and support vector machine (SVM) classifiers (Ma et al., 2019). However, another meta-analysis examining 266 studies found that SVM generally outperforms other methods, with neural networks coming in a close second (Khatami et al., 2016). However, this last meta-analysis didn't discriminate between different neural network architectures, grouping them all in one category.

Transformers, a type of DL architecture using self-attention mechanisms to process sequential data concurrently, have emerged as a breakthrough in DL (Vaswani et al., 2017). They have demonstrated high accuracy in land cover classification tasks. For instance, in classifying raw, unprocessed Sentinel-2 time series data, self-attention and recurrent neural networks have outperformed convolutional neural networks (CNNs) on raw satellite time series (Rußwurm & Körner, 2020).

Despite their success, a systematic review of 90 papers on DL applications in crop type mapping and yield prediction found only two studies employing transformers (Joshi et al., 2023). Moreover, there is a research gap in improving fairness and addressing class imbalance in transformer-based classifications. A survey by Johnson & Khoshgoftaar (2019) revealed that most studies focusing on class imbalance in DL used CNNs, with none utilizing transformers. However, while this survey was published shortly after transformers were introduced, this gap remains unaddressed. To date, no studies specifically tackle the class imbalance problem in crop type mapping using transformers.

1.3. Fairness optimization

Addressing the first dimension of fairness (class imbalance) has been widely investigated in crop type mapping. For example, Waldner et al. (2019) tested various sample balancing methods across crop type mapping datasets to assess how well they minimize bias. However, the second dimension, concerning sensitive attributes independent of crop classes, remains largely unexplored, with only one recent study incorporating fairness in the context of crop yield estimation (E. He et al., 2023). To address this gap, this research aims to explore and evaluate fairness optimization methods in how well they improve both dimensions of fairness in crop type mapping.

Class imbalance is typically addressed using two main approaches: sample balancing and cost-sensitive learning. Sample balancing, also referred to as resampling or data-level methods, involves modifying the training data distribution. This can be done by under-sampling majority classes, over-sampling minority classes, or a combination of both, thereby making the training dataset more balanced (Susan & Kumar, 2021). Cost-sensitive learning, on the other hand, works at the classifier level by assigning higher penalties to misclassifications of minority classes, forcing the classifier to be more attentive to underrepresented classes (Geng & Luo, 2019). Also, given that sample balancing and cost-sensitive methods operate at different levels, they can be combined to form hybrid methods (Buda et al., 2018). The two main approaches have been studied in the context of DL classification, particularly with CNNs. However, so far there has been no research comparing these approaches under a consistent experimental setup (Johnson & Khoshgoftaar, 2019). As such, a thorough comparison of both approaches in the context of a crop type mapping problem using transformers is not a domain specific research gap and contributes to the overall field of ML.

To address fairness in relation to sensitive attributes, several approaches have been proposed. In this research, I will focus on two methods that mirror the logic of sample balancing and cost-sensitive techniques used to tackle class imbalance. One example of sample balancing is demonstrated by Yang et al. (2020), who proposed balancing demographic distributions within a large-scale image dataset to mitigate biases against underrepresented groups. An example of a cost-sensitive approach is provided by Serna et al. (2022), who introduced a method that incorporates demographic information into the loss function to improve fairness in face recognition models. Unlike class imbalance correction methods, which are more generalizable, these fairness optimization approaches are context-specific, requiring tailored interventions based on the sensitive attribute in question.

The aim of this study is to implement cost-sensitive, sample balancing, and hybrid methods specifically designed to improve fairness for sensitive attributes in crop type mapping. These methods will be built on the premise that the performance disparity experienced by the identified sensitive group arises from dynamics like those observed in class imbalance correction methods. Essentially, the research seeks to determine whether increasing the representation of the sensitive group—either within the dataset or during the training process—can alleviate the observed performance disparity.

1.4. Research objectives and questions

The primary goal of this research is to evaluate different fairness optimization methods and their ability to improve both dimensions of fairness—class imbalance and sensitive attribute fairness—using a transformer-based classifier for crop type mapping. To ensure that the observed performance improvements are generalizable across different datasets, these methods will be tested on datasets with varying characteristics, which can be generated by subsampling a larger dataset. This approach also facilitates statistical significance testing to determine whether one method's superiority over others holds consistently across different datasets. The main objective and research question, along with the accompanying sub-objectives and sub-research questions, are outlined below.

Main Objective		**Research questions**
To compare different fairness optimization methods, in terms of sensitive attribute imbalance as well as class imbalance, in the context of crop type mapping using a transformer classifier.	RQ 1	How do the fairness optimization methods compare to each other and the baseline results across datasets with varying characteristics?

Sub-objectives		**Sub-research questions**
Establish datasets with varying characteristics to enable the comparison of fairness optimization methods.	RQ 2	Which characteristics (e.g., number of classes) to vary in the subsampled datasets?
Identify the sensitive attribute suitable for targeted fairness optimization.	RQ 3	Which candidate sensitive attributes have a sensitivity to classification performance and carry societal relevance?
Determine and implement appropriate sample balancing methods to optimize for fairness.	RQ 4-1	In the context of transformers, which sample balancing methods optimize for class imbalance?
	RQ 4-2	How to adapt sample balancing class imbalance correction methods to address fairness of sensitive attributes?
Determine and implement appropriate cost-sensitive methods to optimize for fairness.	RQ 5-1	In the context of transformers, which sample balancing methods optimize for class imbalance?
	RQ 5-2	How to adapt cost-sensitive class imbalance correction methods to address fairness of sensitive attributes?
Determine and implement appropriate hybrid methods to optimize for fairness.	RQ 6-1	In the context of transformers, which hybrid methods optimize for class imbalance?

	RQ 6-2	How to adapt hybrid class imbalance correction methods to address fairness of sensitive attributes?
Determine the impact of dataset characteristics on the performance of baseline models.	RQ 7	How do different dataset characteristics influence the performance of baseline classification models?

2. DATASET

The dataset used in this research is *BreizhCrops*, a benchmark time series dataset for crop type mapping using Sentinel-2 data and crop parcel labels and geometries from the *National Institute of Forest and Geography Information* for 2017 (Rußwurm et al., 2020). It encompasses agricultural parcels in Brittany, France, using mean aggregation of reflectance data for each parcel, inclusive of border pixels. Developed with practitioners in mind, *BreizhCrops* is publicly available and comes with an implementation via a GitHub repository. All the details of the dataset are documented thoroughly in the paper by Rußwurm et al. (2020), further in this chapter I explain specifics relevant to this research: class mapping, testing dataset, the chosen classifier architecture, and the processing level used.

Two pre-configured **class mappings** link crop labels to classes. The selected class mapping for this research, presented by Rußwurm et al. in a workshop during the 2019 International Conference on Machine Learning, consolidates 157 labels into 13 classes, resulting in a total of 763,000 parcels. This ample selection of samples across multiple classes is a highly suitable foundation for creating the subsampled datasets to test the fairness optimization methods.

Regional divisions further enhance the utility of *BreizhCrops*, with the dataset split into four NUTS-3 regions, as shown in Figure 2. This allows for regional splits between training and **testing** datasets, which improves the reliability of performance evaluation by minimizing the overestimation of the model's generalization abilities (Karasiak et al., 2022). In this research, the FRH04 region is consistently used for testing.

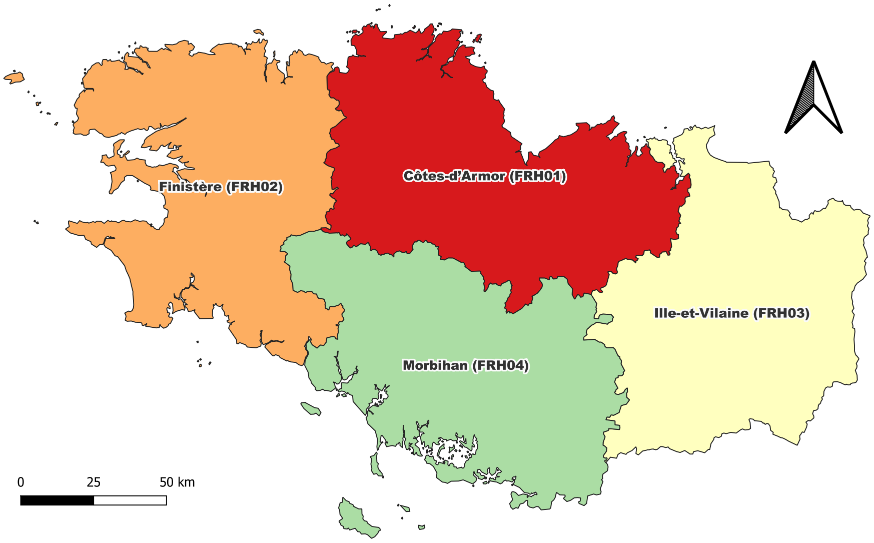

Figure 2 Map of Brittany, France, partitioned into NUTS-3 regions.

BreizhCrops also includes implementations of several DL **classifier architectures**, with the transformer model outperforming all others in the metrics tested (Rußwurm et al., 2020). Given its superior performance and state-of-the-art status, this architecture was selected for use in this research.

BreizhCrops provides data at two **processing levels**: raw reflectances at the top of the atmosphere (L1C) and atmospherically corrected surface reflectances (L2A). For this research, L1C is used because Rußwurm et al. (2020) tuned the hyperparameters for this level, and the same values are applied in all the models in this research. Interestingly, Rußwurm et al. (2020) found no significant performance difference between L1C and L2A, suggesting that it was due to hyperparameters being specifically optimized for L1C. However, a later study showed that L2A can outperform L1C when each processing level has its separate hyperparameter tuning (Rußwurm & Körner, 2020).

3. METHODOLOGY

This chapter outlines the methods used to achieve the research objectives. Figure 3 provides a high-level overview of the methodological framework. Throughout this chapter and the following sections, I will reference auxiliary experiments (AEs) and auxiliary datasets (ADs), which link to Table 5 and Table 6 in the Appendix: Auxiliary experiments and datasets. It is important to note that ADs differ from the experimental datasets (EDs), as ADs were used to support the primary experimental goal of testing different fairness optimization methods across the EDs, the datasets with varying characteristics. The following paragraphs will describe each step of the methodological framework, with references to the relevant subsections in brackets.

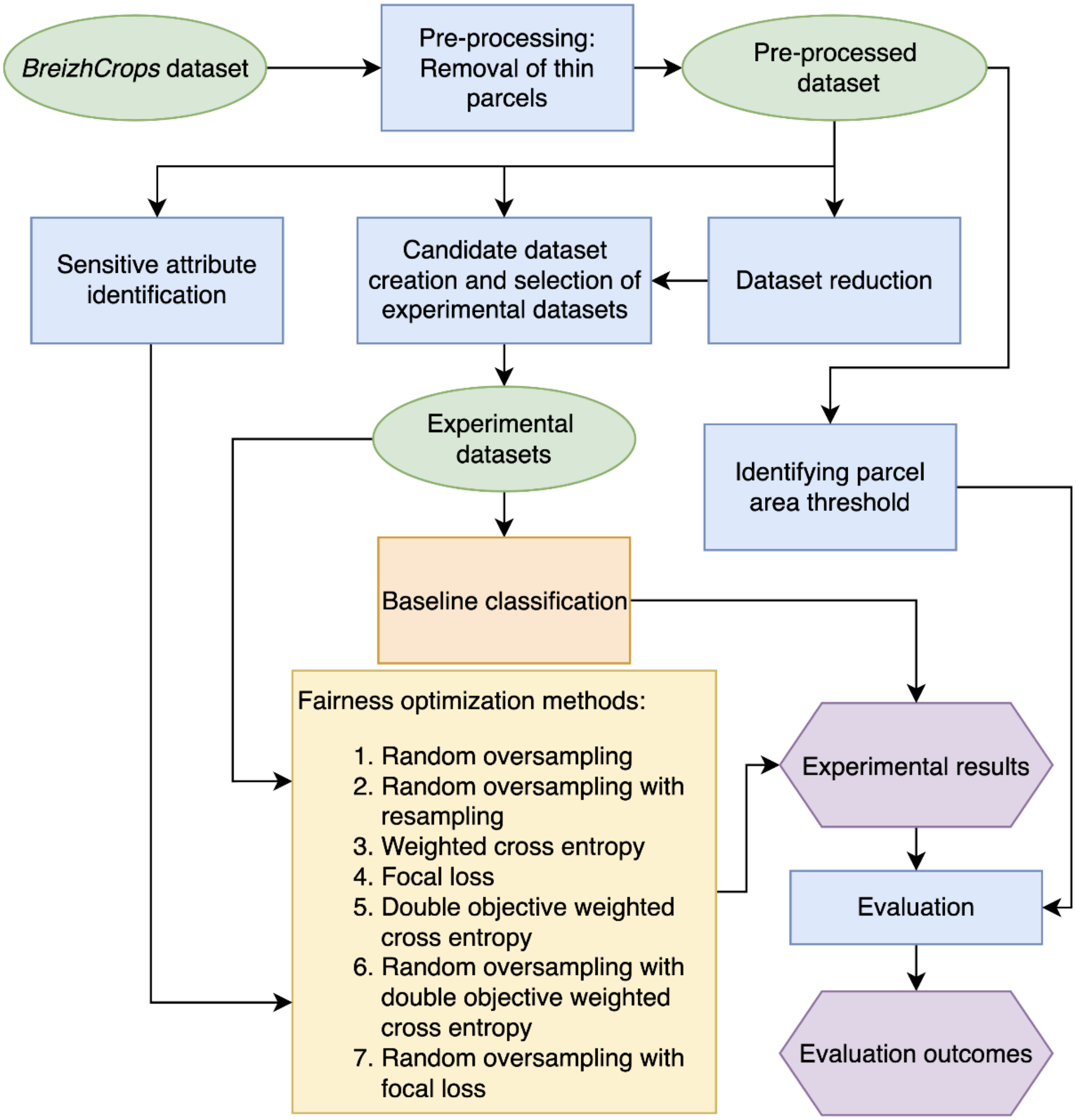
BreizhCrops dataset
Pre-processing: Removal of thin parcels
Pre-processed dataset
Sensitive attribute identification
Candidate dataset creation and selection of experimental datasets
Dataset reduction
Experimental datasets
Identifying parcel area threshold
Baseline classification
Fairness optimization methods:
1. Random oversampling
2. Random oversampling with resampling
3. Weighted cross entropy
4. Focal loss
5. Double objective weighted cross entropy
6. Random oversampling with double objective weighted cross entropy
7. Random oversampling with focal loss
Experimental results
Evaluation
Evaluation outcomes

Figure 3 Methodological framework diagram

The experimental setup section (3.1) covers the overall setup for the experiments, including the training-testing setup, hyperparameters for the transformer model, and other essential details for reproducibility.

The pre-processing and dataset creation section (3.2) outlines the steps undertaken to create the subsampled datasets used to test the fairness optimization methods. First, **pre-processing** of the ***BreizhCrops*** **dataset** (3.2.1) involved removing noisy thin parcels based on their area-to-perimeter ratio, which used the results of the first auxiliary experiment (AE 1). Then, **dataset reduction** (3.2.2) involved the second auxiliary experiment (AE 2), which capped the training and testing set to explore whether drastic reduction in samples maintained reasonable performance while significantly reducing the computational cost. The results of this step informed the **candidate dataset creation** step (3.2.3), where 300 candidate training datasets with varying class distributions and complexity levels were generated, and 10 of these were manually selected to be the **experimental datasets**.

In **sensitive attribute identification** (3.3), candidate socially sensitive attributes were explored for their impact on classification performance, with parcel area identified as the most sensitive to classification performance (AE 4). This led to the application of **fairness optimization methods** (3.4), which consisted of sample balancing, cost-sensitive learning, and hybrid methods.

Finally, the **experimental results** of both the fairness optimization methods and the **baseline classification** were fed into the **evaluation** process (3.6). This evaluation required the **identification of the threshold** separating small and large parcels (3.5), which was identified using piecewise linear regression (AE 5). The **evaluation outcomes** were derived through the use of a statistical evaluation (3.6.3), and a trade-off analysis (3.6.4), enabling a robust assessment of the fairness optimization methods. Additionally, the baseline classification results, along with dataset characteristics, were analysed to assess the impact of dataset variations on the performance of the baseline models (3.6.2).

