remote sensing

MDPI

Article

Crop Mapping without Labels: Investigating Temporal and Spatial Transferability of Crop Classification Models Using a 5-Year Sentinel-2 Series and Machine Learning

Tomáš Rusňák [1,*], Tomáš Kasanický [2], Peter Malík [2], Ján Mojžiš [2], Ján Zelenka [2], Michal Sviček [3], Dominik Abrahám [3] and Andrej Halabuk [1]

[1] Institute of Landscape Ecology, Slovak Academy of Sciences, v.v.i, Štefánikova 3, 814 99 Bratislava, Slovakia; andrej.halabuk@savba.sk
[2] Institute of Informatics, Slovak Academy of Sciences, v.v.i, Dúbravská Cesta 9, 845 07 Bratislava, Slovakia; tomas.kasanicky@savba.sk (T.K.); p.malik@savba.sk (P.M.); jan.mojzis@savba.sk (J.M.); jan.zelenka@savba.sk (J.Z.)
[3] National Agricultural and Food Center (NPPC), Hlohovecká 2, 951 41 Lužianky, Slovakia; michal.svicek@nppc.sk (M.S.); dominik.abraham@nppc.sk (D.A.)
* Correspondence: tomas.rusnak@savba.sk

Abstract: Multitemporal crop classification approaches have demonstrated high performance within a given season. However, cross-season and cross-region crop classification presents a unique transferability challenge. This study addresses this challenge by adopting a domain generalization approach, e.g., by training models on multiple seasons to improve generalization to new, unseen target years. We utilize a comprehensive five-year Sentinel-2 dataset over different agricultural regions in Slovakia and a diverse crop scheme (eight crop classes). We evaluate the performance of different machine learning classification algorithms, including random forests, support vector machines, quadratic discriminant analysis, and neural networks. Our main findings reveal that the transferability of models across years differs between regions, with the Danubian lowlands demonstrating better performance (overall accuracies ranging from 91.5% in 2022 to 94.3% in 2020) compared to eastern Slovakia (overall accuracies ranging from 85% in 2022 to 91.9% in 2020). Quadratic discriminant analysis, support vector machines, and neural networks consistently demonstrated high performance across diverse transferability scenarios. The random forest algorithm was less reliable in generalizing across different scenarios, particularly when there was a significant deviation in the distribution of unseen domains. This finding underscores the importance of employing a multi-classifier analysis. Rapeseed, grasslands, and sugar beet consistently show stable transferability across seasons. We observe that all periods play a crucial role in the classification process, with July being the most important and August the least important. Acceptable performance can be achieved as early as June, with only slight improvements towards the end of the season. Finally, employing a multi-classifier approach allows for parcel-level confidence determination, enhancing the reliability of crop distribution maps by assuming higher confidence when multiple classifiers yield similar results. To enhance spatiotemporal generalization, our study proposes a two-step approach: (1) determine the optimal spatial domain to accurately represent crop type distribution; and (2) apply interannual training to capture variability across years. This approach helps account for various factors, such as different crop rotation practices, diverse observational quality, and local climate-driven patterns, leading to more accurate and reliable crop classification models for nationwide agricultural monitoring.

Keywords: multitemporal classification; Google Earth Engine; within-season crop mapping; domain adaptation; agricultural monitoring; crop monitoring

Citation: Rusňák, T.; Kasanický, T.; Malík, P.; Mojžiš, J.; Zelenka, J.; Sviček, M.; Abrahám, D.; Halabuk, A. Crop Mapping without Labels: Investigating Temporal and Spatial Transferability of Crop Classification Models Using a 5-Year Sentinel-2 Series and Machine Learning. *Remote Sens.* **2023**, *15*, 3414. https://doi.org/10.3390/rs15133414

Academic Editors: Kevin Tansey and Yuanwei Qin

Received: 9 May 2023
Revised: 23 June 2023
Accepted: 30 June 2023
Published: 5 July 2023

1. Introduction

The open data policy of the Sentinel-2 mission has revolutionized field-level crop mapping. The application of diverse machine learning classifiers that can effectively capture

crop-specific spectrotemporal characteristics, ultimately enhancing the performance of multitemporal crop classification for a given season and region [1]. Nevertheless, cross-season and cross-region crop classification presents a unique challenge. These scenarios require models that can effectively generalize across different growing conditions, agricultural practices, and phenological patterns. This challenge has been a topic of interest in the remote sensing community for quite some time. In the field of data science, this issue has often been addressed by domain adaptation, which focuses on adjusting a source model to operate effectively within a specific target domain [2]. For example, recent studies solving crop mapping without labels have utilized the harmonization of varied crop phenology between source and target domains in order to test the transferability of crop classification models across regions [3] or seasons [4]. Another study by Lin et al. [5] adjusted the source and target domains based on the topology of different crops in the spectral feature space to generate crop labels in an unseen target year and applied the trained classification model to these labels for crop mapping.

In our study, we implemented a domain generalization approach, which aims to train a model on multiple source domains, often with diverse characteristics, so that it can generalize well to new target domains without requiring any fine-tuning or adaptation. In fact, this approach is intrinsically linked to remote sensing-based land cover mapping addressing large geographical areas or temporal scales, and early attempts to address this challenge were referred to as signature extension or generalization [6]. Recently, numerous crop mapping studies have emerged, using various terms related to this subject, such as transductive transfer learning (TTL) [7], random forest transfer (RFT) [8], temporal extendibility mapping [9], or transferring decision boundary (TDB) [5]. To minimize terminology misunderstandings, we will refer simply to model transferability or domain generalization capabilities in our study. This defines the desired feature of trained models, which means that even if the test validation data are captured from different years or different geographical regions [10]. Table 1 summarizes various crop classification studies that have utilized satellite products similar to Landsat or Sentinel and investigated the transferability of crop classification models across different datasets.

Table 1. List of the crop classification transferability studies using Sentinel-2 like products. When multiple cases were explored in a single study, only the best-case scenarios were reported.

Transfer Scenario	Satellite Platform	Source Domain	Target Domain	Geographical Location	Crop Configuration/ Nomenclature	Classifier/ Method	Accuracy	Reference
Spatial	Sentinel-2	England, France	10 Countries	Europe	4 crops	RF	89%	[7]
Spatial	Sentinel-2	Zeeland region	Flevo-land, Friesland	Netherland	10 crops	Dynamic Time Warping	69–75%	[3]
Temporal	Landsat	2006–2010 [a]	2006–2010	Kansas	3 crops	RF	83.4%	[9]
Temporal	Sentinel-2	2017–2018	2019	Midwest US, NE China, Hauts-de-France	3, 4, 8 crops	RF	90.7; 89.8; 83.7%	[5]
Temporal	Sentinel-1/ Sentinel-2	2020	2019	Hetao Irrigation District	6 crops	RF	92%	[11]
Temporal	Sentinel-1/ Sentinel-2	2017	2018	Heilongjang	4 crops	RF	91%	[12]
Temporal	Sentinel-2	2016–2019	2020	16 States across USA	3 crops	RF	71.3 [b]	[13]
Temporal	Landsat	2010–2015	2016 [c]	9 States across USA	3 crops	RF	70%	[8]
Temporal	Landsat	2000–2014	2015	Illinois	2 crops	DNN	96%	[14]

[a] The cross-year setup was not generalized across multiple years; instead, one year's data was used for training and another year's data for testing. This was done for all combinations of train/test years, resulting in 25 cross-year classifications. Thus, the overall accuracy represents the mean of all overall accuracies. [b] Sentinel-2 data only. [c] The cross-year setup was not generalized across multiple years; instead, one year's data was used for training and 2016 for testing.

In our study, we address limitations of the previous Landsat-based studies, such as lower temporal frequency and the absence of red-edge spectral bands, by employing temporal composites of Sentinel-2 spectral reflectance products. In particular, we capitalize on a 5-year series of Sentinel products and reference data, enabling a more in-depth exploration of the potential of Sentinel-2 data in across-year crop mapping. Certain studies have reported generalization workflows in relatively homogenous conditions with a relatively simple crop scheme, involving only two or three dominant classes [14]. In our study, we aim to test the performance of classification models in a more diverse agricultural setting that includes all relevant crops, resulting in an 8-class nomenclature. This will provide a more comprehensive understanding of the generalization abilities of the algorithms in complex and varied agricultural landscapes. Lastly, some notable studies, such as Johnson et al. [13], investigated generalization capacity across various spatial domains. However, they focused on using only one machine learning algorithm, namely random forest, which may not provide a comprehensive understanding of the impact of different algorithms on the generalization issue. Therefore, in our study, we aim to test multiple machine learning algorithms to determine whether the choice of algorithm is crucial for classification performance in different transferability scenarios.

In summary, our overarching goal is to develop a robust, efficient, and accurate crop mapping workflow with strong cross-year and region generalization capabilities, ultimately enabling effective nationwide crop monitoring. Accordingly, the specific objectives of this study can be defined as follows:

1. Developing an effective workflow for broad-scale crop mapping using machine learning techniques that can be easily deployed for nationwide agricultural monitoring.
2. Investigating the transferability capacities of the developed crop classification models in both temporal and spatial aspects.
3. Providing analysis-ready datasets to the remote sensing community for further testing and supporting the on-going development of improved methods.

2. Materials and Methods

2.1. Study Area

The two study regions encompass the most agriculturally productive areas in Slovakia, significantly influencing agro-commodity prices and maintaining crucial supply chain connections. Combined, they account for approximately 63% of the total arable land in Slovakia, with the Danubian Lowlands covering around 52% and the eastern Slovakian Lowlands covering approximately 11%.

The Danubian Lowland (later referred to as "Danube"), a part of the greater Pannonian Basin, is located in southwestern Slovakia and covers an area of 9820 km^2 (Figure 1). It consists of two parts: the Danubian Upland, with slightly undulating hilly relief in the north, and the Danubian Flat Depressions, with a predominantly flat relief in the south. The region is the warmest in Slovakia, with an average annual temperature between 9 and 10 °C and a mean annual rainfall between 550 and 700 mm [15]. Favorable climatic and geomorphological conditions, combined with fertile soils, make intensive agriculture the prevailing land use type. The land cover consists mainly of croplands and compact rural settlements. The region's soil cover is unique within Slovakia, featuring a dominant fluvial relief with extensive areas of Chernozems and Fluvisols. The alkalinity of the region's soils is due to the presence of carbonate parent materials and groundwater, with soils predominantly rich in carbonates. The soil texture is primarily loamy.

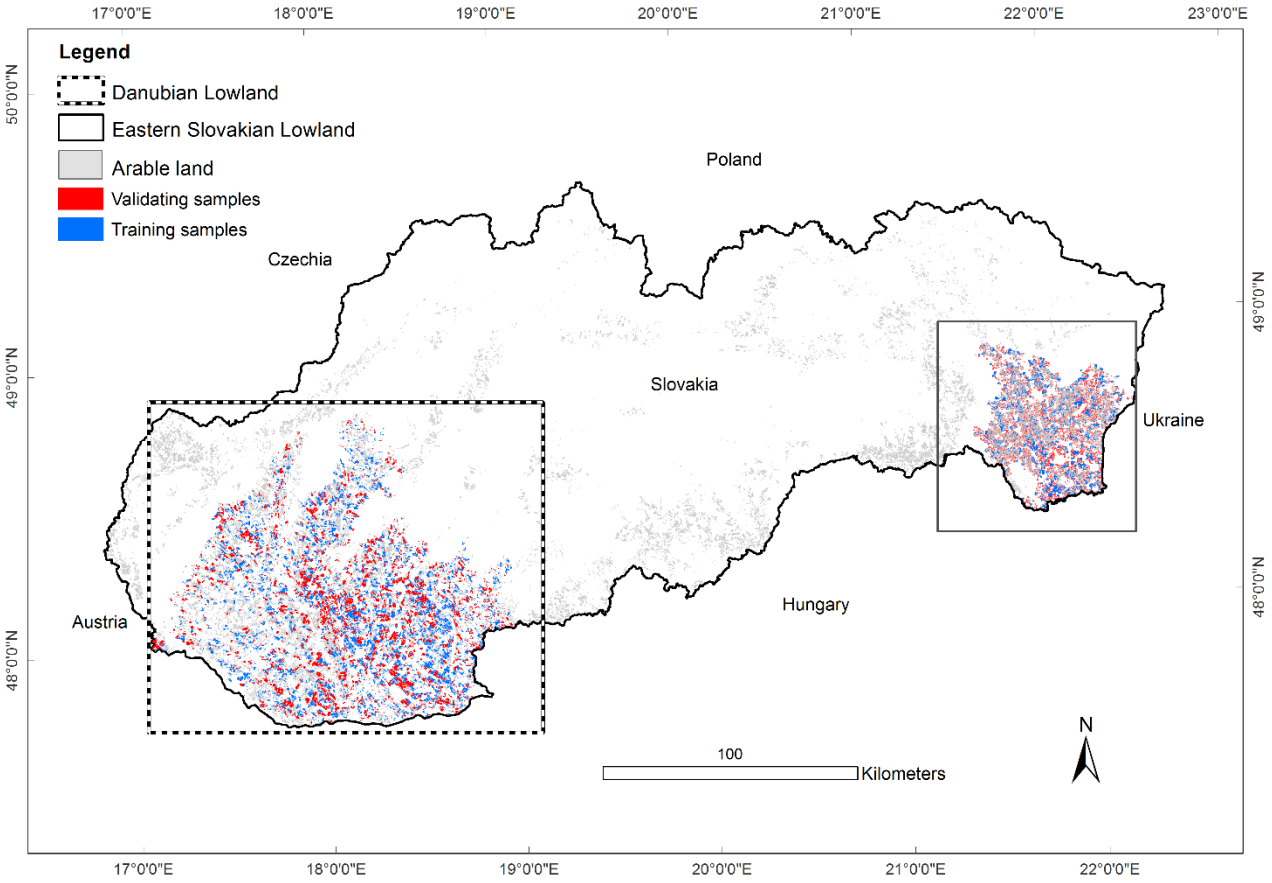

Figure 1. Study areas and distribution of the croplands.