3.1. Experimental setup

This section clarifies the experimental setup used in all experiments to ensure reproducibility and adherence to ML best practices. It explains the transformer classifier architecture used and its hyperparameters, how training was set up with early stopping, and what hyperparameters were tuned through k-fold cross-validation. All computing tasks were conducted on *CRIB*, a geospatial cloud computing platform suited for big data and DL tasks (Girgin, 2021).

All experiments use the transformer classifier from the *BreizhCrops* package, which is adapted from the original transformer architecture from Vaswani et al. (2017). Instead of handling sequence-to-sequence tasks, this version is tailored for sequence-to-label tasks. Full details of this adaptation are explained in detail by Rußwurm & Körner (2020).

Across all experiments, I used the hyperparameters that produced the lowest validation loss on the *BreizhCrops* dataset (Rußwurm et al., 2020). These are, a learning rate of 1.31e-3, a weight decay of 5.51e-8, 64 as the dimensionality of the hidden states, 3 self-attention layers, 1 attention head, an inner-layer dimensionality of 128, and a dropout rate of 0.4. While ideally, hyperparameters should be optimized for each dataset and method combination to achieve the best results, given the computational resources constraint, I use these hyperparameters throughout all experiments, which is a limitation of this study.

To ensure reproducibility in all experiments, I controlled the stochasticity inherent in DL models and other processes. This was achieved by setting a seed for all random processes in Python's random module, NumPy, and PyTorch. To verify that all stochasticity was controlled, I ran ED 1 two times with the baseline setup and the results were confirmed to be identical (AE 3). Note that the baseline classification refers to the barebones version of the model, with fairness optimization methods, using standard cross-entropy loss (Equation 2) as the objective function.

To prevent overfitting and ensure optimal model selection, early stopping was applied with a patience of 20 epochs, using the GMean metric (Equation 1). This metric is chosen as it is highly sensitive to poor performance across all classes. It was calculated on a validation set created by stratifying the training data with an 80/20 split. Training stopped if no improvement in GMean was observed for 20 consecutive epochs, and the model with the highest GMean was selected for final evaluation on the test set.

$$\mathrm{GMean} = \left(\prod_{i=1}^{C} \mathrm{Recall}_i\right)^{\frac{1}{C}}$$

Equation 1

I tuned two hyperparameters using k-fold cross-validation: the maximum small parcel weight for the Double Objective Weighted Cross Entropy (see section 3.4.2.3) and the gamma parameter for Focal Loss (see section 3.4.2.2). This was done by first doing an initial stratified 80/20 split of the data into training and validation sets. Then, the training set was further divided into five folds. In each iteration, one fold was used for validation, and the remaining four for training. This process was repeated for all parameter values, which are detailed in the respective section of each method. The GMean performance from each fold was averaged to determine the optimal hyperparameter values, which were then used in the final training stage.

3.2. Dataset pre-processing and creation

This section outlines the steps to create the EDs used to evaluate the fairness optimization methods. It describes the pre-processing steps to remove noisy samples, the dataset reduction tests, and the methods used to select the EDs.

3.2.1. Pre-processing: Removal of thin parcels

The first Auxiliary Experiment (AE 1) was conducted to identify and eliminate thin parcels, which added substantial noise due to their narrow geometries. Preliminary analysis revealed that many parcels were narrower than 10 metres, the minimum resolution of Sentinel-2 (B2, B3, B4, and B8). These thin parcels hindered classification performance, likely because of their high proportion of edge pixels.

To identify these thin parcels, the area-to-perimeter ratio was used. Lower ratio values indicated parcels that had relatively more perimeter compared to their area, suggesting a thin geometry. I ran a baseline classification using AD 1 to explore how the classification performance varied across different area-to-perimeter ratios.

This is achieved by plotting the macro F1 scores across bins of area-to-perimeter ratios to identify ranges where thin parcels might be hindering performance. To further confirm that the observed performance differences were primarily due to parcel thinness rather than small parcel size, a comparison was made within parcels of similar area ranges. The results of this step are detailed in section 4.1.1.

3.2.2. Dataset reduction

The second auxiliary experiment (AE 2) tested whether reducing the number of samples in both the training and testing datasets would significantly affect classification performance. This was done to find an optimal class-wise sample cap for the training and testing dataset separately, ensuring that computational resources were used efficiently without sacrificing performance.

To conduct this step, three datasets were used, all excluding the thin parcels identified earlier. The first dataset (AD 2) was the full dataset, with no cap on the number of samples per class. The second dataset (AD 3) was created by capping the training set at 10,000 samples per class for each of the 13 classes, reducing the training set size from 534,000 samples to 110,000. The third dataset (AD 4) applied the same 10,000 cap on the training set but also capped the testing set at 3,000 samples per class.

For each dataset, a baseline classification was run to measure the performance of two key metrics: overall accuracy (OA) and the GMean accuracy (Equation 1). The goal was to assess whether a reduced dataset could maintain similar performance while significantly reducing computational resources. The results of this step are showcased in section 4.1.2.

3.2.3. Candidate dataset creation and selection of final datasets

To create the EDs, 300 diverse candidate datasets were generated, from which 10 were manually selected to be the final EDs. Only the training sets were generated, as the testing set, consistently used across all EDs, had already been determined in the previous step.

To generate the final training datasets, 300 candidate datasets were first created by randomly selecting the number of classes (ranging from 4 to 13) and designating a subset as majority classes (those with the most samples), with the number of majority classes randomly chosen to be between one and half of the total classes. Majority classes were always assigned 10,000 samples, except for classes with fewer than 10,000 available samples for the training dataset (fallow, orchards, and protein crops). For non-majority classes, sample sizes were randomly assigned within fixed categories: "Very Low" (500 samples), "Low" (1,000 samples), "Medium" (2,500 samples), and "High" (5,000 samples). Also, it was verified that all available samples were selected across the candidate datasets.

The selected complexity measure was an adaptation of N1 (Ho & Basu, 2002). The full N1 calculation was not applied due to the computation cost of calculating it for all 300 candidate datasets. Instead, the N1 was modified from a metric that captures the complexity of the full dataset to one that quantifies the separability of the classes irrespective of the class imbalance, with N1 varying across datasets due to their different class compositions. This was done by randomly selecting 200 samples per class from each candidate dataset and calculating the N1 on this subset. This assumes that the 200 randomly selected samples per class are representative of the class, and that the sample size is large enough to minimize the effect of outliers. N1 was then calculated by computing pairwise Frobenius distances (Euclidean norm) between all samples, constructing a minimum spanning tree (MST) using Kruskal's algorithm, and dividing the number of inter-class edges in the MST by the total number of samples.

From the 300 candidate datasets, 10 were manually selected based on diversity in class distribution and complexity. The constraints were that these datasets should include all classes, cover all unique class numbers, and that each class served as the majority in at least one dataset. Additionally, I ensured that the chosen datasets had variations in the complexity measure, further ensuring diverse characteristics.

The selected EDs are presented in Table 3 section 4.1.3, along with their imbalance degree, a metric proposed by Ortigosa-Hernandez et al. (2017) for quantifying class imbalance in multi-class datasets. I used the Hellinger distance for this calculation, as recommended by Ortigosa-Hernandez et al. (2017), though other distance measures could be applied.

3.3. Sensitive attribute identification

In this section, I describe the process of identifying candidate attributes that are socially sensitive and assessing their impact on classification performance. The objective is to pinpoint an attribute that exhibits performance disparities across different value ranges, making it a prime candidate for fairness optimization. By identifying these disparities, targeted fairness optimization methods can be applied to minimize performance disparity between groups that can be distinguished using the attribute values. Further in this section, I first explain the selected candidates, and I finish by explaining how I selected the final sensitive attribute out of the candidates.

The first identified candidate attribute was **parcel area**, recognized early on by exploring the original classification results of *BreizhCrops*. It was observed that smaller parcels had a higher rate of misclassification. One possible explanation is that smaller farms may employ different management practices, though the evidence supporting this hypothesis for the study area is limited. The literature provides mixed conclusions on the relationship between farm size and management practices. For example, while Netting (1993) argues that smaller farms worldwide adopt more environmentally friendly practices relative to large scale farmers, Soule (2001) found that farm size and type were not strongly associated with soil and nutrient management in the context of U.S. corn farms. This complicates the hypothesis that smaller farms in this study area necessarily follow different management strategies. Another explanation is the higher ratio of border pixels in smaller parcels, which distorts the mean aggregation of spectral data, leading to less accurate classification. Dean & Smith (2003) showed that removing border pixels improves classification confidence, but it remains unstudied how effectively transformer classifiers mitigate this noise. Despite the uncertainty surrounding the exact cause of misclassification, farm size remains a socially sensitive attribute, as it is linked to economic inequality. As Loughrey et al. (2016) point out, in Western Europe, smaller farms often face structural disadvantages due to unequal access to resources and land concentration, which can hinder new-entrant and small-scale farmers from expanding their operations, thereby perpetuating income disparities within the agricultural sector.

The next candidate attribute was **distance to the nearest city**. This attribute could be sensitive to performance due to air quality, which deteriorates near urban areas (C. Li et al., 2019). This could potentially influence crop growth and the spectral signatures due to the presence of smog. Additionally, proximity to urban centres is known to affect agricultural land values, with wealthier farmers more likely to be located closer to cities (Guiling et al., 2009), making distance to cities relevant from a societal sensitivity standpoint. To assign this attribute to each sample, I calculated the Euclidean distance (in kilometres) from the centroid of each parcel to the nearest city. Only cities with more than 35,000 inhabitants, as per the 2021 official census, were included.

The final group of candidate attributes were **chemical properties of topsoil**, sourced from the European Soil Data Centre (Ballabio et al., 2019). The chemical properties analysed in this study included pH, nitrogen content, carbon-nitrogen ratio, and cation exchange capacity. It was taught these properties could affect crop growth and health by influencing nutrient availability and soil health, which in turn impact the spectral signatures used in classification. Additionally, it can be argued that these properties are societally sensitive, as areas with poor topsoil might consistently be disadvantaged, in both classification accuracy and crop yield, exacerbating existing inequalities. For each parcel, the chemical properties were determined by sampling the value of the chemical property raster at the centroid of the parcel geometry.

With these candidate attributes identified, their performance sensitivity across value ranges was plotted to identify the most sensitive attribute. This is referred to as the fourth auxiliary experiment (AE 4). The experiment used the baseline results of the dataset AD 5, which includes the full training set but shares the same test set as the EDs, ensuring that performance sensitivity is observed in the same test set that will be used to evaluate the fairness optimization methods. To assess the sensitivity of each candidate attribute, I divided their values into 20 quantile-based groups, each containing an approximately equal number of data points. For each group, I calculated the F1 score for each class and averaged these into a macro F1 score. The results are presented as side-by-side plots of the macro F1 scores for all attributes across these groups. Since parcel size showed a crystal-clear sensitivity to performance while other attributes did not, a more mathematically rigorous criterion to determine the most sensitive attribute was deemed unnecessary. The results are presented in section 4.2.

3.4. Fairness optimization methods

In this section, I outline the fairness optimization methods used, they include two novel approaches: Random Oversampling with Resampling (RO-R) and Double Objective Weighted Cross Entropy (DOWCE). These novel methods are specifically designed to address both class imbalance and the imbalance caused by the sensitive attribute of parcel size. The methods are categorized into three types: sample balancing, cost-sensitive, and hybrid methods.

3.4.1. Sample balancing methods

Sample balancing methods work by creating and/or removing samples to meet the desired class sample distribution (Chawla, 2010), which depends on the chosen balancing strategy. I select full balancing, which, since only oversampling is used, means that minority samples are oversampled to match the number in the majority classes.

Creating synthetic samples for time series data is challenging due to temporal dependencies. However, some synthetic oversampling methods for time series exist, such as T-SMOTE (Zhao et al., 2022) and OHIT (Zhu et al., 2022). Upon request, T-SMOTE's source code could not be shared given the researchers' company's strict open-source policy. OHIT's code was available on GitHub, but preliminary testing showed it was unsuitable, as OHIT, like most synthetic oversampling methods, struggles with noisy data (Zhu et al., 2022). Given the cloud and atmospheric noise present in the L1C top-of-atmosphere data that is used, oversampling with OHIT caused a decrease in performance relative to baseline. Therefore, I proceeded with two oversampling methods: Random Oversampling (RO) and the novel RO-R method.

3.4.1.1. Random Oversampling

RO is a straightforward method that balances samples by randomly replicating existing ones (Chawla, 2010), resulting in a dataset consisting of the original samples plus randomly selected duplicates.

3.4.1.2. Random Oversampling with Resampling

This novel method addresses both class and parcel size imbalances by balancing classes while prioritizing the oversampling of small parcels by resampling a randomly oversampled dataset. The procedure is as follows.

First, for a given training dataset, Kernel Density Estimation (KDE) is used to calculate the mode of parcel area within each class. This step helps establish a class-specific threshold to distinguish between small and large parcels.

The next step introduces the algorithm's only hyperparameter: the desired ratio of samples below the KDE mode. This ratio controls the balance of samples that the resampling aims to achieve. The underlying assumption is that samples below the mode correspond to small parcels, and increasing their proportion enhances their representation, potentially improving classification performance for small parcels. For the final experiments, ratios of 35% and 40% below the mode are selected. These values are based on the observation that the original percentage of samples below the mode range from 20% to 30% across classes and increasing beyond 40% was deemed unnecessary for a substantial increase in the representation of small parcel samples.

Based on the desired ratio parameter, the resampling step is performed. Since the input dataset is randomly oversampled, duplicates above the threshold are removed, and below-threshold samples are randomly oversampled until the ratio is met. If no duplicates remain above the threshold, the original samples are preserved, and instead, below-threshold samples are oversampled to achieve the target balance. To ensure that the smallest parcels are prioritized in this process, a weight—defined as the square of the distance from the KDE mode—is applied, favouring the selection of samples farther from the threshold.

Because the resampling logic is applied to majority classes without constraints on reaching the desired ratio, this creates a class imbalance. To address this, the final step randomly oversamples the dataset to restore class balance. The before and after of this method is illustrated in Figure 4.

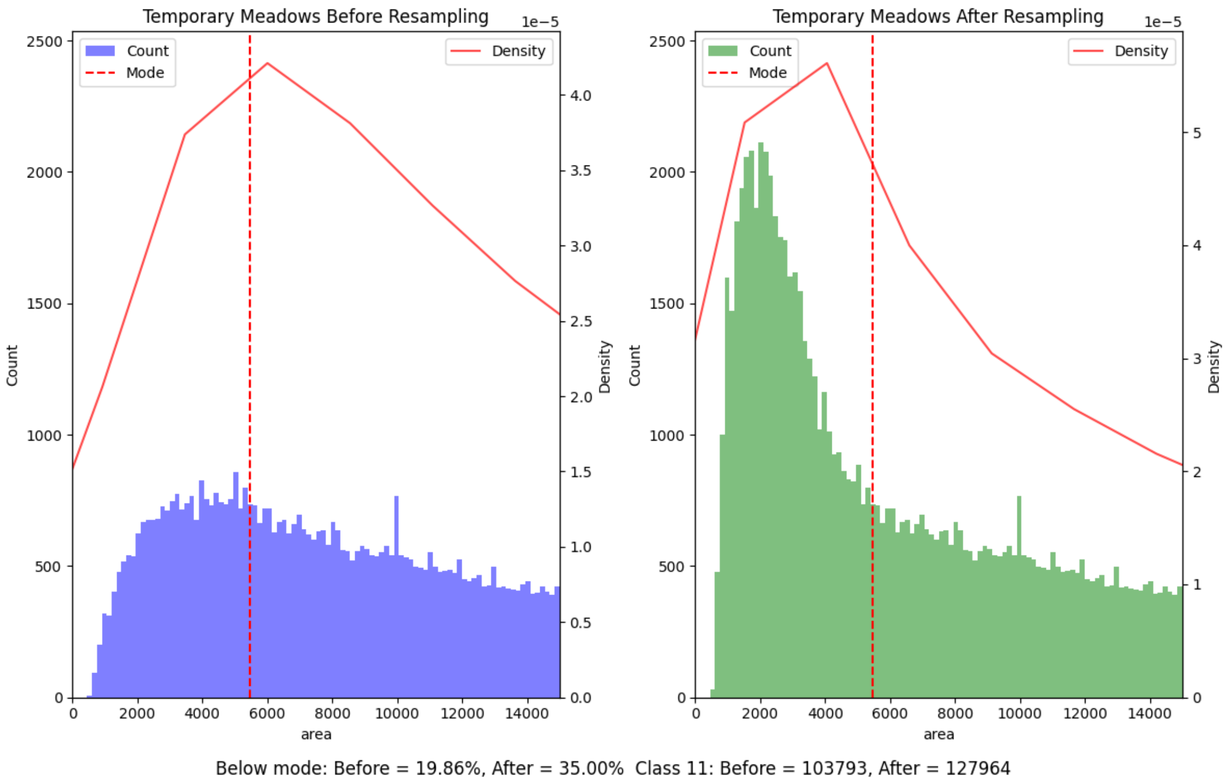

Figure 4 Visual representation of the RO-R method: the before and after parcel area distribution for dataset AD-5 temporary meadows class. Note, the before resampling dataset has had random oversampling applied to it.

It is important to note that while I selected two values for the desired ratio hyperparameter, k-fold cross-validation was not implemented due to time constraints. Instead, I trained the model separately for each parameter value and chose the one with the highest maximum GMean on the validation set. The model with the highest GMean on the validation set was then evaluated on the test set.

3.4.2. Cost-sensitive methods

Cost-sensitive methods address class imbalance by assigning different costs to misclassification errors, assigning a higher cost to misclassifying minority classes than majority classes (Liu & Zhou, 2006). I employed two widely used techniques—Weighted Cross Entropy (WCE) and Focal Loss (FL)—alongside a novel variant of WCE, DOWCE.

3.4.2.1. Weighted Cross Entropy

WCE is a very common loss function that assigns different weights to the classes, instead of the same weight for all classes as it is in a standard cross-entropy loss function. This way, the model can be trained to be less biased against minority classes (Rezaei-Dastjerdehei et al., 2020). WCE loss is a modified version of the standard cross-entropy loss:

$$L_{CE} = -\sum_{c=1}^{K} y_c \log(p_c)$$

Equation 2

Where y_c is the binary indicator if class label is correct for the given observation, p_c is the predicted probability for class c, and K is the number of classes. To account for class imbalance, the WCE loss function introduces weights w_c for each class c:

$$L_{WCE} = -\sum_{c=1}^{K} w_c \cdot y_c \log(p_c)$$

Equation 3

I choose to calculate w_c with the inverse frequency method (H. He & Garcia, 2009):

$$w_c = \frac{N}{K \cdot n_c}$$

Equation 4

Where N is the total number of samples, and n_c is the number of samples in class c.

3.4.2.2. Focal loss

FL (Equation 5) is a more recent loss function developed by Lin et al. (2017), designed to prioritize hard-to-classify samples by down-weighting easy-to-classify ones. This down-weighting is controlled by γ, which I tuned with k-fold cross-validation using values $\gamma = [0.5, 1, 1.5, 2, 2.5, 3]$. For α_c, which is analogous to w_c in WCE, I apply the same class weights as for WCE, using the inverse frequency method (Equation 4). The focus of FL on both hard-to-classify samples as well as class imbalance, theoretically makes it suitable to address both class imbalance and parcel area imbalance, making it an appropriate choice for this research.

$$L_{FL} = -\sum_{c=1}^{K} \alpha_c (1 - p_c)^{\gamma} y_c \log(p_c)$$

Equation 5

3.4.2.3. Double Objective Weighted Cross Entropy

DOWCE is a loss function that extends WCE by incorporating an additional objective to address the imbalance caused by parcel area. This is achieved by introducing a parcel-size-based weight variable $W(A)$, which modifies the loss calculation by adjusting the contribution of each sample based on its parcel area A. As shown in Equation 6, this weight directly multiplies the weighted cross-entropy loss of each sample, effectively scaling the loss depending on how large or small the parcel is.

$$L_{DOWCE} = -\sum_{c=1}^{K} W(A) \cdot w_c \cdot y_c \log(p_c) \qquad \text{Equation 6}$$

The calculation of $W(A)$ involves determining the mode (mode_c) of the parcel area distribution for each class using KDE, as well as identifying the 10th percentile of the area distribution (threshold_c). These two values guide the weighting on a per-class basis. Parcels with areas larger than or equal to the mode are assigned a weight of 1, while parcels smaller than the 10th percentile threshold are assigned the maximum weight W_{max} . The value of W_{max} is tuned through k-fold cross-validation using values $W_{\text{max}} = [1.25, 1.5, 2, 3, 4]$. For parcels with areas between the 10th percentile and the mode, the weight is linearly interpolated based on their size, such that the closer a parcel's area is to the 10th percentile, the higher the assigned weight. The weight function is thus defined as:

$$W(A) = \begin{cases} W_{\text{max}} & \text{if } A < \text{threshold}_c \\ W_{\text{max}} - \left(\frac{A - \text{threshold}_c}{\text{mode}_c - \text{threshold}_c}\right) \times (W_{\text{max}} - 1) & \text{if threshold}_c \leq A < \text{mode}_c \\ 1 & \text{if } A \geq \text{mode}_c \end{cases} \qquad \text{Equation 7}$$

The method is further illustrated in Figure 5, which shows how $W(A)$ varies depending on the parcel size. This parcel area-sensitive weighting, which favours small parcels, along with the class-based weighting, is designed to address both types of imbalances—class and attribute imbalance—hence the name Double Objective Weighted Cross Entropy.

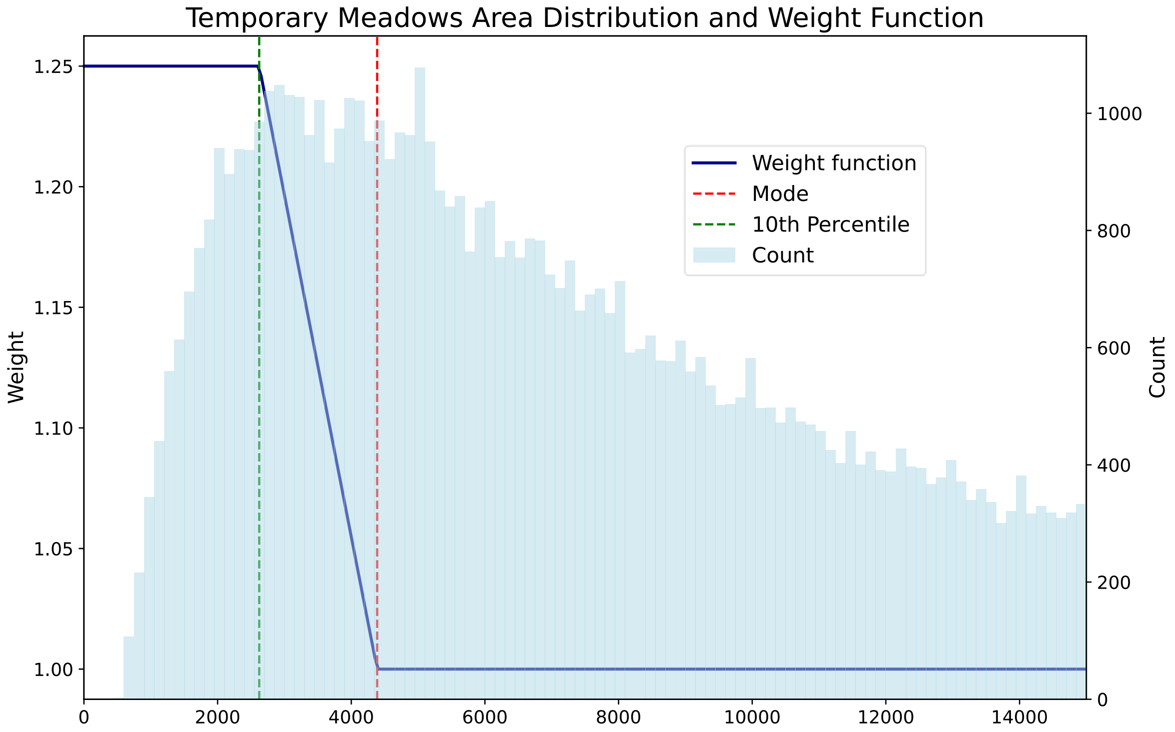

Figure 5 Visual representation of the DOWCE weight function for the temporary meadows class of AD-5

3.4.3. Hybrid methods

Hybrid methods combine sample balancing with cost-sensitive approaches, each applied at different stages of the process. First, sample balancing adjusts the dataset before training, and then cost-sensitive methods are applied during training. Thus, combining these methods involves feeding the balanced dataset into a training process that uses cost sensitive techniques. Two hybrid methods are used:

Random Oversampling with Double Objective Weighted Cross Entropy (RO-DOWCE): In this method, RO balances the class distribution by duplicating minority samples. DOWCE then addresses parcel size imbalance by applying higher penalties for misclassifying smaller parcels. While RO equalizes class sizes, making DOWCE technically single-objective, I continue to use the term "double objective" for consistency, as the implementation remains unchanged.

Random Oversampling with Focal Loss (RO-FL): Similar to the previous hybrid method, RO balances the dataset, addressing the class imbalance. Then, Focal Loss is leveraged to see whether it can help address the parcel-size imbalance by focusing on hard-to-classify samples more, which likely include small parcels.

3.5. Identifying the parcel area threshold

As required by the fairness metric (which is expanded on in section 3.6.1), it is necessary to establish a clear separation between small and large parcels. In the absence of a clear, societally relevant definition of "small parcels" for the study area, I opted for a data-driven approach to determine this threshold. Specifically, I employed a piecewise linear regression model, with parcel area as the independent variable and Macro F1 score as the dependent variable. Then, I define the threshold as the point where the two linear segments of the model intersect. Further, I explain the full procedure to arrive at the threshold.

To construct the piecewise regression, I use the testing samples of the fifth auxiliary dataset's (AD 5) results. Since this dataset forms the basis for undersampling the EDs, deriving the threshold from it ensures that the threshold is generalizable and consistent across all experiments. To ensure robustness and minimize the influence of outliers, the piecewise model is fitted across a range of cut-off values, from 100,000 to 30,000 square meters, decrementing by 1,000 at each step. These cut-off values represent the maximum parcel area included in each analysis. The piecewise linear regression was found to be sensitive to this parameter, so testing multiple cut-offs allows the model to find the optimal threshold by identifying the cut-off value that maximizes the R^2 of the fit. This approach ensures that the threshold is based on the best possible fit for distinguishing between small and large parcels. For each cut-off, the following steps are conducted:

1) **Filtering Data and Dividing into Bins:** The data is filtered to include only parcels below the cutoff value. The filtered data is then divided into evenly spaced bins according to parcel area, ensuring that parcels of similar sizes are grouped together. The number of bins is determined by the square root of the total number of samples.

2) **Calculating Macro F1 for Each Bin:** For each bin, the mean parcel area is calculated, along with the Macro F1 score. These form the independent and dependent variable respectively.

3) **Fitting the Piecewise Model:** A piecewise linear regression is fitted to the Macro F1 scores for that specific cutoff value. The model produces two linear segments, and the R^2 values of the segments are averaged to assess the goodness of fit for the cutoff value.

After testing all cut-off values, the final piecewise regression is chosen to be the one with the highest average R^2. The resulting threshold and piecewise linear regression plot follow in section 4.3.

3.6. Evaluation

In this section, I describe the methods used to evaluate the performance of the fairness optimization methods in mitigating both class imbalance and the imbalance between small and large parcels. The aim is to assess how effectively these methods mitigate the imbalances across datasets, examining trade-offs between performance metrics, and assessing whether the improvements in the performance due to the methods was statistical significance across datasets. Also, I investigate the effects of the dataset characteristics on the baseline model performance, to better understand how the model behaves without any fairness interventions. This section is structured as follows. First, I explain the evaluation metrics employed. Next, I outline the methods used to investigate the effects of dataset characteristics on baseline performance. Then, I outline the statistical evaluation method, which helps determine whether the increases in performance between the methods were statistically significant across datasets. Finally, I present the approach used to analyse trade-offs.

3.6.1. Metrics

Two metrics were chosen for evaluating the results: Macro F1 and the difference in Macro F1 between small and large parcels. These metrics reflect two key aspects of performance. Macro F1 measures performance with sensitivity to class imbalance, giving equal weight to all classes regardless of their sample size. It is important to note that the testing set is imbalanced due to protein crops and orchards, which were the only classes with fewer than 3,000 samples—the cap for the testing set in the testing region (FRH04). The other metric, Macro F1 Difference, measures how well the methods reduce the difference in performance between small and large parcels. Next, I elaborate on the two metrics.

Macro F1 quantifies how well the methods handle class imbalance by providing an unweighted average of the class-wise F1 scores. I use the Averaged F1 method (Equation 8), rather than the F1 of Averages, as it ensures a balanced evaluation across all classes (Opitz & Burst, 2021).

$$\text{Macro F1} = \frac{1}{n}\sum_{i=1}^{n} F1_i \qquad \text{Equation 8}$$

Macro F1 Difference serves as the fairness metric for the sensitive attribute of parcel size. It quantifies the difference between the Macro F1 scores of small and large parcels, as shown in Equation 9. Thus, negative values indicating that the classification favours large parcels and vice versa. To calculate group-specific Macro F1, the dataset is split into two groups, small and large parcels, for which the resulting threshold is described in section 4.3 . Then, confusion matrices are computed separately for each group. The Macro F1 score is then calculated independently for each group by averaging the F1 scores of all classes within that group.

$$\text{Macro F1 Difference} = \text{Macro F1}_{\text{small}} - \text{Macro F1}_{\text{large}} \qquad \text{Equation 9}$$

This custom fairness metric is chosen over more established fairness metrics due to several reasons. First, it provides consistency with the Macro F1 metric used to address class imbalance, ensuring that the evaluation uses a widely used metric for multiclass classification. While Macro F1 Difference does not directly compare error rates across groups like metrics such as equalized odds (Barocas et al., 2023), it still uses both false positives and false negatives. Importantly, it captures the disparity between small and large parcels in a manner that treats each class with equal weight, ensuring a balanced evaluation across the entire dataset. However, this approach presents a limitation. Fairness metrics like equalized odds are typically applied class-wise in multiclass classification (Sabato et al., 2024), allowing for a more detailed analysis of disparities within each class. However, due to the experimental setup, where the methods are tested across datasets with varying class distributions, a more detailed class-wise fairness analysis was not conducted, as the focus was on assessing how well the methods reduce disparities in performance across different datasets and classes.

3.6.2. Effects of dataset characteristics

The performance of the baseline model and its relationship to dataset characteristics were analysed to understand the impact of dataset characteristics on performance without fairness interventions. This analysis method was inspired by Waldner et al. (2019), who did it in the context of crop type mapping using SVM and RF classification models.

I present the effects of dataset characteristics using both evaluation metrics outlined in the previous section: Macro F1, and Macro F1 Difference. The characteristics examined are the number of classes, imbalance degree, complexity, the small-to-large parcel ratio, and the mean area-to-perimeter ratio of small parcels. Regression analyses and R^2 values were used to quantify how much variance in performance metrics can be explained by these characteristics. Additionally, R^2 values from multiple linear regressions were calculated to assess the combined influence of selected characteristics on performance metrics.

3.6.3. Statistical evaluation

In this study, multiple methods were tested across different datasets, allowing for the assessment of whether some methods consistently outperform others across datasets. To conduct this evaluation, the Friedman test (Friedman, 1940) followed by a post-hoc Nemenyi test (Nemenyi, 1963) is a recommended non parametric test v. In this test, statistical significance indicates whether the differences in average rankings across datasets are large enough to conclude that one method generally performs better than another. A non-parametric test was chosen because the conditions required for a parametric test were unlikely to be met. The tests are applied separately for each selected performance metric, with Macro F1 and Macro F1 Difference chosen to evaluate how well the methods addressed the two dimensions of fairness.