The eastern Slovakian lowland (later referred to as "East"), covering 2500 km^2, is the northeastern extension of the Tisza plain and consists of the eastern Slovakian flat depression and the eastern Slovakian upland. The flat depression (1800 km^2) is a tectonic depression with a young structural plain formed by river accumulation during the Pleistocene and Holocene and shaped by aeolian activity, loess deposition, and windblown sands. The Upland (700 km^2) features gently hilly terrain with flat ridges separated by shallow valley-like depressions formed by water-course accumulation. The eastern Slovakian lowlands have Slovakia's most continental climate, characterized by significant temporal weather variability. Long-term meteorological observations indicate suitability for intensive crop cultivation with low irrigation requirements. However, the low gradient and weak water permeability of some soils impede rapid water runoff, causing stagnation in depressions. High groundwater levels influence extensive areas, leading to soil waterlogging, gleiing, and salinization [16]. The majority of soils in the region are used for intensive agriculture. Hydromorphic soils, including Fluvisols, Pseudogleys, and Gleys. Gleiing processes are driven by the dense river network and high groundwater levels. Although the eastern Slovakian lowland lacks highly productive arable soils, it remains a productive agroecosystem. However, soil parameters make crop cultivation more challenging in terms of agronomic and economic aspects.

Differences between the Danubianl lowland and the eastern Slovakian lowland can be observed in soil texture and climate conditions [17]. In terms of soil texture, the Danubian lowland soils are predominantly loamy, with sandier and partially silty fractions in the C-horizon and a higher carbonate content, which contributes to better soil structure. In contrast, the eastern Slovakian lowland soils contain a higher clay fraction, making them heavier, ranging from moderately heavy to very heavy, and often leading to waterlogging, gleiing, and saline soil development in some areas, resulting in a less favorable soil structure. Regarding climate conditions, due to its geographical location, the eastern Slovakian

lowland has the most continental climate in Slovakia, characterized by cold winters and warm summers. On the other hand, the Danubian lowlands have a milder climate but are considered one of the warmest regions in Slovakia. Variations in geographical conditions, such as those in the eastern Slovakian lowlands, can influence the distribution and proportion of predominant crops, as evident in the higher proportion of grasslands only suitable in heavy, gleied soils (Figure 2). Additionally, specific agro-commodity trade relationships may also play a role in shaping the crop distribution in the area, which might be the case for soybeans and sugar beet. Specifically, sugar beet production was not carried out in 2018 and 2019 in eastern Slovakia due to the collapse of regional sugar mill companies. However, in 2020, the farmers experienced a revival with the establishment of new supply chain connections at Hungarian sugar mills.

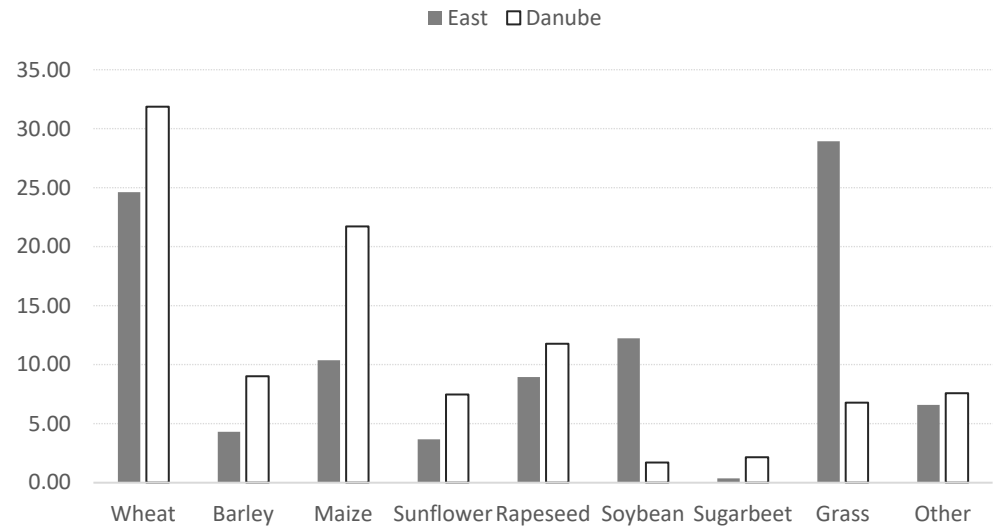

Figure 2. Typical crop proportions (% on *y*-axis) in the two studied regions based on official agricultural statistics. Available online www.statistics.sk (accessed on 25 May 2023).

2.2. Satellite Data

The data collected by the multispectral instrument (MSI) onboard the Sentinel-2A and Sentinel-2B satellites (later referred to as Sentinel-2) were used in the study. In particular, we used Level-2A bottom of atmosphere reflectance products (BOA), which are available in the Google Earth Engine (GEE) catalog (ImageCollection ID: COPERNICUS/S2_SR). This product contains BOA reflectances of 13 Sentinel-2 spectral bands and is accompanied by the scene classification (SCL), including quality indicators such as cloud shadow detection, cloud probabilities, and cirrus mask (https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/processing-levels/level-2, accessed on 20 April 2023). The BOA reflectances were performed at their native spatial resolutions, depending on the spectral characteristics of the respective band. In our analyses, we used Sentinel-2 spectral bands B2, B3, B4, and B8 at native 10 m resolution and bands B5, B6, B7, B8A, B11, and B12 at native 20 m resolution. Three Sentinel-2 tiles were used to fully cover the extent of the study area, namely the 33UYP tile (covering 72.52% of the study area), the 33UXP tile (covering 24.84% of the study area), and the 33UYQ tile (covering 2.64% of the study area) in the case of the Danubian lowland region, and two tiles—34UEV (47.34%) and 34UEU (52.66%) in the eastern Slovakian lowland region. Only products produced between April and August were used to prevent any quality concerns that might have occurred outside of this time period. At the same time, this time range could properly fit the seasonal development of the crops. As a result, we chose all Sentinel-2 products that were accessible on the GEE platform between 2018 and 2022 (totaling 771 images covering Danubian lowland and 215 images covering eastern Slovakia lowland). Our decision to use a 5-year period for this study was primarily due to the availability of consistent, harmonized Sentinel-2 data

collection for our targeted research areas—the Danubian lowland and the eastern Slovak lowland—from 2018.

2.3. Methodology

The general workflow (Figure 3) involved the preprocessing of Sentinel-2 satellite products, creating input features, training different classification models, performing performance analysis based on independent test data, performing feature variable importance analysis, and applying the selected models to images.

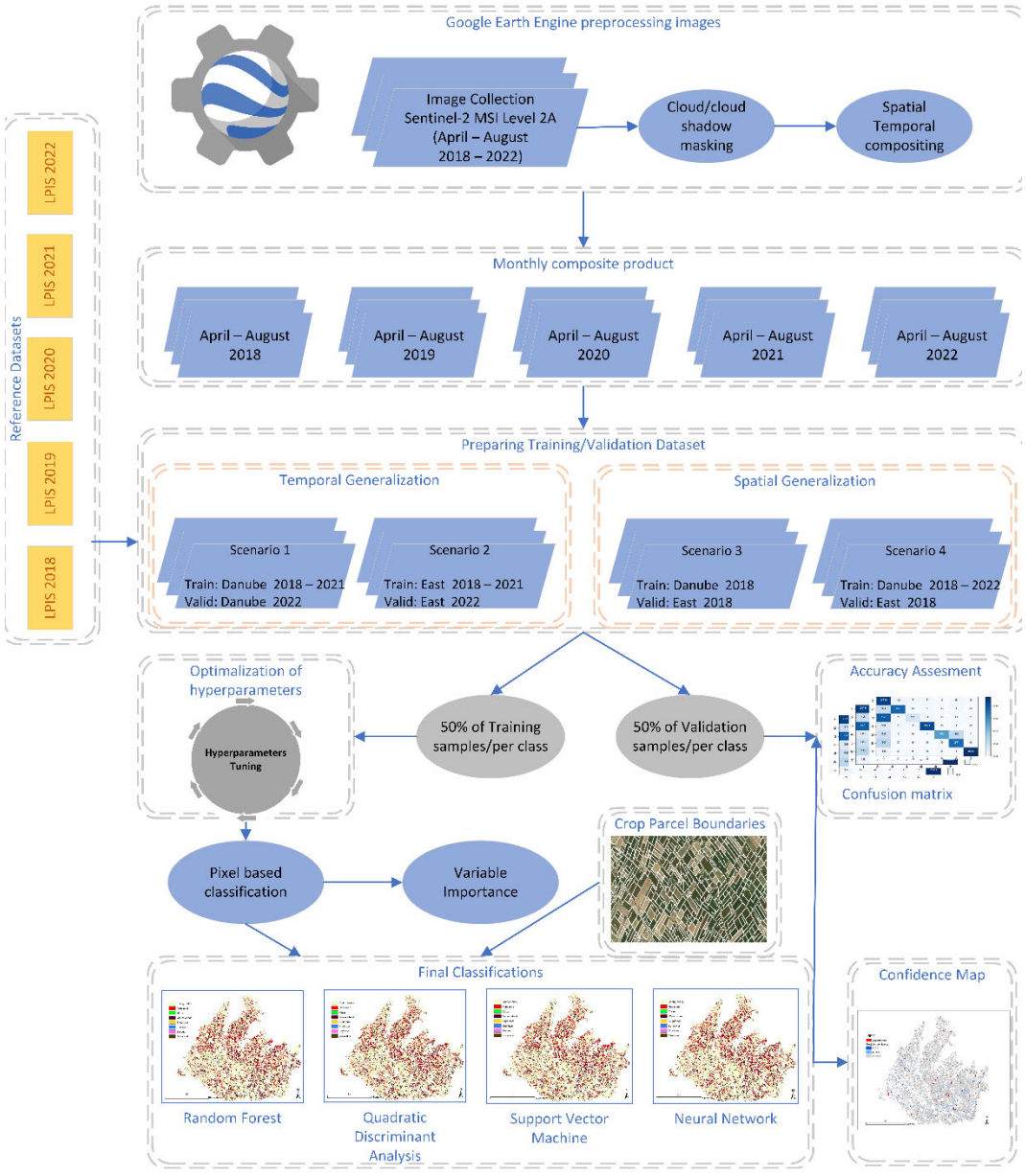

Figure 3. Overall workflow of the study.

- Preprocessing of satellite products.
- Creating a reference dataset by extracting spectro-temporal information.
- Training different classification models.
- Analyzing performance and accuracy assessment.
- Applying models to images.