The evaluation procedure is as follows. First, for each dataset, the methods are ranked based on the selected performance metric. Then the average ranks of the methods across all datasets is calculated. Next, the Friedman test statistic is applied, checking if the differences in average ranks are significantly different from each other. Since the original Friedman statistic has been shown to be too conservative, the Iman-Davenport correction, which takes into account the number of methods and datasets, is applied (Iman & Davenport, 1980). The F-statistic in combination with $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom, where N is the number of datasets, and k is the number of methods, is used to find the critical value in the F distribution table, using a significance level of 0.05. If the F-statistic is greater than the critical value, it indicates that the difference in ranks is statistically significant, meaning that at least one method outperforms the others for the selected metric.

If the Friedman test shows that at least one method performs differently, the post-hoc Nemenyi test is applied to identify which specific methods differ from one another. It works by performing pairwise comparisons between the methods to determine which ones are significantly different. In the Nemenyi test, two methods are considered significantly different if the difference in their average ranks exceeds the critical distance (CD). This distance is determined by several factors: the number of methods k, the number of datasets N, and the critical value q_α which is based on the Studentized range statistic divided by $\sqrt{2}$ (Demšar, 2006), and depends on k and the significance level α, for which I choose $\alpha = 0.05$.

$$CD = q_\alpha \cdot \sqrt{\frac{k(k+1)}{6N}} \qquad \text{Equation 10}$$

The CD, calculated with Equation 10, provides a threshold for determining whether the difference in ranks between two methods is statistically significant. If the difference exceeds the CD, the methods are considered significantly different; if not, they are considered to perform similarly across datasets.

To visualize the results, CD plots are used. These plots show the methods on a horizontal axis based on their average ranks and connect methods that are not significantly different using a CD bar. If two methods are not connected by this bar, their performance is statistically different. To supplement these results, I also created box plots of the ranks across datasets to help understand how spread the ranks were for the methods.

3.6.4. Evaluation of trade-offs

To evaluate trade-offs, I examined the increases of performance metrics from fairness optimization methods relative to the baseline for each dataset, plotting these against each other. These were defined as the difference between the metric values achieved with a given method and the baseline performance:

$$\Delta Macro\ F1_m = Macro\ F1_m - Macro\ F1_{baseline} \qquad \text{Equation 11}$$

$$\Delta \text{Macro F1 Difference}_m = |\text{Macro\_F1\_Difference}_{\text{baseline}}| - |\text{Macro\_F1\_Difference}_m| \qquad \text{Equation 12}$$

Where m denotes a specific method, and $baseline$ refers to the performance score for the baseline model. Note that the gain for Macro F1 Difference is calculated as such because all Macro F1 Difference scores were in the negative. This calculation ensures that this gain value is consistent with the others, where positive values represent an improvement—specifically, a reduction in the disparity between the Macro F1 scores of small and large parcels.

The following pairs are plotted to evaluate the trade-offs: Macro F1 Difference against Macro F1, to assess whether reducing the disparity between small and large parcels affects overall balanced performance; and Macro F1 for large parcels against Macro F1 for small parcels, to explore whether enhancing performance for small parcels compromises the performance of large parcels.

4. RESULTS

This section outlines the results of the applied methods. First, the results from the dataset creation steps are outlined. This is proceeded with a section on the results of the sensitive attribute identification process. Next, I outline the results of the effects of dataset characteristics on the baseline performance. Finally, the main results of this research are presented, that is the evaluation of the fairness optimization methods and the effects of dataset characteristics on baseline performance.

4.1. Dataset creation

This section outlines the results of the dataset creation steps. First, the results of the dataset containing all samples inclusive of thin parcels are presented, identifying a group that can be removed to reduce the noise of the experiments. Next, the results of the dataset reduction test are given, showcasing that the selected caps were a suitable choice. Finally, the candidate dataset creation and experimental dataset selection step is outlined, presenting a table with all the EDs.

4.1.1. Pre-processing: Removal of thin parcels

The results demonstrated that parcels with an area-to-perimeter ratio between 0 and 6 had significantly lower classification performance compared to those with higher ratios. As shown in Figure 6, parcels in this ratio bin exhibited noticeably worse macro-averaged F1 scores, being ~2.4 times worse than the second worst performing bin (6-12).

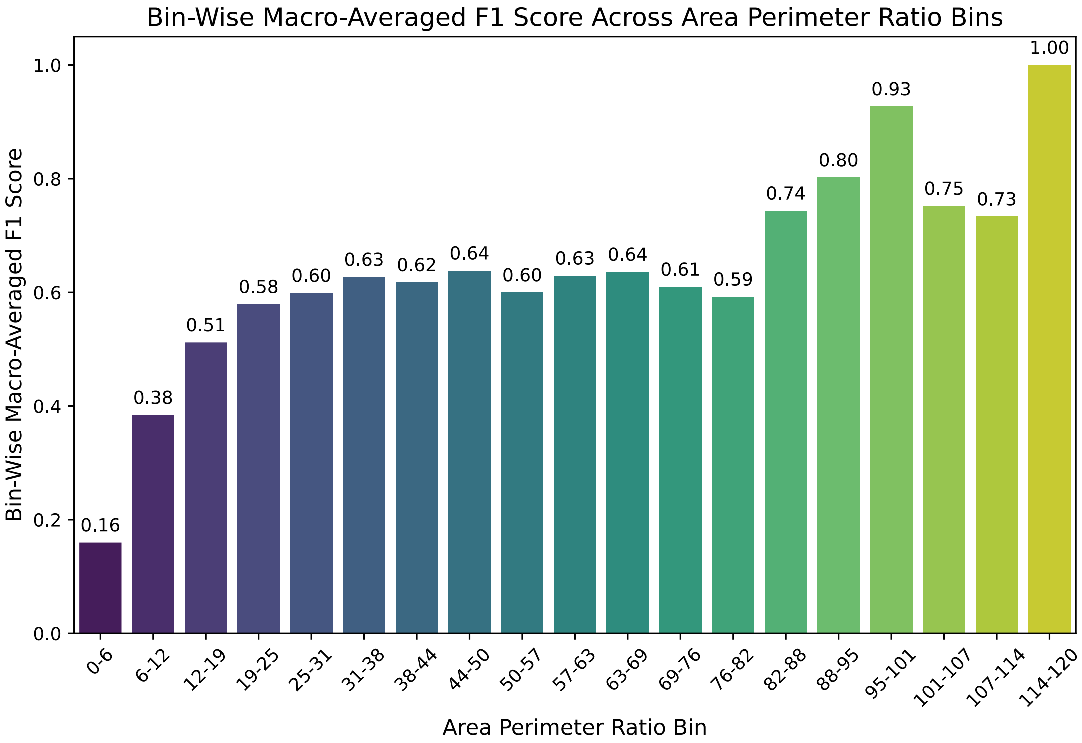

Figure 6 Macro F1 scores across area perimeter ratio bins.

Furthermore, I confirmed that the observed performance difference is largely driven by parcel thinness, rather than small size alone. To demonstrate this, I compared the performance of parcels with an area-to-perimeter ratio inside and outside the 0-6 range, focusing only on parcels within the same area range. This ensured a fair comparison, with both thin parcels (ratio ≤ 6) and those with a higher ratio included while excluding areas that only had one type of parcel (thin or not thin). The results show that parcels with an area-to-perimeter ratio below 6 (thin parcels) achieved a macro-averaged F1 score of 0.17, while parcels with a ratio above 6 achieved a significantly higher score of 0.60.

As a result, all parcels with an area-to-perimeter ratio between 0 and 6 were excluded from the dataset. This exclusion amounted to the removal of approximately 91,000 samples, representing about 12% of the total dataset. However, as shown in Table 1, this reduction was not uniformly distributed across classes. Some classes, such as "miscellaneous," were disproportionately affected, with approximately 74% of their samples being excluded, whereas other classes like barley and wheat experienced minimal reductions, with less than 1% of their samples removed.

Table 1 Class-wise samples before and after thin sample removal

Class	Acronym	Total Samples	Thin Samples	% Thin Samples	Total Samples Without Thin Samples
Barley	Ba	36,905	171	0.463	36,734
Wheat	Wh	89,555	348	0.389	89,207
Corn	Co	153,908	736	0.478	153,172
Fodder	Fo	23,023	399	1.733	22,624
Fallow	Fa	12,157	3,457	28.436	8,700
Miscellaneous	Mi	66,550	49,290	74.065	17,260
Orchards	Or	3,070	497	16.189	2,573
Other cereals	Ce	20,236	149	0.736	20,087
Permanent meadows	Pm	127,813	16,772	13.122	111,041
Protein crops	Pr	3,302	28	0.848	3,274
Rapeseed	Ra	14,732	56	0.380	14,676
Temporary meadows	Tm	182,212	17,522	9.616	164,690
Vegetables or flowers	Vf	30,334	1,833	6.043	28,501
Total		763,797	91,258	11.948	672,539

4.1.2. Dataset reduction

The dataset reduction experiment (AE 2) demonstrated that capping the training samples at 10,000 and the testing samples at 3,000 significantly reduced computational. However, the performance was significantly affected. As shown in Table 2, the full dataset (AD 2) achieved an overall accuracy (OA) of 0.718, while the 10k training cap dataset (AD 3) achieved an OA of 0.647. Despite this reduction in accuracy, the smaller dataset maintained reasonable performance and resulted in an 80% reduction in FLOPS per epoch, making the 10k cap a justifiable trade-off for efficiency.

Table 2 Performance for full and capped datasets

	(AD 2) Full training and testing datasets	(AD 3) 10k cap training, full testing dataset	(AD 4) 10k cap training and 3k cap testing
OA	0.718	0.647	0.659
GMean	0.347	0.578	0.596

Interestingly, the GMean improved from 0.347 for the full dataset (AD 2) to 0.578 for the 10k training cap dataset (AD 3). This improvement was attributed to the sample balancing effect of the 10k cap, which undersampled the majority classes, enhancing the performance of minority classes. The final dataset (AD 4), with a 10k cap on the training set and a 3k cap on the testing set, achieved similar performance to the dataset with the full test set (AD 3). The OA of AD 4 was 0.659, only marginally higher than AD 3, and the GMean increased slightly to 0.596. These minimal differences demonstrated that reducing the testing set to 3,000 samples per class did not significantly affect the evaluation metrics, making it a viable option for reducing computational costs during testing.

There was a notable finding for the per-class performances. For corn, reducing the number of samples from 153,172 in AD 2 to 10,000 in AD 3 resulted in a minor decrease in class recall, from 0.96 to 0.92, and a slight increase in class precision from 0.97 to 0.98. In contrast, for temporary meadows, which are more challenging to classify due to their spectral similarity to permanent meadows, the reduction in samples from 164,690 to 10,000 led to a significant drop in class recall, from 0.63 to 0.36, accompanied by an increase in precision, from 0.67 to 0.72.

Based on these findings, the 10k cap for the training set and the 3k cap for the testing set were selected for all subsequent experiments. The testing set remained consistent across all experiments, ensuring a stable evaluation process, while only including the classes that were present in the respective training datasets for each experiment. Note that for this final testing set, protein crops and orchards contained 651 and 456 samples respectively, as there were no more samples available for the FRH04 region after dropping all thin samples.

4.1.3. Candidate dataset creation and selection of final datasets

The candidate dataset generation produced 300 datasets with varying numbers of classes, complexities, and class distributions. From this pool, 10 diverse datasets were manually selected, ensuring a variety of complexity levels, different numbers of classes, and that most classes were majority at least once.

Details of the final datasets are provided in Table 3, which includes the imbalance degree, complexity, and number of samples. Note that it was not the goal to select datasets where the imbalance degree and number of samples were varied, these are provided to describe the datasets in more detail.

There was a strong positive Pearson's correlation between the number of classes and the complexity measure ($r = 0.87$), indicating that datasets with more classes tend to have higher complexity. There was also a noticeable overlap in the complexity range for datasets with fewer classes. For instance, datasets with 4 classes had complexity values ranging from 0.55 to 0.73, while datasets with 5 classes ranged from 0.61 to 0.76. Due to the limited number of samples available for orchards (which had only 2,117 samples for the training datasets), this class never became a majority in any of the final selected datasets.

Table 3 Description of the experimental datasets. Majority class acronyms are bolded, with acronyms explained in Table 1. The subscripts indicate the number of samples in each class, where (very low) **VL** = 500, (low) **L** = 1,000, (medium) **M** = 2,500, (high) **H** = 5,000, and (very high) **VH** = 10,000.

ID	# classes	Class allocation	# of samples	Complexity	Imbalance degree
ED 1	4	**Fa$_{H}$**-Mi$_{M}$-Pr$_{M}$-Ra$_{VL}$	10500	0.623	2.340
ED 2	5	Fo$_{M}$-Or$_{VL}$-Pr$_{VL}$-Tm$_{H}$-**Vf$_{VH}$**	18501	0.637	2.588
ED 3	6	**Fo$_{M}$**-Mi$_{VL}$-Or$_{L}$-**Pr$_{M}$**-Vf$_{L}$-Wh$_{L}$	8501	0.681	3.300
ED 4	7	**Co$_{VH}$**-Or$_{L}$-Pm$_{VL}$-Ra$_{L}$-Tm$_{VL}$-**Vf$_{VH}$**-Wh$_{M}$	25501	0.716	4.573
ED 5	8	Ba$_{VL}$-**Ce$_{VH}$**-Co$_{H}$-Or$_{VL}$-Pr$_{M}$-Ra$_{M}$-**Tm$_{VH}$**-Vf$_{H}$	36001	0.746	3.567
ED 6	9	Ba$_{VL}$-Fa$_{L}$-Mi$_{M}$-Or$_{L}$-**Pm$_{VH}$**-Pr$_{L}$-Ra$_{H}$-Vf$_{H}$-**Wh$_{VH}$**	36000	0.763	4.565
ED 7	10	Ce$_{M}$-**Co$_{VH}$**-Fa$_{L}$-Mi$_{VL}$-Or$_{VL}$-Pm$_{H}$-Pr$_{M}$-**Ra$_{VH}$**-Vf$_{M}$-**Wh$_{VH}$**	44502	0.805	5.534
ED 8	11	**Ba$_{VH}$**-Ce$_{L}$-Fa$_{VL}$-**Fo$_{VH}$**-Mi$_{H}$-Or$_{L}$-Pm$_{L}$-Pr$_{M}$-Ra$_{H}$-**Tm$_{VH}$**-Wh$_{M}$	48498	0.870	5.549
ED 9	12	Ba$_{M}$-**Ce$_{VH}$**-Co$_{L}$-Fo$_{L}$-Mi$_{VL}$-Or$_{L}$-**Pm$_{VH}$**-Pr$_{VL}$-**Ra$_{VH}$**-Tm$_{L}$-**Vf$_{VH}$**-Wh$_{VL}$	47998	0.844	7.600
ED 10	13	**Ba$_{VH}$**-Ce$_{H}$-**Co$_{VH}$**-Fa$_{H}$-Fo$_{L}$-**Mi$_{VH}$**-Or$_{L}$-**Pm$_{VH}$**-Pr$_{M}$-**Ra$_{VH}$**-Tm$_{L}$-Vf$_{VL}$-Wh$_{VL}$	66501	0.855	7.529

4.2. Sensitive attribute identification

The sensitivity analysis identified parcel area as the most impactful attribute influencing classification performance. By dividing the parcel area values into 20 quantile-based groups, I calculated the macro F1 score for each group. As shown in Figure 7, the results demonstrate that parcel area exhibits the highest sensitivity to performance. The groups are arranged in ascending order, with group 1 containing the smallest parcel areas and group 20 the largest. This analysis reveals that smaller parcels consistently exhibit significantly lower classification accuracy compared to larger parcels. In contrast, other candidate attributes showed relatively low sensitivity to performance.

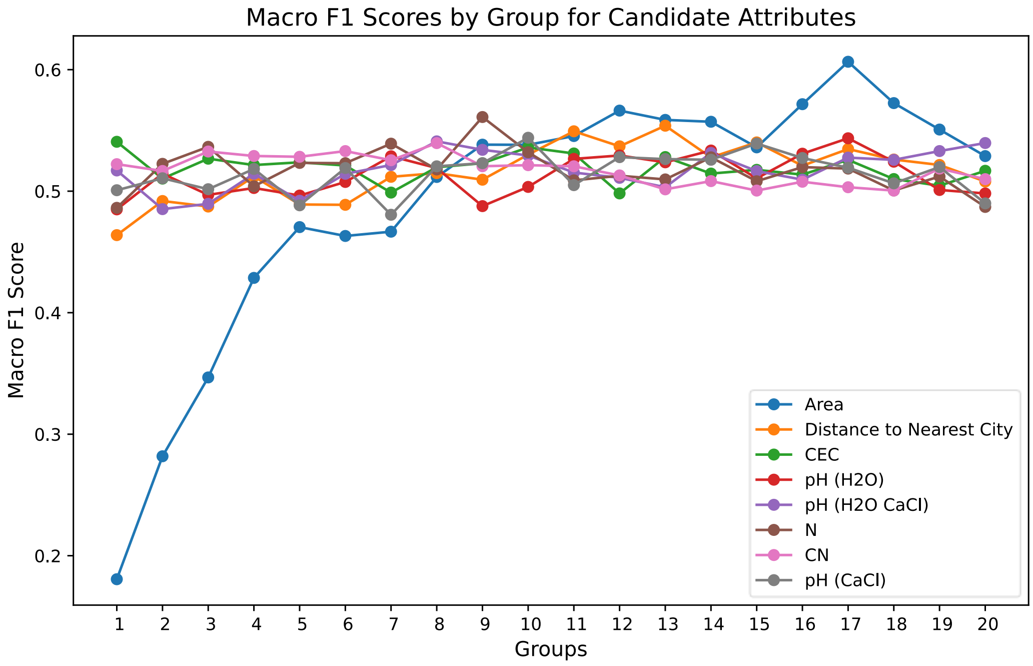

Figure 7 Macro F1 Scores across groups for the candidate sensitive attributes. Details of the chemical attributes are explained by Ballabio et al. (2019)

4.3. Identifying parcel area threshold

The piecewise linear regression model was fitted to the Macro F1 score across a range of cut-off values, from 100,000 to 30,000 square meters, decrementing by 1,000 square meters at each step. For each cut-off value, the piecewise model generated two linear segments, and the R² values were calculated to assess the goodness of the fits. After evaluating the fit for each cut-off, the model with the highest average R² had produced a threshold of **6,170 square meters**. This final model can be seen in Figure 8. The resulting threshold is used to split the dataset in two to apply the metric of Macro F1 Difference, which measures how well the methods reduce the performance difference between small and large parcels.

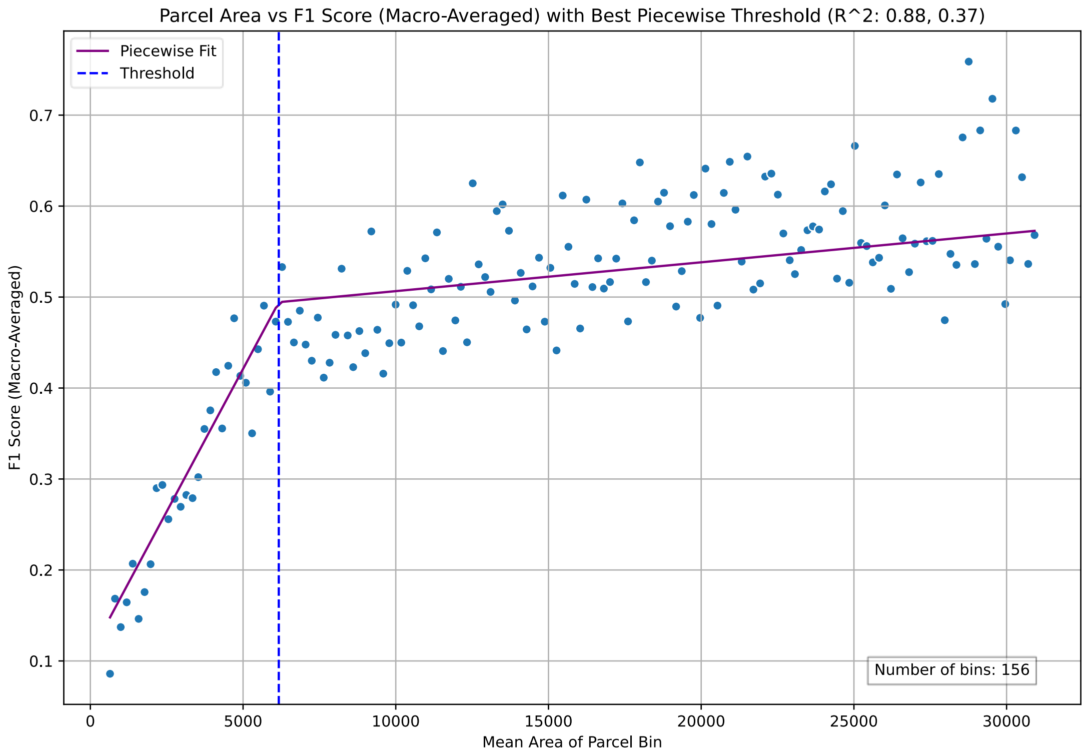

Figure 8 Piecewise linear regression model with the highest average R² value, identifying a threshold of 6,170 square meters for splitting the dataset in small and large parcels.

4.4. Effects of dataset characteristics

The results of the effects of dataset characteristics on baseline performance showed that Macro F1 and Macro F1 Difference ranged from 0.41 to 0.79 and -0.23 to -0.10, respectively (Figure 10). Macro F1 had a strong negative Pearson's correlation with the number of classes ($r = -0.75$). In contrast, Macro F1 Difference had a negligible correlation ($r = 0.09$). The correlation between Macro F1 Difference and Macro F1 was negligible, indicating that class balanced performance is not representative of how well the model handles the performance disparity between large and small parcels.

The relationships between performance metrics and dataset characteristics are shown in Figure 9. The variance of Macro F1 was significantly influenced by the imbalance degree and complexity, which together explained 55% of the variance in Macro F1. In contrast, these characteristics did not explain the variance in Macro F1 Difference. However, the small-to-large area ratio accounted for 12% of the variance in Macro F1 Difference, primarily driven by ED 1, which had an unusually high ratio of 0.90, while the rest of the dataset's ratio ranged from 0.18 to 0.38. Excluding ED 1 reduced the R² to 0.07. Conversely, the mean area-to-perimeter for small parcels was able to explain 18% of the variance. Together, the mean area-to-perimeter ratio for small parcels and small-to-large parcels ratio explained 19% of the variance in Macro F1 Difference.

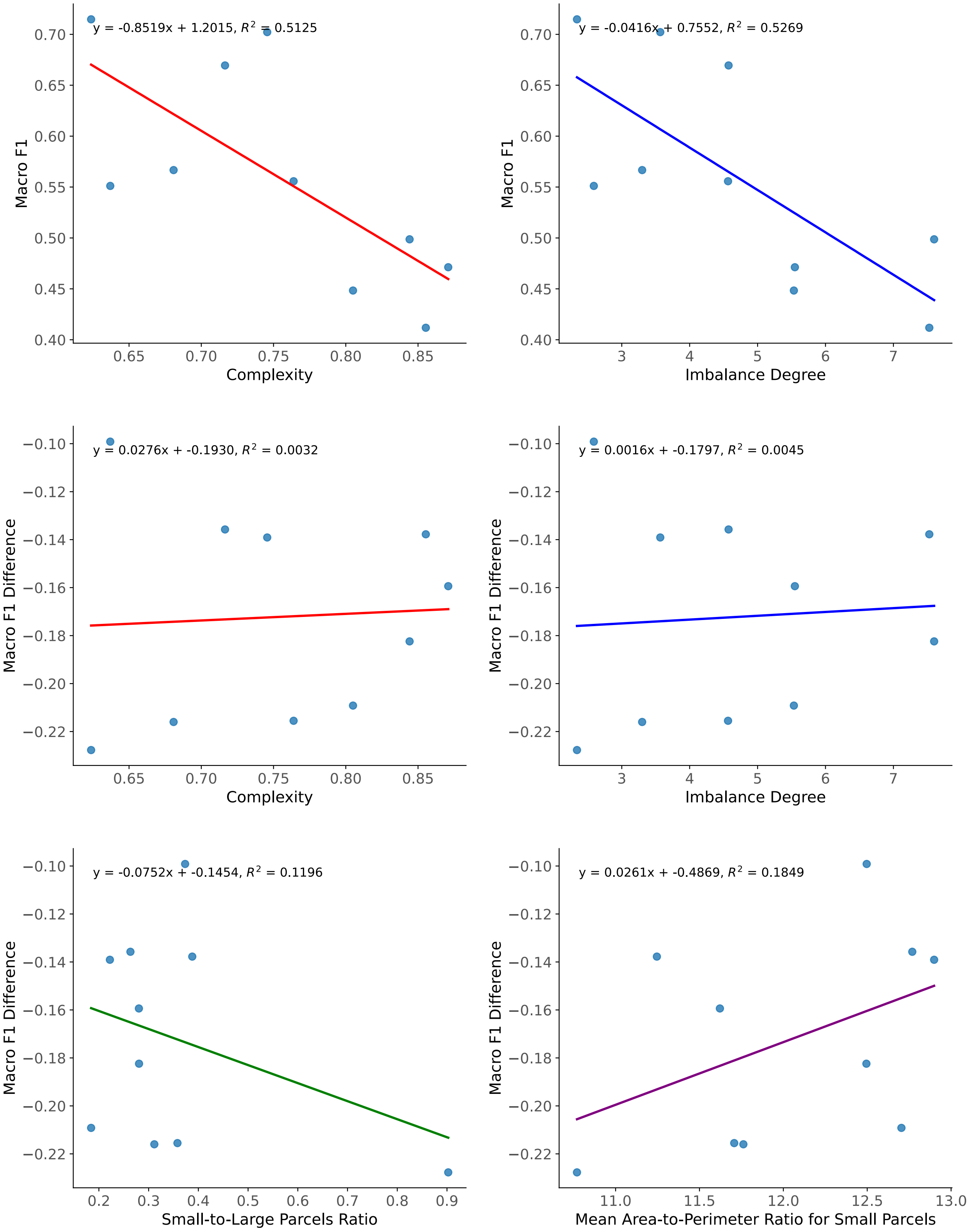

Figure 9 Relationship between performance and dataset characteristics for the baseline model results.

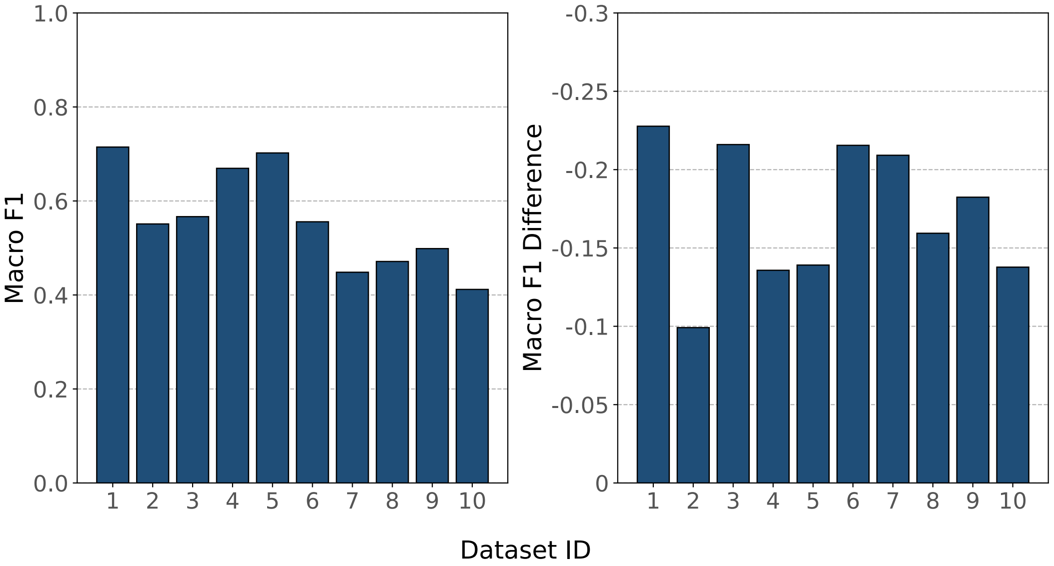

Figure 10 Baseline performance results across datasets.

4.5. Evaluation results of the fairness optimization methods

This section outlines the main findings of how the fairness optimization methods performed across the EDs in addressing both class imbalance and the performance disparity between small and large parcels.

The RO-DOWCE method achieved the highest improvement in **Macro F1**, with an increase of 0.17 for ED 10. It also produced the highest average increase in Macro F1 across all datasets (Figure 11a). Interestingly, out of the 70 dataset-method combinations, only 5 showed a decrease in Macro F1. Notably, 4 of these 5 decreases were from ED 2 and had a maximum Macro F1 increase of just 0.02 with the RO-R method.

Significant differences in performance across methods were observed for the Macro F1 metric (Friedman test; $p < 0.001$). As seen in the critical distance diagrams from the Nemenyi test (Figure 13a), RO-DOWCE was the highest-ranking method, outperforming FL, WCE, DOWCE, and the baseline model significantly, although it was not significantly different from RO-R, RO, and RO-FL. Notably, purely cost-sensitive methods ranked lower overall compared to sample balancing and hybrid methods. The rank distribution box plot in Figure 12a shows that FL and RO had the widest spread in ranks across datasets, ranging from 2 to 8 and 1 to 7 respectively when excluding outliers. In contrast, RO-DOWCE not only ranked 1st most frequently but also had a median rank of 1, while never dropping below the third rank, demonstrating its reliable performance across different datasets.

The highest gain in **Macro F1 Difference** was achieved by the DOWCE method for ED 7. This was an increase from -0.21 to -0.18, indicating that even the largest reduction in performance disparity was modest. Interestingly, for ED 10, all methods resulted in a decrease of Macro F1 Difference. The smallest decrease for this dataset was -0.04, which, in absolute terms, is larger than any increase observed across all other datasets. As seen in Figure 11a, DOWCE had the highest average increase in Macro F1 Difference, but it also had the lowest average increase in Macro F1, which was still substantially positive. Remarkably, as shown in Figure

11b, there were no trade-offs between Macro F1 gains for small and large parcels. RO-DOWCE most reliably improved performance for both parcel types across datasets, but was less effective at reducing performance disparities, as its average gain in Macro F1 Difference was close to zero. Notably, ED 2 remained an outlier, showing negative Macro F1 gains for both parcel sizes across all methods except RO-R. Overall, while most methods improved performance for both small and large parcels, their ability to reduce performance disparities between these groups was limited.

For Macro F1 Difference, significant differences between methods were found (Friedman test; p = 0.03), though these were much less pronounced than for Macro F1. For the Nemenyi test results (Figure 13b), DOWCE achieved the highest rank but was only significantly better than the baseline and RO, while no significant differences were found among the remaining methods. Notably, RO had the lowest average rank, followed by the baseline and WCE, indicating that methods specifically targeting performance disparities between small and large parcels may be more effective. The rank distribution (Figure 12b) shows that RO-R had a wide range, from first to last, making it the only method with such variability when excluding outliers determined by the interquartile range rule.