2.3.1. Preprocessing of Satellite Products

Spatiotemporal compositing followed the obvious steps of so-called pixel-based compositing [18], namely selecting good observations within the compositing period, prioritizing, and assigning target values at the pixel level. We selected the good observations by masking clouds, cloud shadows, and cirrus provided in Level-2A, accompanied by the SCL (scene classification) product. A simple median compositing method was used for producing temporal composite products over monthly time periods, beginning from April 2018 to August 2022. The median method prioritizes observations closest to the central tendency within a given period and is more resilient against outliers (e.g., unfiltered cloud remnants) than the mean, though it is only effective when the majority of selected observations are of good quality. In particular, we computed median values for each spectral band and pixel separately for a given period using the GEE "imageCollection" function. The same approach has been widely used in many crop multitemporal classification studies. Monthly composites from April to August in a given year were used to create seasonal time series for each year: 2018, 2019, 2020, 2021, and 2022.

2.3.2. Creating Training and Test Datasets

The land parcel information system (LPIS) serves as the spatial foundation for implementing the EU Common Agricultural Policy at the national level. While some agricultural data, including crop types per parcel, is publicly available, we selected specific data that could effectively train the classification algorithms according to the project's objectives. To minimize the potential issue of crop mixture within parcels, we applied the following filters to the LPIS data: Only parcels in which more than 90% of the declared crop types were selected for a given year were selected, and less common crop types (less than 10%) were filtered out. From this selection process, we identified eight crop types: rapeseed, barley, wheat, sunflower, corn, sugar beet, soybean, and permanent grasslands. To ensure a balanced distribution of crop classes, we conducted random sampling per year and extracted 2000 pixels per class/year, except for 2020 and 2022, where we could extract only 1698 and 1458 pixels per class, respectively, due to observational limitations. This resulted in a total of 9156 pixels/case. This dataset had been randomly split into training (50%) and test (50%) datasets (for each year/class), allowing independent validation. Three main classification scenarios were tested to explore the generalization capacities of different ML models using the so-called held-out concept. In this approach, we held out specific subsets of data, such as certain spatial regions or temporal periods, to evaluate how well the trained models can generalize to previously unseen conditions and variations. Each scenario was tested for each year. Table 2 provides an example for the tested year 2022.

Table 2. Examples of testing scenarios when target unseen year represent 2022 season.

	Train Region	Train Years	Test Region	Test Year
Scenario 1	Danube	2018, 2019, 2020, 2021	Danube	2022
Scenario 2	East	2018, 2019, 2020, 2021	East	2022
Scenario 3	Danube	2022	East	2022
Scenario 4	Danube	2018, 2019, 2020, 2021, 2022	East	2022

Scenario 1: Scenario 1 investigates the temporal generalization capacity of machine learning (ML) classification models by assessing their inter-year performance in the Danubian lowland region using training data from four years and reserving an unseen year for testing purposes. Using a total of five years of data, this temporal generalization assessment was iterated four times, each time holding out a different year for testing while training on the remaining four years.

Scenario 2: Scenario 2 shared the same objective as Scenario 1, focusing on temporal generalization capacity, but applied to a different region—the eastern Slovakian lowland.

Scenario 3: Scenario 3 examined the spatial generalization capacity of ML classification models by evaluating their performance across distinct spatial domains, training on data from the Danubian lowlands, and testing on unseen data from the eastern Slovakian lowlands within the same year.

Scenario 4: Scenario 4 assessed both temporal and spatial generalization capacities of ML classification models by evaluating their performance across distinct spatial regions and different years, training on five years of data from the Danubian lowlands and testing on each year of unseen data from the eastern Slovakian lowlands.

These scenarios emulate realistic crop mapping situations where timely reference label data might be partially or completely missing. Scenarios 1 and 2 simulate situations where labeled data exists for certain years but not others, reflecting real-life long-term crop monitoring constraints. These scenarios help us evaluate the temporal generalization capacity of machine learning classifiers, which is crucial for within-season crop mapping when current-season labeled data may be unavailable. Scenarios 3 and 4, on the other hand, mimic conditions of spatial variability in labeled data availability. They assess the spatial generalization capabilities of our models across diverse geographic conditions, mirroring large-scale crop monitoring contexts. Scenario 3 represents cases with single-year labeled data, while Scenario 4 covers instances with multiple years of labeled data. Overall, each scenario offers valuable insights into the robustness of ML models in the temporal and spatial dimensions of crop mapping, reflecting the challenges of real-world crop mapping projects.

The training and testing datasets were prepared accordingly to create 4 distinct training and 4 testing datasets, each representing a specific classification scenario for evaluating the generalization capacities of different ML models.

2.3.3. Training Different Models

In the preliminary analysis, we evaluated the performance of various ML classification models available in MATLAB's Classification Learner toolbox. Based on the overall accuracy measure estimated on the Scenario 1 dataset, we selected the following four best-performing classification models for further analyses: quadratic discriminant analysis (QDA), support vector machines (SVM), random forests (RF), and neural networks (NN). For all classification tasks, we employed 5-fold cross-validation and used Bayesian optimization to automatically determine and apply the optimal hyperparameters for relevant models. For all classification tasks, we employed 5-fold cross-validation and used Bayesian optimization to automatically determine and apply the optimal hyperparameters for relevant models [19]. This approach ensured that the selection of hyperparameters was more objective and systematic and less influenced by individual biases or prior knowledge. By using Bayesian optimization, the algorithm was able to systematically explore the hyperparameter space and find the best combination of hyperparameters for each model and dataset. This approach was particularly important in ensuring the reliability and generalizability of the models. The hyperparameters selected during the training process for all classification scenarios and years are presented in Table S1 of Supplement S2. From this tuning process, we compiled statistics, which are outlined in Table 3. This information can help guide the determination of plausible ranges for key hyperparameters. However, it is important to acknowledge that, due to the complexity of hyperparameter tuning, it should be deemed specific to a given classification problem and data.

Table 3. Aggregated Statistics from Hyperparameter Tuning.

Algorithm	Hyperparameter	Type/Statistic	Frequency/Value
SVM	Kernel function	Gaussian	10 *
		Quadratic	7 *
		Cubic	3 *
	Box constraint level [1]	Mean	445.14
		Min	4.36
		Max	961.49
	Kernel scale	Mean	13.02
		Min	5.26
		Max	23.83
Neural Network	Fully connected layers	Layer 1	9 *
		Layer 2	9 *
		Layer 3	2 *
	Activation function [2]	Tanh	12 *
		Relu	6 *
		Sigmo	2 *
	First layer size	Mean	154
		Min	10
		Max	298
	Second layer size	Mean	64
		Min	11
		Max	176
	Third layer size	Mean	23
		Min	16
		Max	29
	Regularization strength (Lambda)	Mean	2.72×10^{-6}
		Min	4.49×10^{-8}
		Max	7.62×10^{-6}
RF	Number of learners	Mean	360
		Min	32
		Max	500
	Number of predictors to sample	Mean	16
		Min	3
		Max	46
	Max. number of splits	Mean	13,837
		Min	1104
		Max	39,619

[1] For Gaussian function. [2] Tanh—Hyperbolic tangent activation, ReLU-Rectified linear unit activation, Sigmo-Sigmoid activation. * Frequency depicts the number of times a given parameter was used out of the 20 classification tasks (scenario/year).

Random forests (RF), introduced by [20], are an ensemble learning method frequently utilized and highly popular in remote sensing applications. By combining multiple decision trees, it creates a more accurate and robust model. This method is particularly advantageous due to its intuitive and easily explainable nature, in contrast to more complex models like neural networks. Random forests employ ensemble techniques such as bagging to improve predictive accuracy and reduce overfitting. Each tree in the ensemble is built using a random subset of the training data and a random subset of features, contributing to the diversity of the individual trees. The random forest algorithm enhances the stability and accuracy of classification models, addressing noise in training data and overfitting issues. Furthermore, RF is capable of handling large datasets, missing values, and can perform variable importance estimation, making it a versatile tool for various applications. Within the Bayesian optimization work-flow, three hyperparameters were optimized: the number of trees, the number of predictors, and the number of splits.

Discriminant analysis (DA) is a statistical technique used in pattern recognition and classification tasks. It focuses on identifying linear or quadratic combinations of features that best separate instances into different classes or groups. This method aims to maximize the between-class variance while minimizing the within-class variance, resulting in effective

class discrimination. There are two main types of discriminant analysis: linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). LDA assumes equal covariance matrices for the classes and finds a linear combination of features for optimal separation, while QDA allows for different covariance matrices and results in a quadratic decision boundary. Discriminant analysis is commonly employed in remote sensing to classify and analyze data based on its underlying patterns [21]. In fact, QDA is often referred to as maximum likelihood classification within the remote sensing community due to its implementation in widely used image processing software. QDA captures spectral characteristics in remote sensing data and classifies observations based on the class that maximizes likelihood. Its ability to represent quadratic decision boundaries makes QDA suitable for handling complex, high-dimensional data and providing accurate classification results in diverse scenarios. In the course of the optimization process, we examined various discriminant analysis types and covariance matrix configurations to identify the optimal model for our classification tasks. In each case, QDA emerged as the top performer; hence, we will refer to QDA in the remainder of the paper.

Support Vector Machines (SVM) are powerful and robust predictive methods based on statistical learning commonly used in remote sensing applications [22]. SVM maps training examples to points in high-dimensional space and finds the optimal separating hyperplane, maximizing the margin between categories. This approach effectively reduces overfitting and performs well when classes are clearly separated in the multidimensional feature space. Additionally, SVM is known for its adaptability to diverse data distributions and its capability to handle large feature spaces. For multiclass classification with three or more classes, SVM divides the problem into binary classification subproblems, using one SVM learner per subproblem. However, as a non-probabilistic classifier, SVM does not provide probabilistic explanations for classification results. During the optimization process, hyperparameters to be optimized include the type of kernel functions, box constraint level, kernel scale, and multiclass strategy.

Neural networks (NN) are a set of algorithms inspired by the neural interactions in the brain, designed to recognize patterns in data by simulating the way neurons process and transmit information. NNs have been successful in handling a wide range of tasks, including remote sensing and crop classification, owing to their ability to model complex relationships and learn from large volumes of data [23]. These networks typically consist of multiple layers, including input, hidden, and output layers, with each layer containing a set of interconnected nodes or neurons. The choice of network architecture, such as the number of layers and neurons, depends on the complexity of the data and the desired level of model performance. Aside from their versatility, NNs can also be fine-tuned to adapt to specific data types and distributions, enhancing their applicability across diverse domains. Neural networks for classification consist of fully connected layers, each applying a weight matrix, bias vector, and activation function. The final layer uses SoftMax activation for classification scores and predicted labels. We left the following hyperparameters to be fine-tuned during the optimization process: the number of fully connected layers, layer sizes, activation functions, and regularization strength.

2.3.4. Analyzing Performance and Accuracy Assessment

After the training phase (Section 2.3.3), the models were directly applied to the entire image (cropland-masked) using the selected hyperparameters per each case (scenario/year) following a pixel-based approach. No additional transfer learning or tuning was employed in this step. Subsequently, each classification output was cross-validated on tested cases, which represent randomly selected pixels from left-aside parcels per each crop type/year, as defined in Section 2.3.2. This validation step allowed us to assess the accuracy and reliability of the classification results. We used overall accuracy and F1 statistics as standard measures of accuracy and classification performance to comprehensively evaluate and compare various classification tasks [24]. Overall accuracy is the proportion of correctly classified instances out of the total instances, while F1 score is the harmonic mean of the

user's accuracy and the producer's accuracy, providing a balance between these two metrics for assessing class-specific classification performance. Interpreting ML classification models can be challenging due to their complexity. We employed feature importance analysis to facilitate a better understanding of the model's decision-making process and the role of specific features in predicting crop classifications, ultimately enhancing the overall interpretability of the results. Specifically, we used permutation-based feature importance analysis [25], a technique that evaluates the importance of each feature by randomly shuffling its values and measuring the impact on the model's performance. Furthermore, a specialized analysis was conducted to investigate the within-season classification performance. In this instance, monthly composites were incrementally added as input features to reflect the progression of the season. For example, April spectral reflectances were first considered, followed by April and May, and then April, May, and June, and so on. This particular approach enables the identification of the earliest time point at which a general classifier can produce satisfactory results, which is a standard approach across the community [12].

2.3.5. Applying Models to Images

All four classification models from Scenario 1 were applied to the Sentinel-2 monthly composite series from 2022 to exemplify their practical application in crop mapping. Pixel-based estimates were assigned to LPIS (land parcel identification system) parcels using a majority rule approach, effectively regionalizing the classification results for more coherent crop mapping. Additionally, a confidence map was generated to visualize the uncertainty of the classification map, which was based on an ensemble of model outputs at the decision level [26]. This was achieved by superimposing the four alternative outputs and identifying the frequency of agreement among the models. We established three confidence levels for parcel classification. The lowest confidence level indicates that only one model accurately classified the parcel, the medium confidence level signifies that two or three models successfully performed, and the highest confidence level is reached when all models correctly classified the parcel. In addition to the standard accuracy evaluation at an aggregated level using error matrices (Section 2.3.4), the confidence map enables a visual plausibility check to assess the spatial pattern of misclassifications. We paid particular attention to detecting potential spatial trends in errors, analyzing whether misclassifications demonstrated consistent geographic tendencies or associations with specific types of terrain or sub-regions. This allowed us to assess the spatial representativeness of our sampling and determine whether certain subregions were consistently linked to misclassification.

3. Results

3.1. Observational Quality

Monthly median composites of the spectral reflectance products may be less effective when there are fewer than 3 observations per monthly compositing period. Hence, we begin by analyzing the quality of the input data to gain insights that will inform the interpretation of our results later on. Figure 4 illustrates the variability in observation quality of the input monthly composites across different years, seasons, and regions. Generally, summer observations are of higher quality than spring observations due to differences in cloud cover and atmospheric conditions. Figure 4 also emphasizes the distinct regional differences observed during specific months, such as May 2021, June 2020, and May 2019. These differences may provide valuable insights for consideration when evaluating the spatiotemporal transferability scenarios of crop classification models. Moreover, balancing trade-offs between observation quality and full coverage is essential regarding the delivery of wall-to-wall crop mapping products.

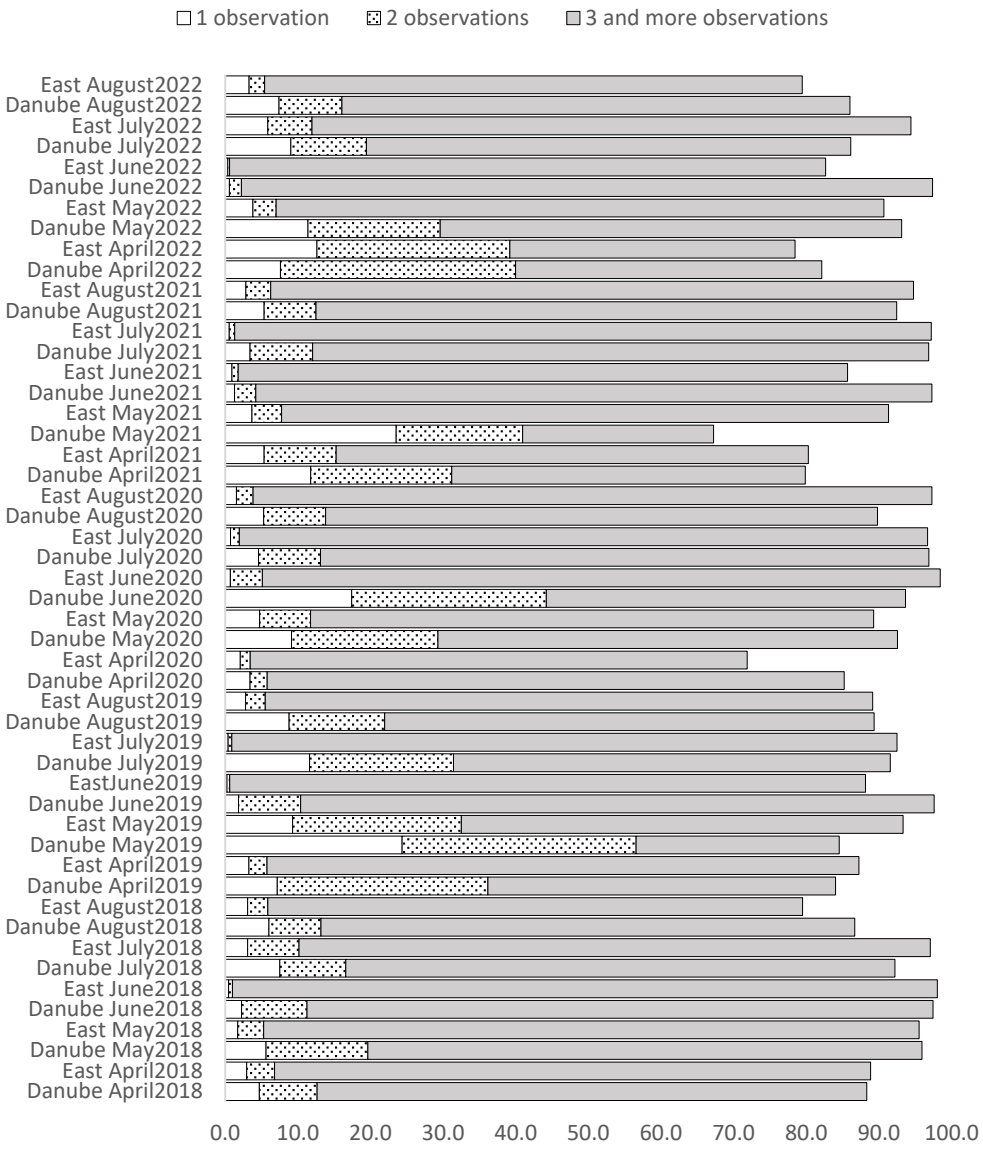

Figure 4. Observational quality of input monthly median composites. Number of observations used for median compositing and proportion of total coverage (% on *x*-axis) of the study regions.

3.2. Crop Specific Spectral Response across Regions and Seasons

Figures 5–8 display the mean spectral profiles of crops for the entire training dataset. These figures offer a visual representation of the unique spectral characteristics of each crop type and their interseasonal variations, helping to better understand crop-specific spectro-temporal patterns. Cereals, rapeseed, and grass display distinct spectral differences in the red edge (RE) and near-infrared (NIR) ranges from spring until July, when they are harvested. After the harvest, the spectral patterns in the NIR range tend to become noisy. However, the grass spectral response in the visible spectral range (VIS) maintains a more consistent profile after the harvest, which may contribute to its good classification performance. This observation holds true and remains consistent across both regions under study. Rapeseed displays the most distinctive spectral profile, which remains consistent across regions and seasons. The higher interseasonal variation in the NIR spectral range in April and June may be attributed to the varied coverage of young canopies after winter (particularly noticeable in eastern Slovakia) and differing maturation stages in June. Conversely, wheat and barley exhibit relatively similar spectrotemporal patterns. April appears to be the optimal period for distinguishing between them, especially in the RE and NIR spectral ranges in both regions. However, noticeable differences can also be observed

in the VIS spectral range in eastern Slovakia. This characteristic can be attributed to a higher proportion of delayed sowing of barley in eastern Slovakia, which may be linked to prolonged unfavorable conditions such as waterlogging. This observation may also contribute to the higher interseasonal variability of barley's spectral response during its early development in May in eastern Slovakia.

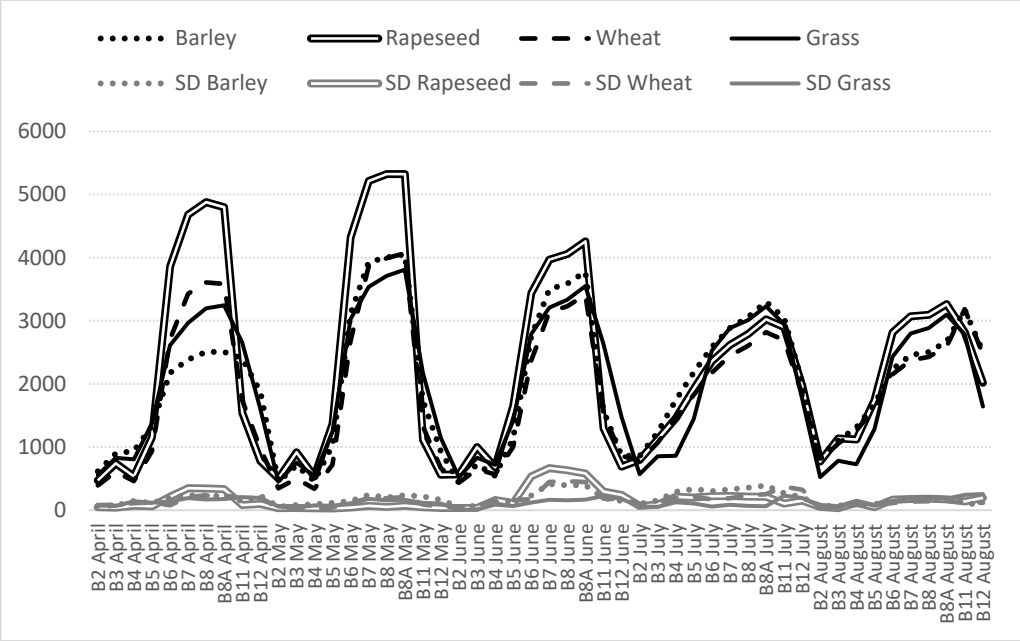

Figure 5. Long-term (2018–2022) means and Standard Deviations (SD) of spectral bands for Danubian lowland cereal crops, rapeseed, and grasslands. *y*-axis represents reflectance (0–1) values multiplied by 10,000. X axis represents Sentinel-2 spectral band for specific month.

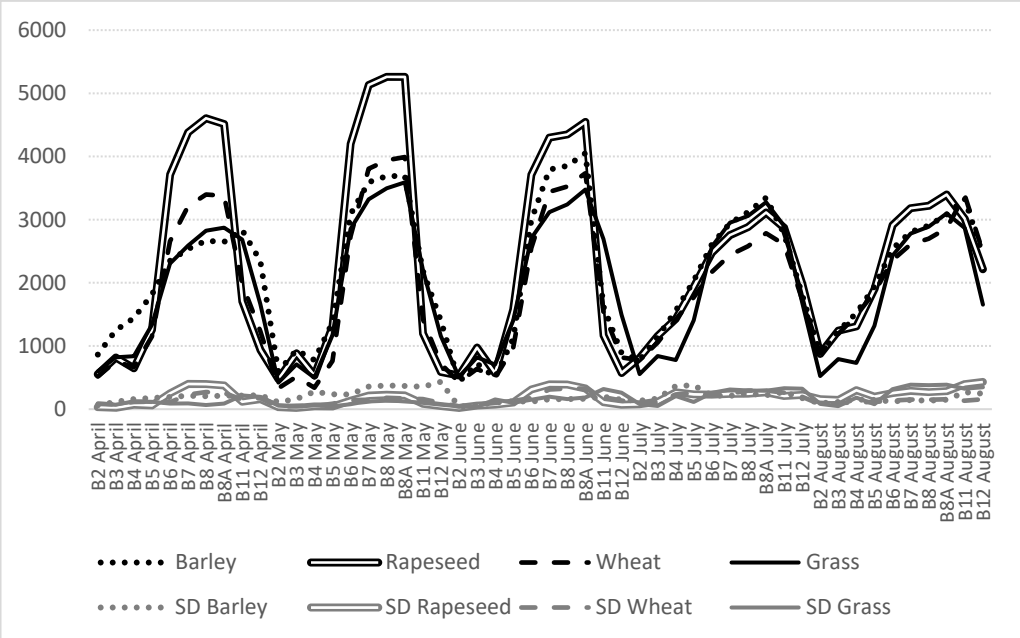

Figure 6. Long-term (2018–2022) means and standard deviations (SD) of spectral bands for East Slovakian lowland cereal crops, rapeseed, and grasslands. *y*-axis represents reflectance (0–1) values multiplied by 10,000. X axis represents Sentinel-2 spectral band for specific month.