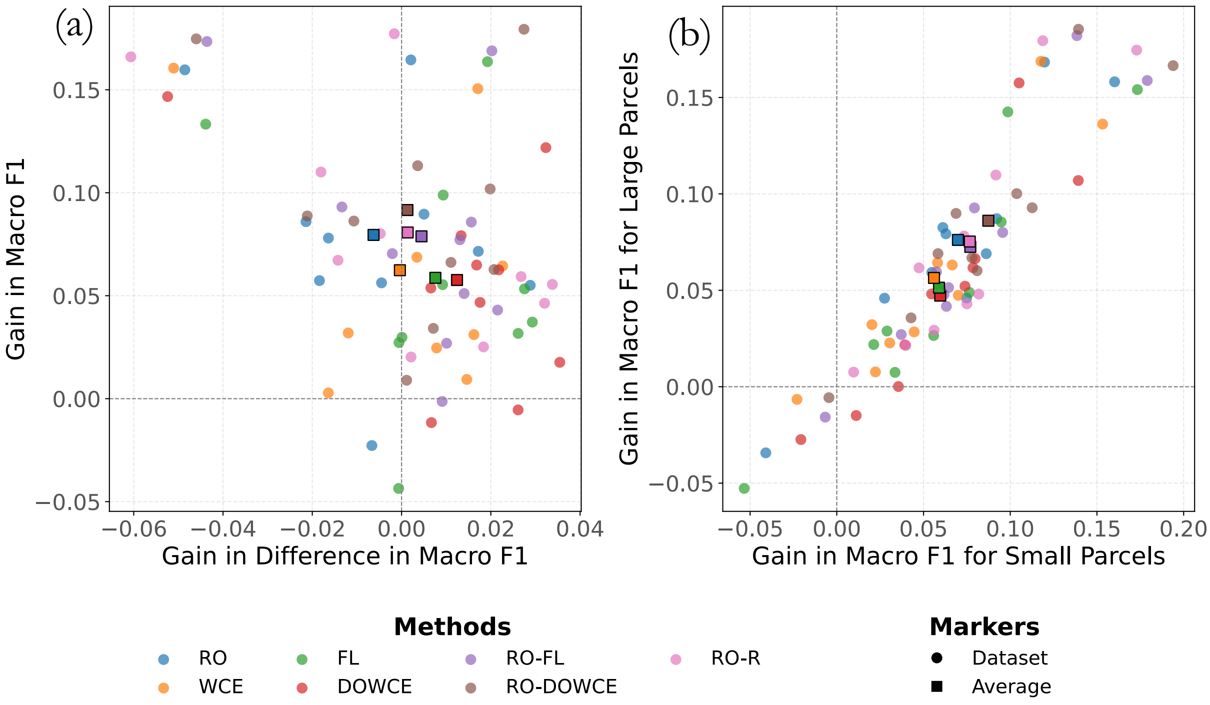

Figure 11 Trade-offs in performance gains across metrics. (a) Gain in Macro F1 versus gain in Macro F1 Difference; (b) Gain in Macro F1 for large parcels versus gain in Macro F1 for small parcels. Each plot shows the results for all datasets as well as the average gain per method across datasets.

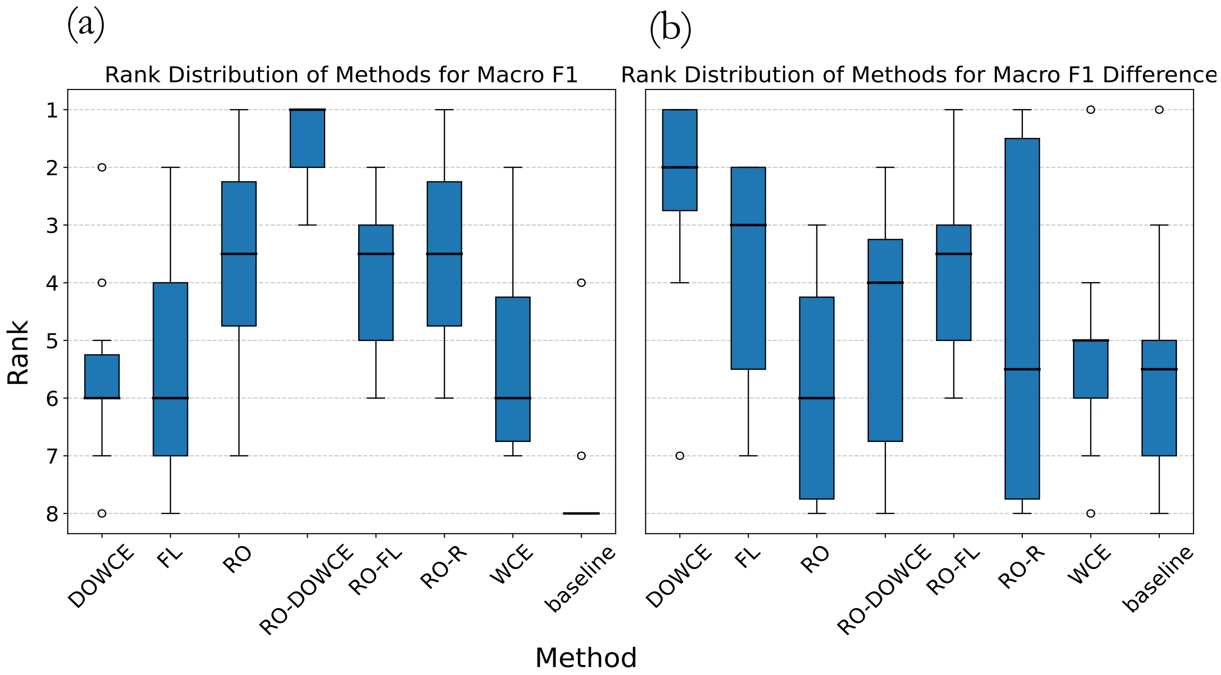

Figure 12 Box plots showing the rank distribution of methods for (a) Macro F1 and (b) Macro F1 Difference. Outliers are identified using the interquartile range rule.

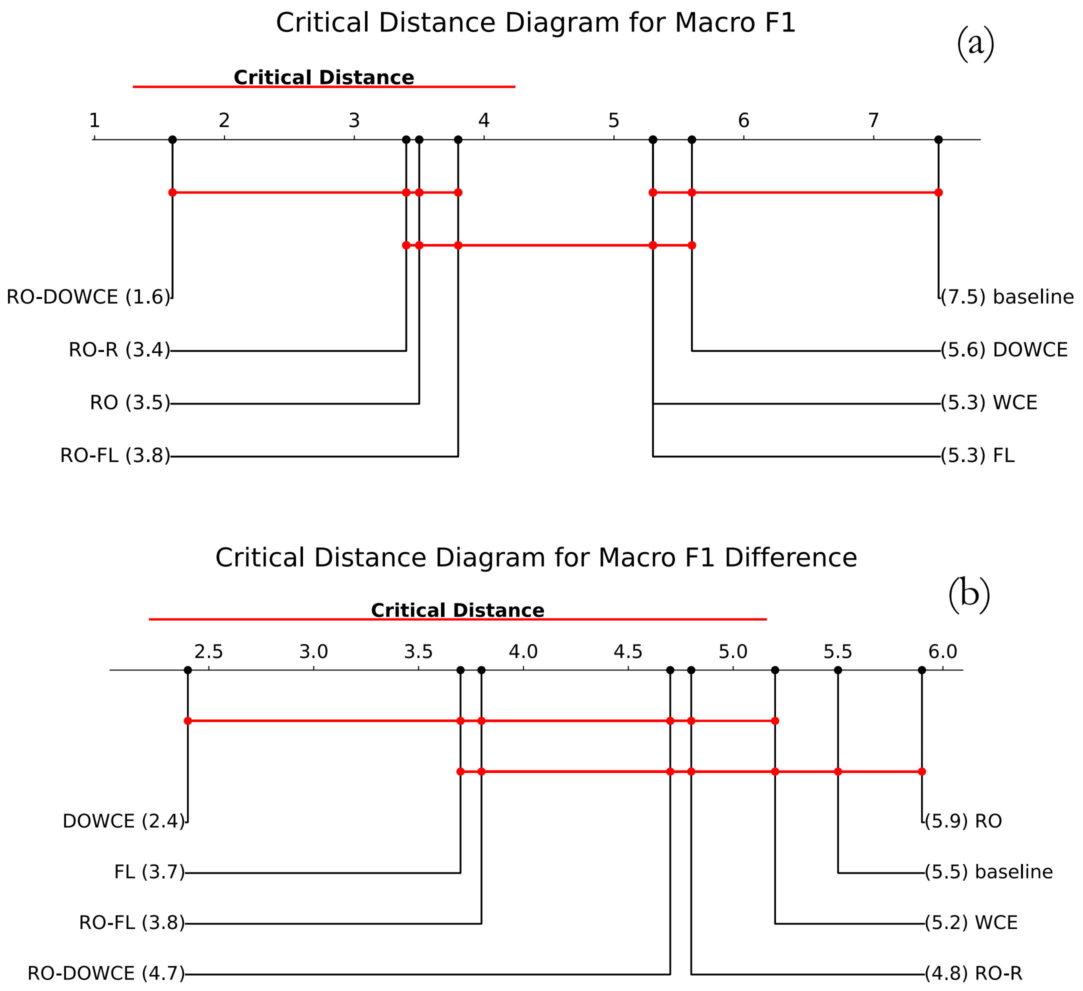

Figure 13 Nemenyi critical distance diagrams comparing balancing methods and including the baseline, for (a) Macro F1 and (b) Macro F1 Difference. Methods connected by red horizontal lines are not significantly $(\alpha = 0.05)$ different from each other. The x-axis represents the average rank values.

5. DISCUSSION

The central goal of this study was to improve the fairness of crop classification by addressing both class and sensitive attribute imbalances using a transformer classifier. To achieve this, several research questions were answered. Datasets were created with varying numbers of classes, distributions, and complexity, using subsampled datasets from the larger *BreizhCrops* dataset to ensure a comprehensive evaluation (RQ 2). Parcel size was selected as the key sensitive attribute due to significant performance disparities between small and large parcels (RQ 3). RO was chosen as the sample balancing method, and RO-R was developed to specifically address parcel size imbalance by increasing the proportion of small parcel samples (RQ 4-1 & RQ 4-2). WCE and FL were selected as cost-sensitive methods, while the novel DOWCE was designed to target parcel size imbalance by increasing the loss for smaller parcels (RQ 5-1 & RQ 5-2). Hybrid methods, RO-DOWCE and RO-FL, were introduced to address both class and sensitive attribute imbalances (RQ 6-1 & RQ 6-2). The diverse datasets allowed for the generalizability of the methods to be statistically tested, ensuring that the findings could be applied to different crop type mapping scenarios.

Answering RQ 7 provided insights into the effects of dataset characteristics on baseline performance aligned with expectations, showing a decrease in Macro F1 scores as complexity and imbalance degree increased, consistent with findings by Buda et al. (2018) and Waldner et al. (2019). However, these characteristics did not explain variations in Macro F1 Difference. While the ratio of small-to-large parcels partially explained the performance disparity, its explanatory power was limited. In contrast, the mean area-to-perimeter ratio of small parcels explained a substantially larger portion of the performance disparity. Combining these characteristics added only marginal explanatory power. This indicates that the thinness of the parcels could be one of the main drivers of performance disparity between small and large parcels, although a significant portion of the disparity remains unexplained. One hypothesis is that the higher edge pixel proportion in smaller parcels, for which the area-to-perimeter ratio serves as a proxy of, contributes to this disparity. This aligns with challenges related to the mixed pixel problem (Jian-jun, 2006). To test this hypothesis, future studies could isolate the effects of edge pixels by excluding them in the mean aggregation and observing the performance disparity thereafter. If the performance disparity persists after exclusion, it would suggest that other characteristics of small parcels other than mixed pixels contribute to decreased performance.

The key results of this research answered RQ 1: "How do the fairness optimization methods compare to each other and the baseline results across datasets with varying characteristics?". Contrary to the expectation of a trade-off—where improving performance for the sensitive group (small parcels) would negatively impact the privileged group (large parcels), as observed in class-based imbalances (Waldner et al., 2019), it was found that improvements in performance for small parcels did not lead to reductions in performance for large parcels. Instead, the performance increase was similar for both groups. The minimal reduction in the performance disparity between small and large parcels, despite the focus on small parcels, raises questions about the factors limiting further improvement. One possibility is that the higher prevalence of edge pixels in small parcels introduces additional noise, constraining performance gains. While the transformer model has demonstrated robustness to top-of-atmosphere noise in this dataset (Rußwurm et al., 2020), it is unclear how well it handles the noise introduced by edge pixels in mean-aggregated parcels. Although both large and small parcel accuracy improved with most methods, this could largely be attributed to class balancing. However, the fact that RO-DOWCE outperformed other methods could suggest that it handled noise better. By focusing on noisy small parcels, the model could have improved its ability to manage edge pixel noise, helping it to better learn the core features that distinguish crop types. This noise handling boosted performance across all parcel sizes by allowing the model to better focus on key class features. However, despite this, the model still faced challenges in fully overcoming the noise, as the improvement in small parcels remained limited. Further studies could demonstrate whether the improved performance of RO-DOWCE is due to better noise management by explicitly addressing noise—such as by removing edge pixels from the mean aggregation. This research could also help determine whether performance disparities are primarily driven by edge pixel noise or other characteristics of small parcels. One possible way to explore this is by analysing the predicted probabilities of the samples throughout the training process. If small parcels consistently show low predicted probabilities for the true class, and this pattern remains even after removing edge pixels, it would suggest that factors beyond noise are contributing to the classification difficulty. This approach would also allow for a class-wise analysis, revealing whether this difficulty pattern is consistent across different crop types or more pronounced in specific classes.

Notably, RO-DOWCE achieved the highest average rank across datasets for the Macro F1 measure, although the difference was not statistically significant when compared to RO-R, RO, and RO-FL. However, RO-DOWCE never dropped below third rank, mostly placing first or second, which exceeded expectations as it was initially thought that the method might lead to reduced performance for large parcels due to its focus on small parcel fairness. Additionally, DOWCE alone performed only marginally better than the baseline, ranking second lowest. These results indicate that assigning higher penalties for misclassifications of challenging groups may benefit overall performance when combined with class balancing methods but not when using cost-sensitive methods alone. Furthermore, the purely class balancing method that targeted small parcel imbalance, RO-R, scored the second highest average rank for Macro F1, indicating that increasing representation of small parcels in the training data does not significantly decrease performance but can even increase it in some datasets when using a transformer classifier. However, given the lack of statistical significance, it cannot be concluded that RO-DOWCE is superior to common methods such as RO-FL and RO. Given this lack of significant differences, practitioners might consider using RO-FL, as it emphasizes harder-to-classify samples without the need to identify specific groups, potentially contributing to fairness in a more universal way. Additionally, practitioners may want to test various methods, including synthetic oversampling methods. However, in doing so, they should be careful of the challenges associated with synthetic oversampling for time-series data, such as the need for non-noisy data to produce high-quality synthetic samples (Zhu et al., 2022). Instead, practitioners may opt to use RO, which avoids this caveat as well as being easier to understand than synthetic oversampling methods. As Buda et al. (2018) point out, RO does not cause overfitting in CNNs as it does in classical ML models. Furthermore, Waldner et al. (2019) demonstrated that RO achieved the highest average rank across crop type mapping datasets for the SVM model. These findings, along with the results from this study, suggest that RO is a robust and practical solution for addressing class imbalance in time-series classification tasks for DL and SVM models.

It is vital to remind of a major limitation of this study that impacts the interpretation of the results. Ideally, hyperparameter tuning would have been performed for each model-dataset combination; however, due to limitations in computational resources, this was not feasible. As a result, it remains unclear how much additional performance could be squeezed from some methods compared to others. It is entirely possible that some methods might benefit more from fine-tuning than others, impacting the average ranks of the methods. It is also worth noting that the DOWCE loss function's parcel-size-based weight variable calculation depends on a fixed threshold based on the 10th percentile of parcel area, and a linear interpolation between the threshold and the mode. While W_{max} was tuned, the interpolation and threshold were not optimized. Given that RO-DOWCE shows promise with its first-place average rank for Macro F1, tuning these additional hyperparameters could further distinguish this method from the rest.

Given the lack of conclusive evidence on whether the novel methods targeting sensitive attribute imbalance effectively decrease the disparity between sensitive and privileged groups, it might be beneficial to carry out a similar study using a sensitive attribute that is not associated with increased noise. However, identifying such an attribute for crop type mapping has proven challenging in the context of Brittany, France. Other regions, especially in the developing world, might exhibit more noticeable sensitivities, particularly because smallholder farmers in developing countries often face significant resource constraints (Meemken & Bellemare, 2020).

However, many smallholder farmers in the developing world do not have homogeneous parcels due to intercropping practices, making classification a challenging task (Jin et al., 2017; Samberg et al., 2016; Sinha et al., 2022). Additionally, smallholder farms are usually too small to be captured by coarse-resolution satellite imagery as, for example, in Africa, 25% of fields are less than 0.5 acres (Burke & Lobell, 2017). These factors necessitate a more targeted approach to address sensitive attribute fairness in the developing world, requiring finer-resolution imagery and considerations for intercropping (Estes et al., 2022).

While crop type mapping initiatives often focus on the developing world—for example, *GEOGLAM* (Becker-Reshef et al., 2023)—there are highly relevant applications in the developed world as well. To connect my research to a real-world case of crop type mapping relevant to my study area, the next section will present a case study on compliance checks in the EU. This will elaborate on the stakeholder and wickedness dimension of this research and further explore my findings by basing recommendations on them.

5.1. A wicked problem case study: compliance checks in the EU

In this section, I evaluate whether this research addresses the 'wicked problem' of unfairness in crop type mapping by focusing on the conflict between stakeholders in the case study of compliance checks for Common Agricultural Policy (CAP) subsidies. Given the lack of theoretical consensus on the term 'wicked problem,' it is important to clearly define its use. Following Termeer et al. (2019), I apply the concept analytically by reducing it to two key dimensions: conflict and uncertainty. The effectiveness of my research in addressing the 'wicked problem' thus depends on how well it tackles these dimensions. I further define conflict as referring to the disagreements between stakeholders and the lack of consensus on solutions, while uncertainty relates to the knowledge gaps—where improving understanding of the uncertainty may sometimes help reduce conflict.

Since my technical research primarily addresses the uncertainty dimension, I focus here on the stakeholder conflict dimension. I select the case study of CAP compliance monitoring in the EU, which aligns well with the location of the crop type mapping dataset used for the research: Brittany, France. To explore this conflict, I conduct a stakeholder analysis. First, I outline the project of CAP compliance monitoring through remote sensing. Next, I perform a stakeholder inventory, grouping them based on shared interests. Finally, I analyse the conflicts between these stakeholders and give recommendations, some of which consider my experimental results.

5.1.1. CAP compliance monitoring system

The EU provides subsidies to farmers through the CAP. The current compliance monitoring recommendation is to randomly visit 5% of declared plots. Given the potential to replace this system with remote sensing monitoring, the EU has published a technical guidance document on this transition (Devos et al., 2018). This shift is an active research area, for example, Lozano-Tello et al. (2021) suggests that compliance models incorporate several years of data to improve accuracy.

This proposed system would classify crop parcels and check if they match the crop type declared by the farmer. Such system will be subject to two types of errors: false positives (FP), where compliant parcels are flagged as non-compliant, and false negatives (FN), where non-compliant parcels get flagged as compliant. These errors are more likely for minority classes and sensitive groups unless effective bias mitigation is applied.

Devos et al. (2018) proposes that flagged non-compliance could be verified through field inspections or farmer-provided photos. Since it is a not known how countries are implementing this system, I imagine the best- and worst-case scenarios of each type of error. In the best-case scenario, a FP results in extra public expenditure and inconvenience for farmers, who may be repeatedly asked to verify compliant parcels due to biased classification results. In the worst-case scenario, repeated FPs could financially burden farmers by delaying subsidy payments, leading to legal disputes, while eroding trust in the CAP and the EU.

As for FNs, in the best-case scenario, minor non-compliance may be overlooked, but in some cases, e.g. parcels belonging to multi-year eco-schemes, future classifications may catch violations, allowing retroactive fines. In the worst case, significant misuse of public funds could occur due to major subsidy misallocations, further diminishing trust in the system and creating perceptions of unfairness among compliant farmers, as well as taxpayers, undermining the credibility of EU agricultural policies.

5.1.2. Stakeholders

I identify key **stakeholders** and group them based on their shared interests, which I speculate on. Note that I am purposefully choosing to focus on three main stakeholder groups to streamline the conflict analysis. Other stakeholders, such as EU taxpayers or Environmental groups, are excluded to keep the stakeholder conflict analysis more succinct.

- **Under-represented farmers**: These are farmers whose parcels have relatively low representation in the data, either because they grow niche minority crops, or have other attributes that make their parcels harder to learn. This makes this group more vulnerable to both types of errors. This group is also one that the CAP aims to make more prevalent given the emphasis on the transition to more sustainable agriculture, through focusing on biodiversity and soil health incentivized through eco-schemes (European Commission, 2022).

  - **Interests:** Prioritize minimizing FPs (being incorrectly flagged) for them, to avoid any annoyances or subsidy delay. However, this is not a monolith group. Some might prioritize reducing FPs for their specific niche (e.g. small, ecological farm), while some might focus on overall fairness while also disagreeing over what specific groups to prioritize. This can lead to internal tensions about how fairness should be prioritized and whether minimizing FPs for certain crops comes at the expense of others. This group might also be interested in minimizing FNs for all groups, as they would want a fair allocation of funds. However, their primary concern is likely ensuring they are not unfairly penalized, and as such might be willing to tolerate a higher rate of FNs.

- **Well-represented farmers:** These are farmers who grow common crops and do not have under-represented attributes. These farmers are less likely to experience both errors.-They are also more likely to have larger political power due to lobbying of **farmer interest groups**, another key stakeholder. These groups represent the rights of all farmers but often prioritize large-scale profitable farmers. For example, Copa-Cogeca, the strongest interest group for European farmers, has faced criticism due to prioritizing high tech large-scale farming at the expense of moving towards more sustainable farming (Savage & Lei Win, 2023).
  - **Interests:** Prioritize minimizing FPs but might be more tolerant than under-represented groups due to larger resources to manage the administrative burden of being incorrectly flagged. Might be willing to tolerate higher FNs in return of lower FPs.
- **EU member states:** Are the political bodies responsible for implementing the CAP, and ensuring compliance of farmers to ensure the EU CAP subsidies are fairly allocated. This group manages communication with farmers and resolves disputes, such as flagged noncompliance, and would be responsible for implementing the compliance monitoring system through remote sensing with the help of **technology providers** such as public scientific institutions and private businesses. They follow the guidance and regulations set by the directorate-general for Agriculture and Rural Development of the European Commission (**DG AGRI**), another stakeholder. I assume conflicts between DG AGRI and EU member states are unlikely for this topic, as they both have a vested interest in ensuring fair subsidy allocation.
  - **Interests:** Likely to prioritize minimizing both FPs and FNs. FPs as they might lead to administrative burden and loss of trust in the system. FNs as it is critical to avoid subsidy misuse and to enforce the law, making it likely that they are more of a priority than FPs.

5.1.3. Conflicts

Given the identified stakeholder groups, I will now elaborate on the potential conflicts that may arise between them, which are summarized in Table 4. One key conflict may occur within the under-represented farmers group, which includes diverse sub-groups such as small-scale, eco-friendly or minor crop producers. Although they share the overarching goal of minimizing FPs—to avoid being unfairly flagged for non-compliance—this creates tension regarding which sub-groups should be prioritized in fairness interventions. In turn, this could fragment the group, making advocacy from within the group harder, as a unified demand for fairness intervention becomes more challenging to formulate. Due to this, member states aiming for a fair system will struggle to define the fairness objective and will have to find compromises by neglecting some groups in favour of others.

Table 4 Stakeholder conflict matrix. The first conflict represents a conflict within the stakeholder group.

Stakeholder/s		Conflict description
Under-represented farmers		Disagreements over what FPs to prioritize minimizing given the vastness of different under-represented farmer groups. This creates tension over which specific groups should be prioritized.
Under-represented farmers	Well-represented farmers	Under-represented farmers fear that fairness interventions will be undermined due to well-represented farmers' influence, while well-represented farmers seek to protect their performance and resist changes that may cause significant trade-offs.
Well-represented farmers	EU member states	Well-represented farmers resist monitoring system changes that could reduce their performance for the benefit of under-represented farmers, seeing it as prioritizing a small minority. EU member states and DG AGRI must balance this with the goal of promoting sustainable farming.
Under-represented farmers	EU member states	Under-represented farmers seek tailored fairness interventions to reduce FPs, but these changes may increase system complexity and costs. EU member states prioritize efficiency and balancing FPs and FNs to avoid extra administrative burdens while ensuring law enforcement.

The next conflicts, between under-represented farmers, well-represented farmers, and EU member states, revolve primarily around performance trade-offs. Under-represented farmers, disproportionately affected by FPs, seek fairness interventions to reduce these errors and ensure equitable treatment. However, well-represented farmers, who benefit from the system's high accuracy, resist changes that could introduce performance trade-offs, fearing that reducing FPs for under-represented farmers could lead to increased FPs for them. This tension is further complicated by power imbalances between groups. Well-represented farmers have greater sway in policy decisions given farmer interest groups prioritization of them. This can lead to under-represented farmers feeling that their calls for fairness are overshadowed by the majority's resistance to changes that might negatively affect their performance.

EU member states, responsible for implementing the monitoring system, are caught between CAP's sustainability goals, which favour under-represented farmers, and pressure from well-represented farmers to avoid penalizing the majority. They must also manage administrative burdens and the cost of fairness improvements, which could increase system complexity while only addressing the needs of a small portion of under-represented farmers. Another layer of conflict involves balancing FPs, which create administrative burdens, and FNs, which lead to subsidy misallocation. While farmers may push for more tolerance of FNs, disagreements could arise over how much non-compliance should be allowed. On top of that, each farmer group might have a preference to reduce FNs for the other group, perceiving unequal FNs rates as unfair, given that it makes it easier for one group to escape non-compliance. EU member states must weigh these competing priorities to ensure both fairness and law enforcement.

5.1.4. Recommendations

In this section, I discuss potential solutions to the identified conflicts and connect my research findings to the case study to further elaborate on these solutions. One potential approach to meet the diverse needs of under-represented farmers is to incorporate feedback mechanisms into the compliance system. This would enable ongoing fine-tuning of the system based on real-world outcomes, ensuring that fairness is continuously improved and that new emerging under-represented groups are considered. This approach could reduce uncertainty in the system, possibly reducing conflict. To further ensure the minimization of conflicts, and to encourage feedback, the feedback system would benefit from targeted comprehensible communication of the compliance monitoring system (van der Burg et al., 2021).

To minimize conflicts between under-represented and well-represented farmers, the compliance system could employ fairness optimization strategies that improve equity without introducing performance trade-offs—a type of fairness known as "*no unnecessary harm fairnes*s" (Kaplow & Shavell, 1999). This fairness definition would encourage balancing minimization of FPs for under-represented farmers without introducing more FPs for well-represented farmers.

My research provides specific insights into the compliance system's technical challenges, particularly the issue of performance disparities caused by the thinness of parcels. Parcels with a high proportion of edge pixels introduce noise, impacting classification accuracy—especially for smaller farms, which comprise nearly two-thirds of all EU farms (Eurostat, 2020). A viable mitigation strategy could involve excluding edge pixels from the mean aggregation process, thereby reducing noise. However, given that many small parcels consist mainly of edge pixels in Sentinel-2 imagery, a complete exclusion might be impractical. Instead, to address this issue, it would be wise to experiment with higher-resolution imagery, either across all parcels or specifically for those mainly composed of mixed pixels. However, higher-resolution remote sensing data is very costly, so alternative strategies such as maintaining the field visit compliance system for small parcels could be considered instead. This approach resonates with the ethical considerations highlighted by Barocas et al. (2023), who suggests that, in cases where fairness cannot be guaranteed, it may be more appropriate to rely on alternative methods instead.

Additionally, the research highlighted the benefits of class imbalance correction methods in crop type mapping in Brittany, France, which significantly decreased performance disparities across classes. However, achieving the desired class balance requires a comprehensive evaluation on a test set representative of real-world class distributions to evaluate the trade-offs between overall and minority class accuracy. In applying "*no unnecessary harm fairnes*s", the implementation could leverage Minimax Pareto optimization to find an approach that maximizes minority class accuracy without reducing OA (Nagpal et al., 2024). However, while this approach might be technically feasible, it is hard to say whether it will lead to resolution of conflicts between minor and major crop farmers.

Finally, while this research found no significant sensitivity to topsoil chemical attributes within the *BreizhCrops* dataset, the applicability of this finding across other EU regions remains to be proven. Given the substantial variability in soil attributes in some provinces (Ballabio et al., 2019), further exploration of performance sensitivities related to these attributes could be beneficial to ensure that the system performs fairly across different topsoil profiles. Additionally, it would be valuable to audit performance sensitivities for other sensitive attributes, such as the type of subsidy received or the farmer's income (Rossi, 2019), to ensure that economically underprivileged groups are not disproportionately affected by higher error rates.

6. CONCLUSION

This study applied various fairness optimization methods to address both class imbalance and sensitive attribute imbalance in diverse, mean-aggregated crop type mapping datasets using a transformer classifier. Parcel size was selected as the sensitive attribute due to significant performance disparities between small and large parcels. To tackle these imbalances, two novel methods (DOWCE and RO-R) were specifically developed to address both class and sensitive attribute imbalances, while FL was incorporated because its focus on hard-to-classify samples could theoretically tackle both imbalances as well. These, along with established methods (RO, WCE) and hybrid approaches (RO-DOWCE, RO-FL), were evaluated across the datasets to assess their effectiveness in addressing the imbalances.

The results demonstrated that sample balancing and hybrid methods were more efficient at dealing with class imbalance than cost-sensitive approaches. RO-DOWCE achieved the highest average rank across datasets, and never dropped below third rank. Furthermore, the other novel method, RO-R, scored the second highest average rank. Conversely, the imbalance between small and large parcels was not sufficiently addressed by any method. Instead, all methods improved the performance of both small and large parcels without significantly reducing the performance disparity between the groups. This indicates that higher representation, either in the data or during training, was not sufficient to close the performance gap as initially assumed. It is possible that RO-DOWCE's superior performance was partly due to better noise management, particularly regarding edge pixels in small parcels, which may have allowed the model to better focus on core class features. However, the persistent performance gap suggests that mixed pixels, which are more prevalent in smaller parcels, might still be the limiting factor. Future research should isolate the impact of edge pixels to better understand their role in parcel size imbalance and identify whether other factors aside from mixed pixels are limiting the improvement for this sensitive group.

As AI continues to play a critical role in agricultural monitoring and decision-making, ensuring fairness in crop type mapping is not just a technical challenge but a moral imperative. Addressing imbalances between major and minor crops, as well as large and small parcels, is vital to preventing biased outcomes that could disproportionately affect sensitive groups. While this research has made progress in identifying and addressing key fairness dimensions, the discovery of other sensitive groups are essential steps in ensuring fair AI applications in agriculture.