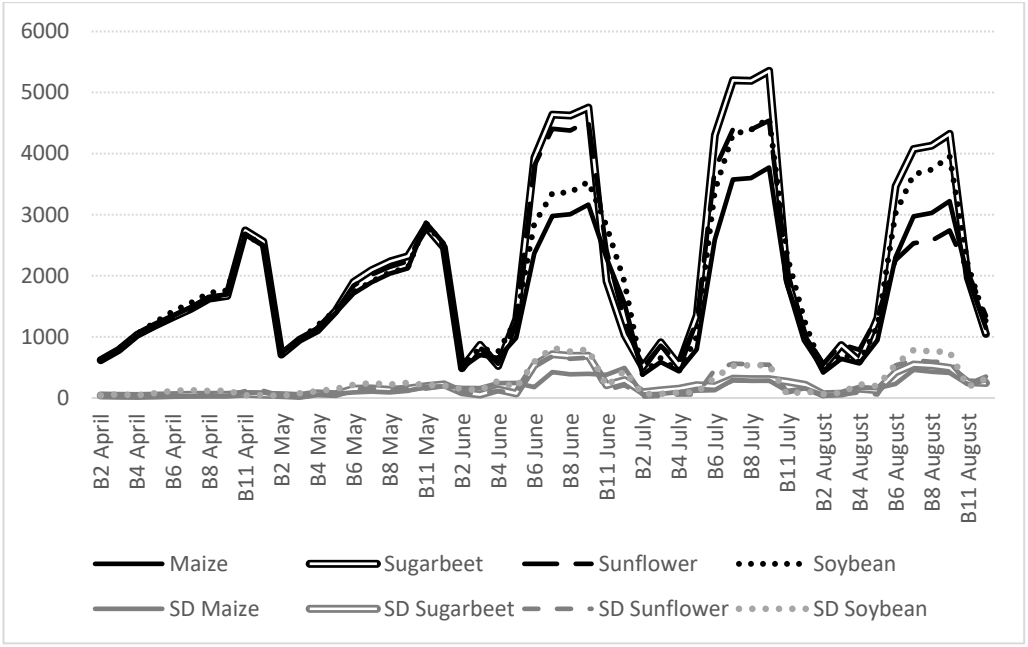

Figure 7. Long-term (2018–2022) means and Standard Deviations (SD) of spectral bands for Danubian lowland summer crops. *y*-axis represents reflectance (0–1) values multiplied by 10,000. X axis represents Sentinel-2 spectral band for specific month.

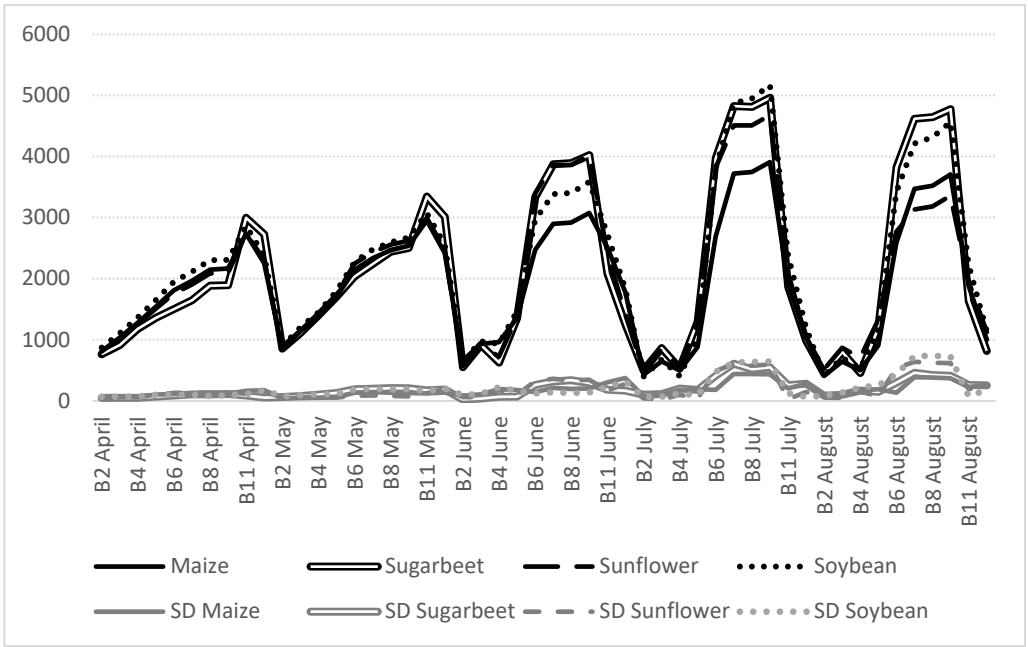

Figure 8. Long-term (2018–2022) means and Standard Deviations (SD) of spectral bands for East Slovakian lowland summer crops. *y*-axis represents reflectance (0–1) values multiplied by 10,000. X axis represents Sentinel-2 spectral band for specific month.

On the other hand, summer crops exhibit distinct spectral differences in the red-edge to NIR range during the summer months, starting in June. In fact, they are typically sown from mid-April to mid-May, during which time they appear as bare soil, making their differentiation impossible at that stage. Maize stands out as the most distinctive summer crop, exhibiting consistently lower reflectance in the red-edge to NIR spectral range during June and July in both regions. However, notable interseasonal variation can be observed in July in eastern Slovakia. Sugar beet appears to be most distinctive in the red-edge to NIR spectral range in the Danubian lowlands, although notable inter-

seasonal variation is visible in June (Danubian lowlands). Soybean displays particularly different spectrotemporal patterns across regions, exhibiting significantly higher reflectance in the red edge to NIR region in eastern Slovakia compared to the Danubian lowlands. Nevertheless, soybean also exhibits considerable interseasonal variation, which might reduce its distinguishable characteristics across years.

Although distinctive characteristics can be discerned from the spectral mean data, the complex distribution patterns might further affect the transferability of classification models. For instance, we illustrate the spatiotemporal distribution patterns of the highly variable soybean (Figure 9) and the more consistent grasslands (Figure 10), highlighting the differences in their respective classification performances. In our study, we utilized Sentinel 2 spectral band 6, as it was identified as one of the most important features for classification models. We observed notable differences in its distribution both across regions and seasons. It seems that in 2018 and 2019, certain factors, such as drought or later sowing dates, could have contributed to lower reflectance values in this spectral region compared to East Slovakia, where this effect was not observed. These distinct distribution patterns may affect the spatiotemporal transferability of the classification models. In contrast, we noticed a relatively consistent spectrotemporal pattern in the distribution of spectral band 6 for grasslands. This observation could contribute to the relatively good transferability of classification models across seasons and regions when classifying grasslands. In any case, we did not identify any coherent patterns related to observational quality (Figure 4), suggesting that factors other than observational quality may play a role in influencing the difference in spectrotemporal distributions across regions.

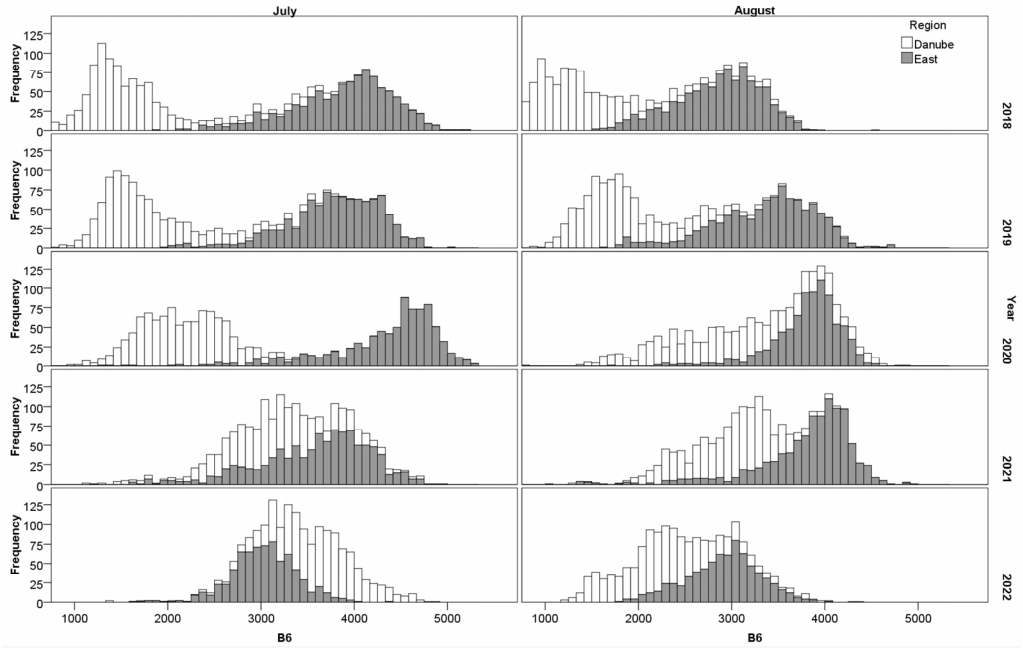

Figure 9. Variability in the spectral band B6 distributions of soybeans during July and August across different years and regions. *x*-axis represents reflectance (0–1) values multiplied by 10,000.

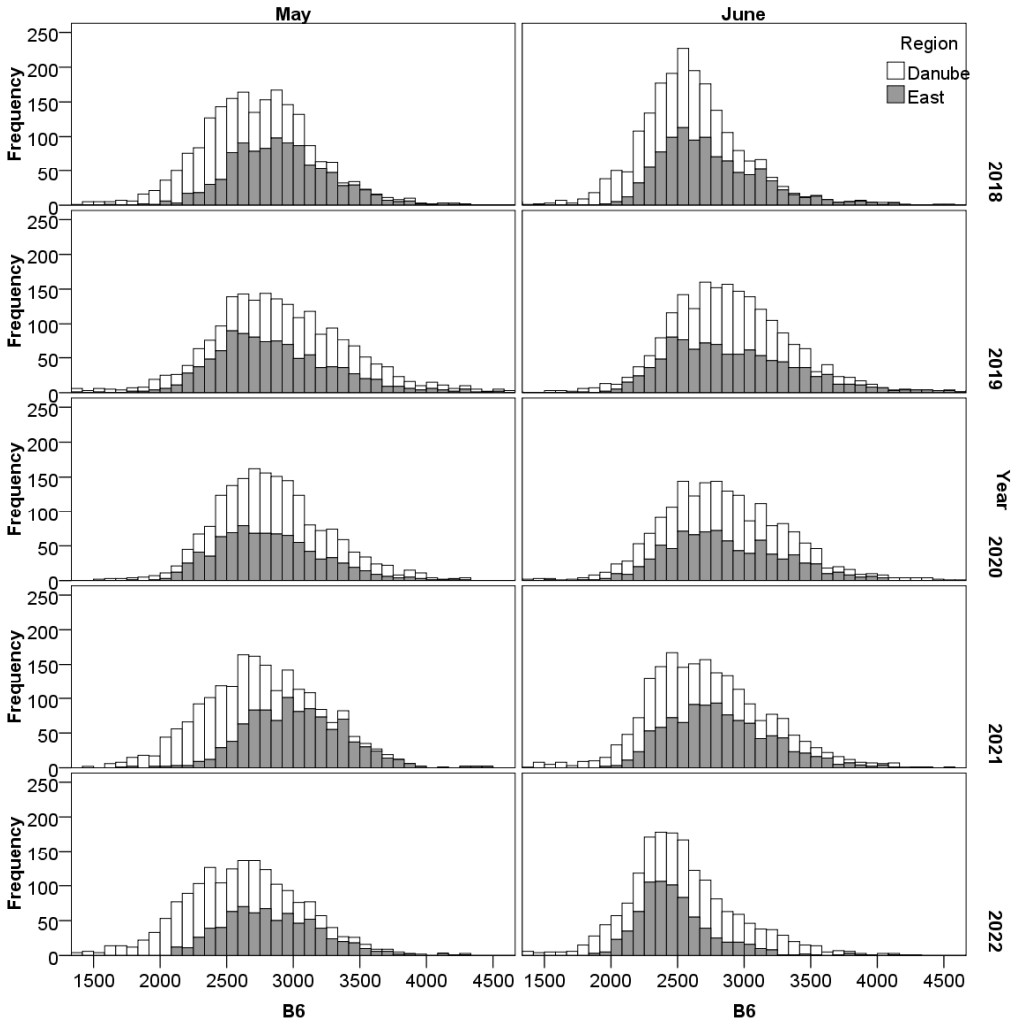

Figure 10. Variability in the spectral band B6 distributions of grasslands during July and August across different years and regions. *x*-axis represents reflectance values (0–1) multiplied by 10,000.

3.3. Temporal Transferability

Scenarios 1 and 2 focus on examining the transferability capabilities of the classification models. The findings from these scenarios reveal that the transferability of classification models across years differs between regions (Figure 11). However, the interseasonal patterns in both regions were quite similar, with the only exception being in 2018, where a notably better performance was observed in the Danubian lowlands compared to the eastern Slovakia region. Again, we did not identify any relationship to the observational quality (Figure 3), suggesting that factors other than the observational quality may play a role in influencing the difference in across year classification performance. The Danubian lowlands demonstrate better performance, with the best-performing classifier achieving overall accuracies ranging from 91.53% in 2022 to 94.3% (Table 4). In eastern Slovakia, the best-performing classification model demonstrates a range of overall accuracies from 85% to 91.9% (Table 5). Regardless, the interseasonal variability in performance for the worst-case models is proportionally higher, which highlights the importance of conducting a multi-classifier analysis.

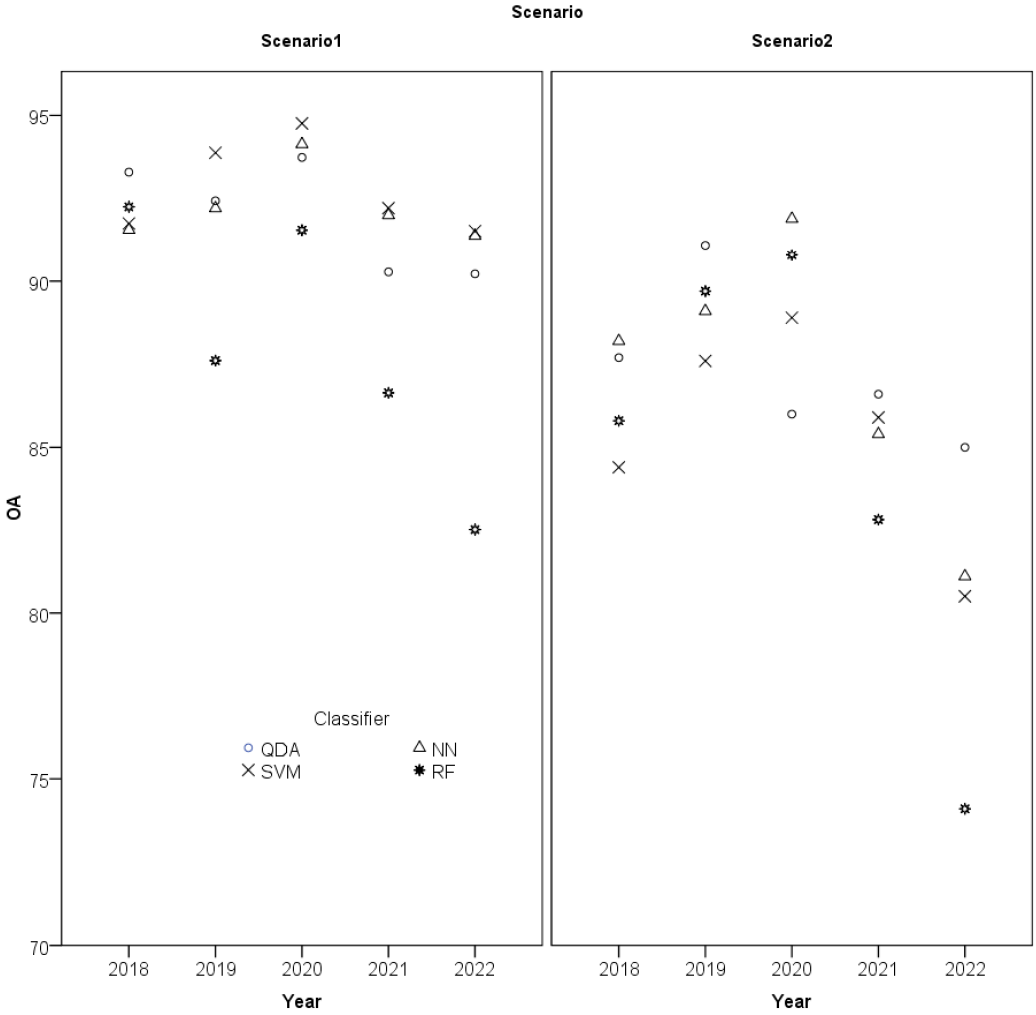

Figure 11. Overall accuracies (OA in % on Y Axis) across years: Scenario 1 represents the temporal transferability analysis in the Danubian lowland, while Scenario 2 pertains to the East Slovakian lowland region.

Table 4. Overall accuracies (%) across years of the Scenario 1. Scenario 1 represents the temporal transferability analysis in the Danubian lowland region.

Scenario 1	Train Danube without 2018	Train Danube without 2019	Train Danube without 2020	Train Danube without 2021	Train Danube without 2022	Mean
	Test Danube 2018	Test Danube 2019	Test Danube 2020	Test Danube 2021	Test Danube 2022	
QDA	93.30	92.44	93.74	90.28	90.22	92.00
SVM	91.76	93.88	94.76	92.22	91.53	92.83
NN	91.57	92.22	94.14	92.01	91.39	92.27
RF	92.26	87.61	91.56	86.64	82.50	88.11
Mean	92.22	91.54	93.55	90.29	88.91	

Table 5. Overall accuracies (%) across years of the Scenario 2. Scenario 2 represents the temporal transferability analysis in the East Slovakian lowland region.

Scenario 2	Train East without 2018	Train East without 2019	Train East without 2020	Train East without 2021	Train East without 2022	Mean
	Test East 2018	Test East 2019	Test East 2020	Test East 2021	Test East 2022	
QDA	87.70	91.10	86.00	86.60	85.00	87.28
SVM	84.40	87.60	88.90	85.90	80.50	85.40
NN	88.20	89.10	91.90	85.40	81.10	87.14
RF	85.80	89.70	90.80	82.80	74.10	84.64
Mean	86.53	89.38	89.40	85.18	80.18	

The performance of the worst-case models ranged from 74.10% to 86.0% in Eastern Slovakia and from 82.5% to 91.6% in the Danubian lowlands. Interestingly, except for 2018 in the Danubian lowlands and 2018, 2019, and 2020 in eastern Slovakia, the random forest classifier performed substantially worse compared to the other algorithms. In general, in the Danubian lowlands, all classifiers except random forest performed equally well. On the other hand, in eastern Slovakia, neural networks and quadratic discriminant analysis consistently outperformed support vector machines and random forest classifiers.

In the Danubian lowlands, class-specific performance showed that rapeseed, grasslands, and sugar beet were consistently well-classified across seasons, indicating stable transferability for these crop types (Figure 12). Other summer crops were also classified at acceptable levels, with the exception of 2022, where classification performance was relatively lower, notably for soybeans and maize. We attribute this lower performance to the extensive drought that occurred in 2022. There is no reason to suspect the potential negative effect of unsuitable input data, considering that there were sufficient observations for the summer monthly composites, as evidenced in Figure 3. Given that cereals consistently exhibited poorer performance in 2018, we investigated whether there were any patterns of unsuitable observation conditions (Figure 3). Since we did not identify any substantial divergence in the availability of sufficient observational data during the key months (April and May), we believe that this performance decline was likely due to reasons other than input data quality.

Crop-specific classification performance is greatly affected by the uniqueness of the spectrotemporal signatures. While certain patterns may be visible in the mean series (Figures 5–8), the spectrotemporal distribution pattern driving the classification is considerably more intricate and challenging to visualize, as demonstrated in the soybean (Figure 9) and grassland (Figure 10) cases. Ultimately, the separability of the multidimensional feature space can be inferred from the misclassification error matrices, which play a crucial role in understanding and evaluating the model's performance and its ability to distinguish between different crop types. As an example, we provide the error matrices for the worst-case scenario of 2022 (Table 6) and the best-case scenario of 2020 (Table 7) from Scenario 1, although the misclassification patterns are consistent across all scenarios (data not shown). Here, the evident commission errors of misclassifying soybean as maize and sunflower result in poorer F1 statistics for these classes (Table 6). In the case of the 2022 drought, there is a possibility that soybeans affected by the drought may exhibit spectral properties that are erroneously recognized as maize or sunflower, but this was not observed in 2020 (Table 7). On the other hand, in 2020, barley was classified erroneously as wheat, leading to lower F1 statistics for these cereal classes and an underestimation of barley as well as an overestimation of wheat.

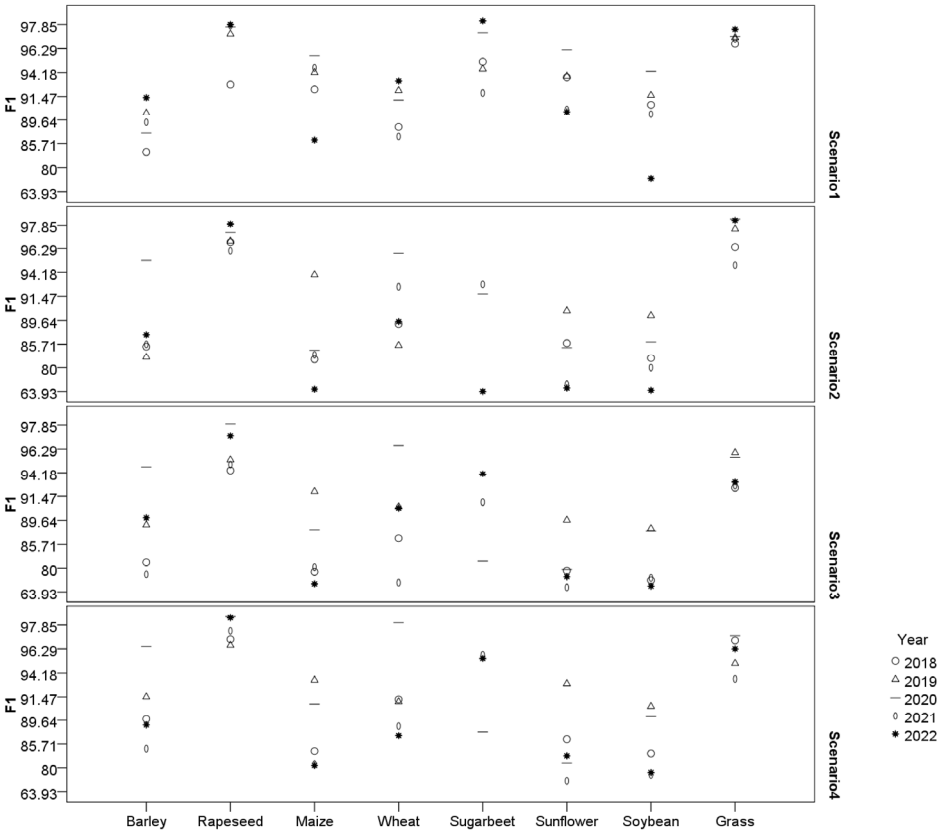

Figure 12. Crop-Specific Performance Using F1 Statistics (in % on Y Axis) for All Scenarios: Scenario 1 represents the temporal transferability analysis in the Danubian Lowland, while Scenario 2 focuses on the temporal transferability analysis in the East Slovakian Lowland region. Scenario 3 involves the spatial transferability scenario, using 1 year of data from the Danubian Lowland for training and the same year of data from the East Slovakian Lowland for testing. Scenario 4 encompasses all years (2018–2022) of data from the Danubian Lowland for training while testing each individual year of data from the East Slovakian Lowland separately. Tabulated data are provided in Table S2, Supplement S2.

Table 6. Error matrices of the worst-case scenario 2022 for Scenario 1. Scenario 1 represents the temporal transferability analysis in the Danubian Lowland.

		Prediction								
	SVM	**Barley**	**Rapeseed**	**Maize**	**Wheat**	**Sugar Beet**	**Sunflower**	**Soybean**	**Grass**	∑
Reference	Barley	975	2	2	18	1	0	0	2	1000
	Rapeseed	21	977	0	1	1	0	0	0	1000
	Maize	26	2	922	13	1	10	11	15	1000
	Wheat	80	1	0	913	0	1	0	5	1000
	Sugar beet	5	15	4	0	971	5	0	0	1000
	Sunflower	8	0	4	0	1	967	8	12	1000
	Soybean	16	0	193	18	1	160	603	9	1000
	Grass	3	0	1	0	2	0	0	994	1000
	∑	1134	997	1126	963	978	1143	622	1037	
	OA	91.5								
	KIA	0.90								

Table 7. Error matrices of the best case 2020 for Scenario 1. Scenario 1 represents the temporal transferability analysis in the Danubian Lowland.