REFERENCES

Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023). JUSTICE AS FAIRNESS: A HIERARCHICAL FRAMEWORK OF RESPONSIBLE AI PRINCIPLES. *ECIS 2023 Research-in-Progress Papers*. https://aisel.aisnet.org/ecis2023_rip/79

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., & Panagos, P. (2019). Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. *Geoderma*, *355*, 113912. https://doi.org/10.1016/j.geoderma.2019.113912

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/

Becker-Reshef, I., Barker, B., Whitcraft, A., Oliva, P., Mobley, K., Justice, C., & Sahajpal, R. (2023). Crop Type Maps for Operational Global Agricultural Monitoring. *Scientific Data*, *10*(1), 172. https://doi.org/10.1038/s41597-023-02047-9

Becker-Reshef, I., Justice, C., Sullivan, M., Vermote, E., Tucker, C., Anyamba, A., Small, J., Pak, E., Masuoka, E., Schmaltz, J., Hansen, M., Pittman, K., Birkett, C., Williams, D., Reynolds, C., & Doorn, B. (2010). Monitoring Global Croplands with Coarse Resolution Earth Observations: The Global Agriculture Monitoring (GLAM) Project. *Remote Sensing*, *2*(6). https://doi.org/10.3390/rs2061589

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proceedings of the National Academy of Sciences*, *114*(9), 2189–2194. https://doi.org/10.1073/pnas.1616919114

Chawla, N. V. (2010). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 875–886). Springer US. https://doi.org/10.1007/978-0-387-09823-4_45

Cheng, A., Mayes, S., Dalle, G., Demissew, S., & Massawe, F. (2017). Diversifying crops for food and nutrition security—A case of teff. *Biological Reviews*, *92*(1), 188–198. https://doi.org/10.1111/brv.12225

Dean, A. M., & Smith, G. M. (2003). An evaluation of per-parcel land cover mapping using maximum likelihood class probabilities. *International Journal of Remote Sensing*, *24*(14), 2905–2920. https://doi.org/10.1080/01431160210155910

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*(1), 1–30. https://dl.acm.org/doi/10.5555/1248547.1248548

Devos, W., Lemoine, G., Milenov, P., & Fasbender, D. (2018). Technical guidance on the decision to go for substitution of OTSC by monitoring. *Publications Office of the European Union: Luxembourg*. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC112918/jrc112918_published_1.pdf

Elrahman, S. M. A., & Abraham, A. (2013). A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing*, *1*, 332–340. https://ias04.softcomputing.net/jnic2.pdf

Estes, L. D., Ye, S., Song, L., Luo, B., Eastman, J. R., Meng, Z., Zhang, Q., McRitchie, D., Debats, S. R., Muhando, J., Amukoa, A. H., Kaloo, B. W., Makuru, J., Mbatia, B. K., Muasa, I. M., Mucha, J., Mugami, A. M., Mugami, J. M., Muinde, F. W., … Caylor, K. K. (2022). High Resolution, Annual Maps of Field Boundaries for Smallholder-Dominated Croplands at National Scales. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.744863

European Commission. (2022). *COMMON AGRICULTURAL POLICY FOR 2023-2027 28 CAP STRATEGIC PLANS AT A GLANCE*. https://agriculture.ec.europa.eu/document/download/a435881e-d02b-4b98-b718-104b5a30d1cf_en?filename=csp-at-a-glance-eu-countries_en.pdf

Eurostat. (2020). *Farm indicators by legal status of the holding, utilised agricultural area, type and economic size of the farm and NUTS 2 region* [Dataset]. https://ec.europa.eu/eurostat/databrowser/view/ef_m_farmleg/default/table?lang=en

FAO. (2019). *OUR PRIORITIES - The Strategic Objectives of FAO*. https://www.fao.org/3/mg994e/mg994e.pdf

Friedman, M. (1940). A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, *11*(1), 86–92. https://doi.org/10.1214/aoms/1177731944

Fritz, S., See, L., Bayas, J. C. L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van Der Velde, M., Verdin, J., Wu, B., Yan, N., You, L., Gilliams, S., … McCallum, I. (2019). A comparison of global agricultural monitoring systems and current gaps. *Agricultural Systems*, *168*, 258–272. https://doi.org/10.1016/j.agsy.2018.05.010

Geng, Y., & Luo, X. (2019). Cost-Sensitive Convolution based Neural Networks for Imbalanced Time-Series Classification. *Intelligent Data Analysis*, *23*(2), 357–370. https://doi.org/10.3233/IDA-183831

Girgin, S. (2021, October 20). *Using FOSS to develop and operate a geospatial computing platform*. https://doi.org/10.5281/zenodo.6025282

He, E., Xie, Y., Liu, L., Chen, W., Jin, Z., & Jia, X. (2023). Physics Guided Neural Networks for Time-Aware Fairness: An Application in Crop Yield Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12). https://doi.org/10.1609/aaai.v37i12.26664

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2008.239

Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 289–300. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.990132

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the fbietkan statistic. *Communications in Statistics - Theory and Methods*, *9*(6), 571–595. https://doi.org/10.1080/03610928008827904

Jian-jun, W. (2006). Review on Un-mixing Mixed-pixel of Remotely Sensed Data. *Research of Soil and Water Conservation*, *13(5)*, 103–105.

Jin, Z., Azzari, G., Burke, M., Aston, S., & Lobell, D. B. (2017). Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. *Remote Sensing*, *9*(9). https://doi.org/10.3390/rs9090931

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Joshi, A., Pradhan, B., Gite, S., & Chakraborty, S. (2023). Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. *Remote Sensing*, *15*(8). https://doi.org/10.3390/rs15082014

Kaplow, L., & Shavell, S. (1999). The Conflict between Notions of Fairness and the Pareto Principle. *American Law and Economics Review*, *1*(1/2), 63–77. https://dx.doi.org/10.2139/ssrn.161269

Karasiak, N., Dejoux, J.-F., Monteil, C., & Sheeren, D. (2022). Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, *111*(7), 2715–2740. https://doi.org/10.1007/s10994-021-05972-1

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, *177*, 89–100. https://doi.org/10.1016/j.rse.2016.02.028

Khoury, C. K., Bjorkman, A. D., Dempewolf, H., Ramirez-Villegas, J., Guarino, L., Jarvis, A., Rieseberg, L. H., & Struik, P. C. (2014). Increasing homogeneity in global food supplies and the implications for food security. *Proceedings of the National Academy of Sciences*, *111*(11), 4001–4006. https://doi.org/10.1073/pnas.1313490111

Li, C., Wang, Z., Li, B., Peng, Z.-R., & Fu, Q. (2019). Investigating the relationship between air pollution variation and urban form. *Building and Environment*, *147*, 559–568. https://doi.org/10.1016/j.buildenv.2018.06.038

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., & Benediktsson, J. A. (2019). Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(9), 6690–6709. https://doi.org/10.1109/TGRS.2019.2907932

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2017.324

Liu, X., & Zhou, Z. (2006). The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study. *Sixth International Conference on Data Mining (ICDM'06)*, 970–974. https://doi.org/10.1109/ICDM.2006.158

Loughrey, J., Donnellan, T., & Lennon, J. (Eds.). (2016). *The Inequality of Farmland Size in Western Europe*. https://doi.org/10.22004/ag.econ.236341

Lozano-Tello, A., Fernández-Sellers, M., Quirós, E., Fragoso-Campón, L., García-Martín, A., Gutiérrez Gallego, J. A., Mateos, C., Trenado, R., & Muñoz, P. (2021). Crop identification by massive processing of multiannual satellite imagery for EU common agriculture policy subsidy control. *European Journal of Remote Sensing*, *54*(1), 1–12. https://doi.org/10.1080/22797254.2020.1858723

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Meemken, E.-M., & Bellemare, M. F. (2020). Smallholder farmers and contract farming in developing countries. *Proceedings of the National Academy of Sciences*, *117*(1), 259–264. https://doi.org/10.1073/pnas.1909501116

Moskolaï, W. R., Abdou, W., Dipanda, A., & Kolyang. (2021). Application of Deep Learning Architectures for Satellite Image Time Series Prediction: A Review. *Remote Sensing*, *13*(23). https://doi.org/10.3390/rs13234822

Nagpal, R., Shahsavarifar, R., Goyal, V., & Gupta, A. (2024). Optimizing fairness and accuracy: A Pareto optimal approach for decision-making. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00508-4

Nemenyi, P. (1963). *Distribution-free multiple comparisons*. Princeton University. http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:6406278

Netting, R. McC. (1993). *Smallholders, Householders: Farm Families and the Ecology of Intensive, Sustainable Agriculture*. Stanford University Press.

Oneto, L., & Chiappa, S. (2020). Fairness in Machine Learning. In L. Oneto, N. Navarin, A. Sperduti, & D. Anguita (Eds.), *Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)* (pp. 155–196). Springer International Publishing. https://doi.org/10.1007/978-3-030-43883-8_7

Opitz, J., & Burst, S. (2021). *Macro F1 and Macro F1* (arXiv:1911.03347). arXiv. https://doi.org/10.48550/arXiv.1911.03347

Ortigosa-Hernández, J., Inza, I., & Lozano, J. A. (2017). Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, *98*, 32–38. https://doi.org/10.1016/j.patrec.2017.08.002

Orynbaikyzy, A., Gessner, U., Mack, B., & Conrad, C. (2020). Crop Type Classification Using Fusion of Sentinel-1 and Sentinel-2 Data: Assessing the Impact of Feature Selection, Optical Data Availability, and Parcel Sizes on the Accuracies. *Remote Sensing*, *12*(17). https://doi.org/10.3390/rs12172779

Pallante, G., Drucker, A. G., & Sthapit, S. (2016). Assessing the potential for niche market development to contribute to farmers' livelihoods and agrobiodiversity conservation: Insights from the finger millet case study in Nepal. *Ecological Economics*, *130*, 92–105. https://doi.org/10.1016/j.ecolecon.2016.06.017

Pellegrini, L., & Tasciotti, L. (2014). Crop diversification, dietary diversity and agricultural income: Empirical evidence from eight developing countries. *Canadian Journal of Development Studies / Revue Canadienne d'études Du Développement*, *35*(2), 211–227. https://doi.org/10.1080/02255189.2014.898580

Rajendran, S., Afari-Sefa, V., Shee, A., Bocher, T., Bekunda, M., dominick, I., & Lukumay, P. J. (2017). Does crop diversity contribute to dietary diversity? Evidence from integration of vegetables into maize-based farming systems. *Agriculture & Food Security*, *6*(1), 50. https://doi.org/10.1186/s40066-017-0127-3

Renard, D., & Tilman, D. (2019). National food production stabilized by crop diversity. *Nature*, *571*(7764), 257–260. https://doi.org/10.1038/s41586-019-1316-y

Rezaei-Dastjerdehei, M. R., Mijani, A., & Fatemizadeh, E. (2020). Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function. *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, 333–338. https://doi.org/10.1109/ICBME51989.2020.9319440

Rossi, R. (2019). *Understanding farmer income*. European Parliament, Think Tank. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2019)637924

Rußwurm, M., & Körner, M. (2020). Self-attention for raw optical Satellite Time Series Classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *169*, 421–435. https://doi.org/10.1016/j.isprsjprs.2020.06.006

Rußwurm, M., Lefèvre, S., & Körner, M. (2019). *BreizhCrops: A Satellite Time Series Dataset for Crop Type Identification*. ICML Time Series Workshop

Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S., & Körner, M. (2020). BreizhCrops: A Time Series Dataset for Crop Type Mapping. *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, *43*(B2), 1545–1551. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1545-2020

Sabato, S., Treister, E., & Yom-Tov, E. (2024). *Fairness and Unfairness in Binary and Multiclass Classification: Quantifying, Calculating, and Bounding* (arXiv:2206.03234; Version 2). arXiv. https://doi.org/10.48550/arXiv.2206.03234

Samberg, L. H., Gerber, J. S., Ramankutty, N., Herrero, M., & West, P. C. (2016). Subnational distribution of average farm size and smallholder contributions to global food production. *Environmental Research Letters*, *11*(12), 124010. https://doi.org/10.1088/1748-9326/11/12/124010

Savage, S., & Lei Win, T. (2023, June 29). The truth behind Europe's most powerful farmers lobby. *POLITICO*. https://www.politico.eu/article/copa-cogeca-farmering-lobby-europe/

Serna, I., Morales, A., Fierrez, J., & Obradovich, N. (2022). Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, *305*, 103682. https://doi.org/10.1016/j.artint.2022.103682

Shadman Roodposhti, M., Aryal, J., Lucieer, A., & Bryan, B. A. (2019). Uncertainty Assessment of Hyperspectral Image Classification: Deep Learning vs. Random Forest. *Entropy*, *21*(1). https://doi.org/10.3390/e21010078

Sinha, A., Basu, D., Priyadarshi, P., Ghosh, A., & Sohane, R. K. (2022). Farm Typology for Targeting Extension Interventions Among Smallholders in Tribal Villages in Jharkhand State of India. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.823338

Soule, M. J. (2001). Soil Management and the Farm Typology: Do Small Family Farms Manage Soil and Nutrient Resources Differently than Large Family Farms? *Agricultural and Resource Economics Review*, *30*(2), 179–188. https://doi.org/10.1017/S106828050000112X

Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. *Engineering Reports*, *3*(4), e12298. https://doi.org/10.1002/eng2.12298

Tahaei, M., Constantinides, M., Quercia, D., & Muller, M. (2023). *A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI* (arXiv:2302.05284). arXiv. http://arxiv.org/abs/2302.05284

Termeer, C. J. A. M., Dewulf, A., & Biesbroek, R. (2019). A critical assessment of the wicked problem concept: Relevance and usefulness for policy science and practice. *Policy and Society*, *38*(2), 167–179. https://doi.org/10.1080/14494035.2019.1617971

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., Haert, F. C. van der, Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., Wever, W. de, … Clopath, C. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15871-z

UN. (2015). *20130 Agenda for Sustainable Development*. https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

UNESCO. (2022). *Recomendation on the Ethisc of Artifical Intellegence*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

van der Burg, S., Wiseman, L., & Krkeljas, J. (2021). Trust in farm data sharing: Reflections on the EU code of conduct for agricultural data sharing. *Ethics and Information Technology*, *23*(3), 185–198. https://doi.org/10.1007/s10676-020-09543-1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Waldner, F., Chen, Y., Lawes, R., & Hochman, Z. (2019). Needle in a haystack: Mapping rare and infrequent crops using satellite imagery and data balancing methods. *Remote Sensing of Environment*, *233*, 111375. https://doi.org/10.1016/j.rse.2019.111375

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558. https://doi.org/10.1145/3351095.3375709

Zhao, P., Luo, C., Qiao, B., Wang, L., Rajmohan, S., Lin, Q., & Zhang, D. (2022). T-SMOTE: Temporal-oriented Synthetic Minority Oversampling Technique for Imbalanced Time Series Classification. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2406–2412. https://doi.org/10.24963/ijcai.2022/334

Zhu, T., Luo, C., Zhang, Z., Li, J., Ren, S., & Zeng, Y. (2022). Minority oversampling for imbalanced time series classification. *Knowledge-Based Systems*, *247*, 108764. https://doi.org/10.1016/j.knosys.2022.108764

DATA AVALAIBILITY

The *BreizhCrop* dataset, which forms the basis of this research, is publicly available at:

https://github.com/dl4sits/breizhcrops

The supplementary code used to conduct research, is provided in the following repository:

https://github.com/ilyagorb/fair_breizhcrops

This repository also includes additional results, such as individual results for each dataset-method combination.

APPENDIX: AUXILIARY EXPERIMENTS AND DATASETS

Table 5 Auxiliary datasets. Note that AD-1 is the only dataset with thin parcels

id	Used in experiment/s	Description	Total Training Samples	Total Testing Samples
AD 1	AE 1	Full dataset (with thin parcels)	606239	157558
AD 2	AE 2	Full dataset	534891	137648
AD 3	AE 2	10k training cap	110074	137648
AD 4	AE 2	10k training, 3k testing cap	110074	34107
AD 5	AE 4, AE 5	Full training, 3k testing cap	534891	34107

Table 6 Auxiliary experiments

id	Used dataset/s	Objective	Explained in section	Results section	Main Findings
AE 1	AD 1	Removal of Thin Parcels	3.2.1	4.1.1	Thin parcels with an area-to-perimeter ratio < 6 had significantly lower performance.
AE 2	AD 2, AD 3, AD 4	Dataset Reduction	3.2.2	4.1.2	Caps maintained reasonable performance while reducing the computational cost.
AE 3	ED 1	Stochasticity test	3.1	NA	Stochasticity controlled with seed, results identical in repeated runs for the same seed.
AE 4	AD 5	Sensitive attribute identification	3.3	4.2	Parcel area identified as the most sensitive attribute.
AE 5	AD 5	Identifying parcel area threshold	3.5	4.3	6,170 metres identified as the threshold separating small and large parcels.

DECLARATION OF AI USE

In accordance with the AI guidelines from the University of Twente:

> *During the preparation of this work, I utilized several AI assisted tools to support this thesis. ChatGPT was used to aid with code-writing, and due diligence was applied to ensure the accuracy and correctness of the code. Additionally, I used Consensus, Semantic Scholar, and Bing Co-Pilot to help find relevant literature. Grammarly was also employed to correct grammatical errors, and both ChatGPT and Grammarly was occasionally used to help brainstorm with reformulating sentences to ensure clarity and conciseness.*
>
> *I take full responsibility for the content of this thesis.*