		Prediction								
	SVM	**Barley**	**Rapeseed**	**Maize**	**Wheat**	**Sugar Beet**	**Sunflower**	**Soybean**	**Grass**	∑
Reference	Barley	794	25	9	142	6	1	13	10	1000
	Rapeseed	0	995	0	0	4	1	0	0	1000
	Maize	0	0	971	6	3	1	16	0	1000
	Wheat	9	10	3	976	1	0	1	0	1000
	Sugar beet	0	0	5	0	987	3	5	0	1000
	Sunflower	4	0	4	0	18	949	23	2	1000
	Soybean	2	0	25	0	6	13	952	2	1000
	Grass	3	6	5	15	2	4	8	957	1000
	∑	812	1036	1022	1139	1027	972	1021	971	
	OA	94.8								
	KIA	0.90								

3.4. Spatial Transferability

Scenario 3 focuses on examining the transferability capabilities of the classification models in a spatial context. The performance varied between particular years; e.g., the best-performing classification model demonstrates a range of overall accuracies from 84.9% in 2018 to 91.54% in 2019, while performance for the worst-case models ranged from 72.10% in 2022 to 88.75% in 2020 (Table 8), which highlights the importance of conducting a multi-classifier analysis. The performance of classifiers was quite similar, except for RF in 2022, 2021, and 2018, when RF achieved considerably lower overall accuracy values. Interestingly, when all seasons from the Danube region were used for training in Scenario 4, a consistently better performance except for 2021 was achieved compared to Scenario 3 (Figure 13). This improvement in classification performance was particularly noticeable in 2018 and 2022 (Table 9), further emphasizing the potential benefits of incorporating interseasonal variance in spectral data distributions when addressing spatial differences between regions. In this scenario, the crop-specific performance also showed a slight improvement, particularly for rapeseed, wheat, and sugar beet, as illustrated in Figure 12. Conversely, crops such as barley, maize, sunflower, and soybean did not experience substantial benefits from this strategy (Figure 12). In any case, similar to scenarios 1 and 2 (temporal transferability), it was found that rapeseed, grasslands, and sugar beet (with the exception of 2020) can be classified across regions at acceptable levels. The substantial differences in spectral data distribution, as illustrated in Figure 9 for soybean, may contribute to the highest misclassification rates observed among soybean, maize, and sunflower. Specifically, the classification model tended to overestimate soybean at the expense of maize and sunflower in 2021 (Table 10). This discrepancy could be attributed to the distinct development of soybean in Eastern Slovakia during 2021, which may have brought its spectrotemporal profile closer to those of maize and sunflower, thereby affecting the typical differences in their spectrotemporal signatures as these high commission error rates were not observed in other years (see Table 9 as an example). Moreover, similar to the temporal transferability scenarios, wheat and barley were misclassified only between each other. Therefore, if an aggregated "cereals" class were assigned, the transferability of the classification models would perform well for cereals both across seasons and regions. Sugar beet generally exhibited distinct spectrotemporal signatures, with the exception of 2020 (Table 11), when sugar beet was overestimated at the expense of sunflower. This was not the case in other years or in temporal transferability scenarios 1 and 2.

Table 8. Overall accuracies (%) across years of the Scenario 3. Scenario 3 explores the spatial transferability across regions. Classifiers are trained and tested using the same years but in different regions.

Scenario 3	Train Danube 2018	Train Danube 2019	Train Danube 2020	Train Danube 2021	Train Danube 2022	Mean
	Test East 2018	Test East 2019	Test East 2020	Test East 2021	Test East 2022	
QDA	81.10	88.60	88.75	87.35	81.34	85.43
SVM	84.90	88.66	90.30	87.41	85.29	87.31
NN	82.16	91.54	90.27	87.34	83.80	87.02
RF	76.90	89.29	89.15	80.65	72.10	81.62
Mean	81.27	89.52	89.62	85.69	80.63	

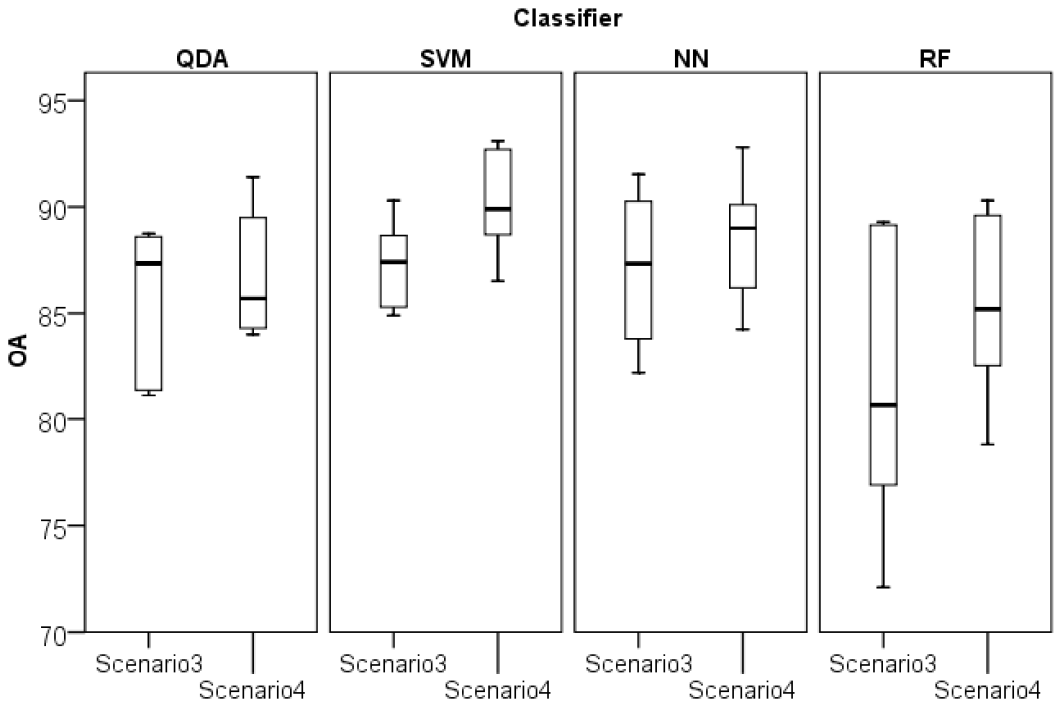

Figure 13. Classification performance on different scenarios.

Table 9. Overall accuracies (%) across years of the Scenario 4. Scenario 4 explores the spatial transferability across regions. Scenario 4 includes all years (2018–2022) of data from the Danubian lowland for training, while testing on each individual year of data from the East Slovakian lowland separately.

Scenario 4	Train Danube 2018–2022	Train Danube 2018–2022	Train Danube 2018–2022	Train Danube 2018–2022	Train Danube 2018–2022	Mean
	Test East 2018	Test East 2019	Test East 2020	Test East 2021	Test East 2022	
QDA	85.70	91.40	89.50	84.00	84.30	86.98
SVM	89.90	93.10	92.70	86.52	88.70	90.18
NN	90.10	92.80	89.00	84.24	86.20	88.47
RF	85.20	90.30	89.60	82.50	78.80	85.28
Mean	87.73	91.90	90.20	84.32	84.50	

Table 10. Error matrices of the worst case 2021 for Scenario 4. Scenario 4 encompasses all years (2018–2022) of data from the Danubian Lowland for training, while testing on each individual year of data from the East Slovakian Lowland separately.

		Prediction								
	SVM	**Barley**	**Rapeseed**	**Maize**	**Wheat**	**Sugar Beet**	**Sunflower**	**Soybean**	**Grass**	∑
Reference	Barley	889	0	24	0	2	15	10	60	1000
	Rapeseed	16	956	2	1	0	17	4	4	1000
	Maize	7	0	748	0	9	24	200	12	1000
	Wheat	172	2	2	791	1	4	4	24	1000
	Sugar beet	0	0	0	0	1000	0	0	0	1000
	Sunflower	1	0	19	0	66	616	266	32	1000
	Soybean	11	3	40	1	3	9	928	5	1000
	Grass	0	0	0	0	0	0	6	994	1000
	∑	1096	961	835	793	1081	685	1418	1131	
	OA	86.53								
	KIA	0.85								

Table 11. Error matrices of the best case 2020 for Scenario 4. Scenario 4 encompasses all years (2018–2022) of data from the Danubian lowland for training, while testing on each individual year of data from the East Slovakian lowland separately.

		Prediction								
	SVM	**Barley**	**Rapeseed**	**Maize**	**Wheat**	**Sugar Beet**	**Sunflower**	**Soybean**	**Grass**	∑
Reference	Barley	832	2	7	2	3	1	2	0	1000
	Rapeseed	12	836	0	0	0	0	0	1	1000
	Maize	4	2	735	0	41	26	38	3	1000
	Wheat	11	2	0	832	0	0	1	3	1000
	Sugar beet	0	0	0	0	849	0	0	0	1000
	Sunflower	4	0	10	8	154	606	50	17	1000
	Soybean	6	0	11	4	39	3	765	21	1000
	Grass	1	0	2	0	2	0	0	844	1000
	∑	870	842	765	846	1088	636	856	889	
	OA	92.74								
	KIA	0.92								

3.5. Feature Importance

Analyzing feature importance aids in the interpretability of machine learning models, leading to a better understanding of the unseen factors that drive the classifiers. By shedding light on the key features that contribute to distinguishing crop classes, we can gain insights into the underlying spectrotemporal characteristics that are crucial for accurate classification. To maintain focus on the most obvious scenario, we have chosen to analyze feature importance exclusively for scenario 1 for the last year of 2022. This scenario represents a typical crop-type mapping situation within a region of interest, using past labeled data to classify the upcoming year with unseen reference data. Figure 14 displays the spectrotemporal input features that were most frequently ranked among the top 10 for each classifier. In summary, spectral band 6 from May was ranked as the most useful, followed by band 8A from August and band 6 from July. Upon examining the spectral characteristics of the features, it was established that the most useful ones were primarily the narrowband red edge (B6 and B7) and near-infrared (NIR) band B8A, followed by the B11 short-wave infrared (SWIR) band. Importantly, the broadband 8B and visible (VIS) bands were not considered important in the classification process. Another interesting observation from our analysis was that all periods played a crucial role in the classification process, with July emerging as the most important and August as the least important

among them. Additionally, we have examined the earliest period in a growing season when reliable crop classification is achievable, enabling prompt within-season crop mapping. Figure 15 demonstrates the trade-off between timeliness and accuracy for all classifiers, highlighting the balance between obtaining results earlier in the season and maintaining classification performance. Interestingly, this trade-off varied slightly among classifiers, with noticeable problems in random forest (RF) decreasing when more drought-affected features were incorporated later in the season. In any case, our findings showed that acceptable performance could be achieved as early as June, with only slight improvements towards the end of the season. However, this was the case for an anomalous drought year, which might degrade the added value of the months with unusual spectrotemporal responses from the summer crops. For comparison, we conducted additional analysis using data from the "standard" year 2020 and observed a more expected pattern, with a consistent increase in performance towards the end of the season, reaching saturation in July (Figure 16).

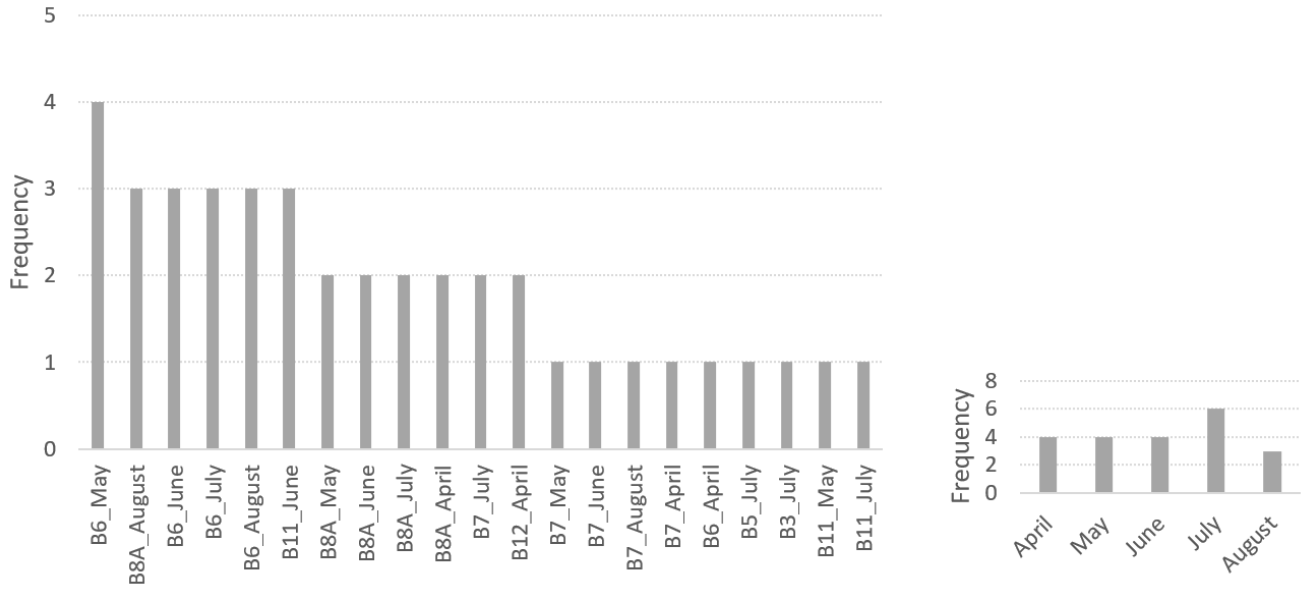

Figure 14. Frequency at which specific features are identified among the top 10 most important features in all considered classifiers for Scenario 1 and 2022 tested year (A) and their aggregation according to spectral bands (B) and monthly period (C).

3.6. Confidence Map

Applying classification models to satellite products enables the creation of spatially comprehensive crop maps, providing a detailed representation of crop distribution across study regions. Firstly, we employed classification models on input monthly composites using a pixel-based approach. Subsequently, we utilized known parcel boundaries to aggregate pixel-based outputs according to declared parcel boundaries, employing a majority rule method. In this study, we showcase the worst-case example from scenario 1 in 2022 to provide a conservative perspective on the mapping results, allowing for a more cautious assessment of the model's performance and potential limitations. Clearly, the mapping coverage is constrained by the quality of observations throughout the entire season. In the case of 2022, coverage reached approximately 80% of the total arable land in both regions. The classification maps generated in our study appeared to be consistent with expected spatial patterns of crop distribution (Figure 17). There were no implausible spatial trends observed, indicating that the classification models were able to effectively capture and represent the spatial distribution of the different crop types in the study area. Upon closer examination of Figure 18, it is evident that there are no noticeable erroneous spatial clusters present. This further supports the effectiveness of the classification models in accurately

capturing the spatial distribution of crop types without introducing spatial errors or biases in the input data. However, we need to notice several factors contributing to crop misclassification, and understanding these factors is essential for improving the transferability of crop classification models. Some of these factors include natural drivers, such as crop phenology, and climate extremes such as drought, heatwaves, or prolonged rain causing waterlogging, which can impact interseasonal classification accuracy. Within-class variability due to genetic differences among crop varieties, including their variable response to drought, can result in spectral variability, making it challenging to distinguish between similar crops. This variability can be further exacerbated by irrigation infrastructure and patterns, introducing additional complexity to classification efforts, particularly for summer crops. Lastly, variations in agronomic practices, like sowing and harvest dates, or factors such as pests, diseases, or other disturbances can alter the spectral properties of crops, potentially leading to misclassification. These factors introduce uncertainties throughout the whole classification workflow. We have tried to address these uncertainties through the methodology we implemented, which includes balanced spatial sampling, Bayesian optimization for hyperparameter selection, and an ensemble of multiple models. In particular, the use of a multi-classifier approach enables the determination of parcel-level confidence in mapping. By assuming that higher confidence arises when multiple classifiers provide similar results, this approach enhances the reliability of the crop distribution maps. This practical solution is particularly suited for decision-makers in operational applications. In any case, a comprehensive uncertainty assessment would be beneficial; however, this would require a specific experimental design for each source of uncertainty and extend beyond the scope of this study.

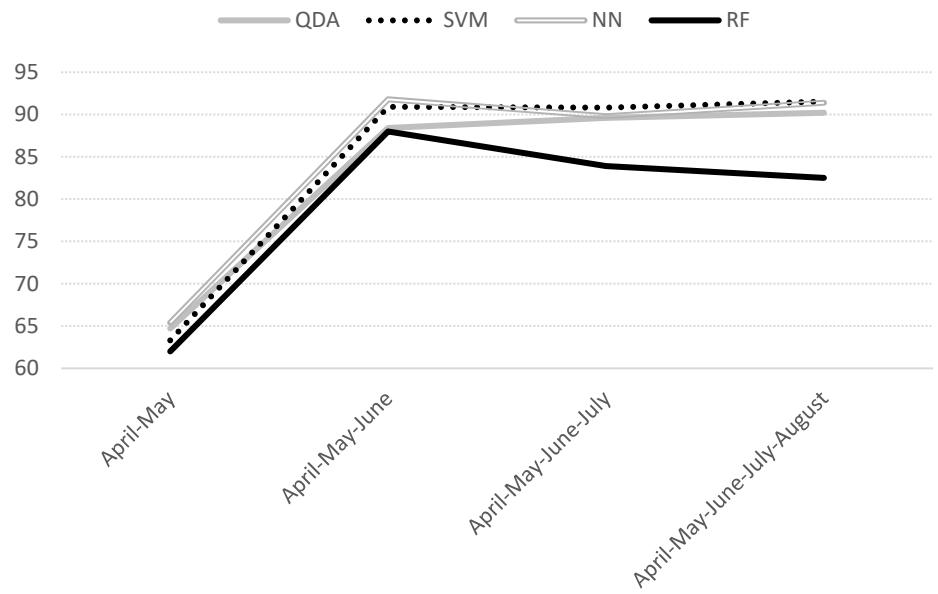

Figure 15. Evolution of overall accuracies (OA in % on Y Axis) with sequential increases in input features within the season 2022.

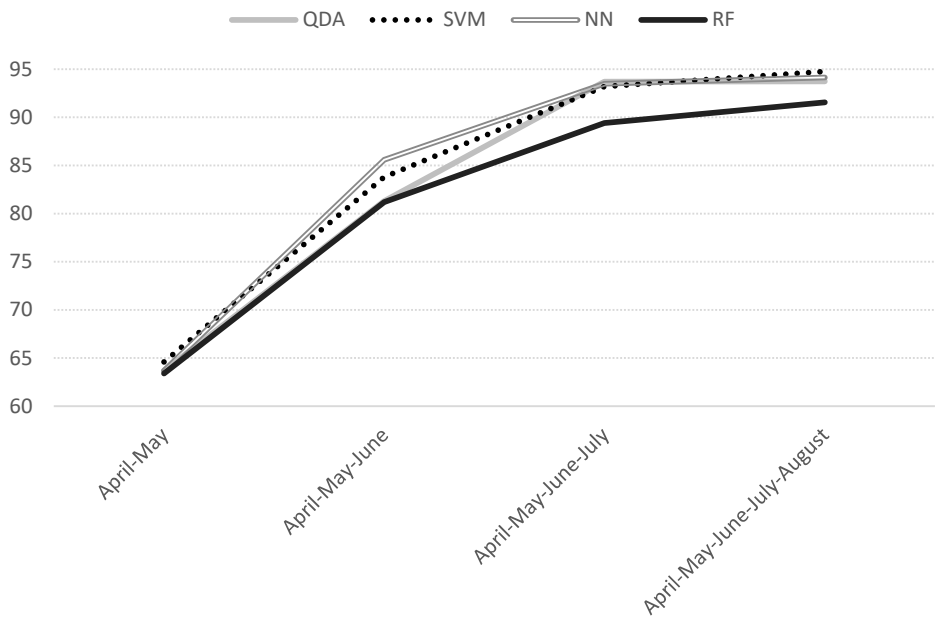

Figure 16. Evolution of overall accuracies (OA in % on Y Axis) with sequential increases in input features within the season 2020.

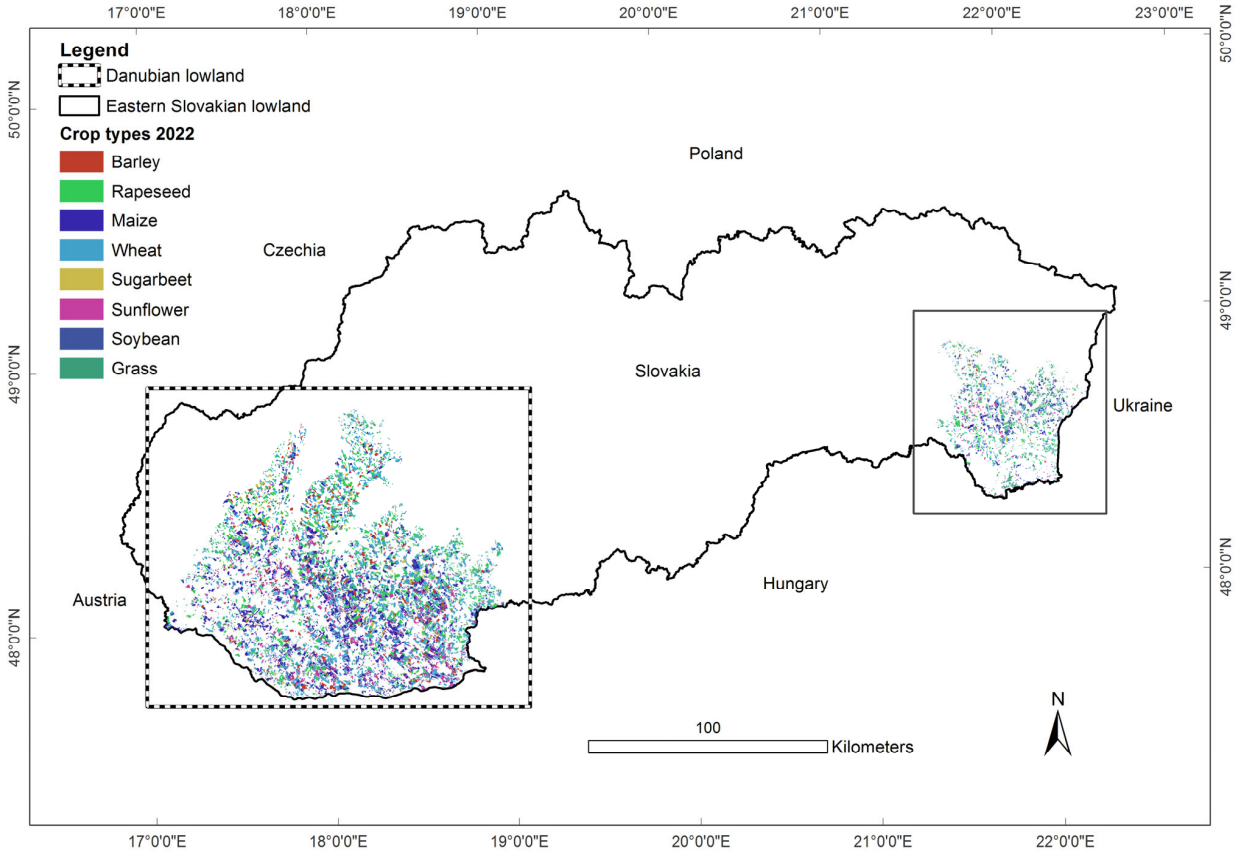

Figure 17. Spatial distribution of crops for 2022 in Danubian lowlands and Eastern Slovakia. The classification model for Danubian lowlands followed the Scenario 1, e.g., the train data included reference crop labels in Danubian region spanning years 2018, 2019, 2020, and 2021. The classification model for eastern Slovakia followed Scenario 2, e.g., the train data included reference crop labels in eastern Slovakia's lowlands spanning years 2018, 2019, 2020, and 2021. Hence, year 2022, with prolonged summer drought, was not used for the training.

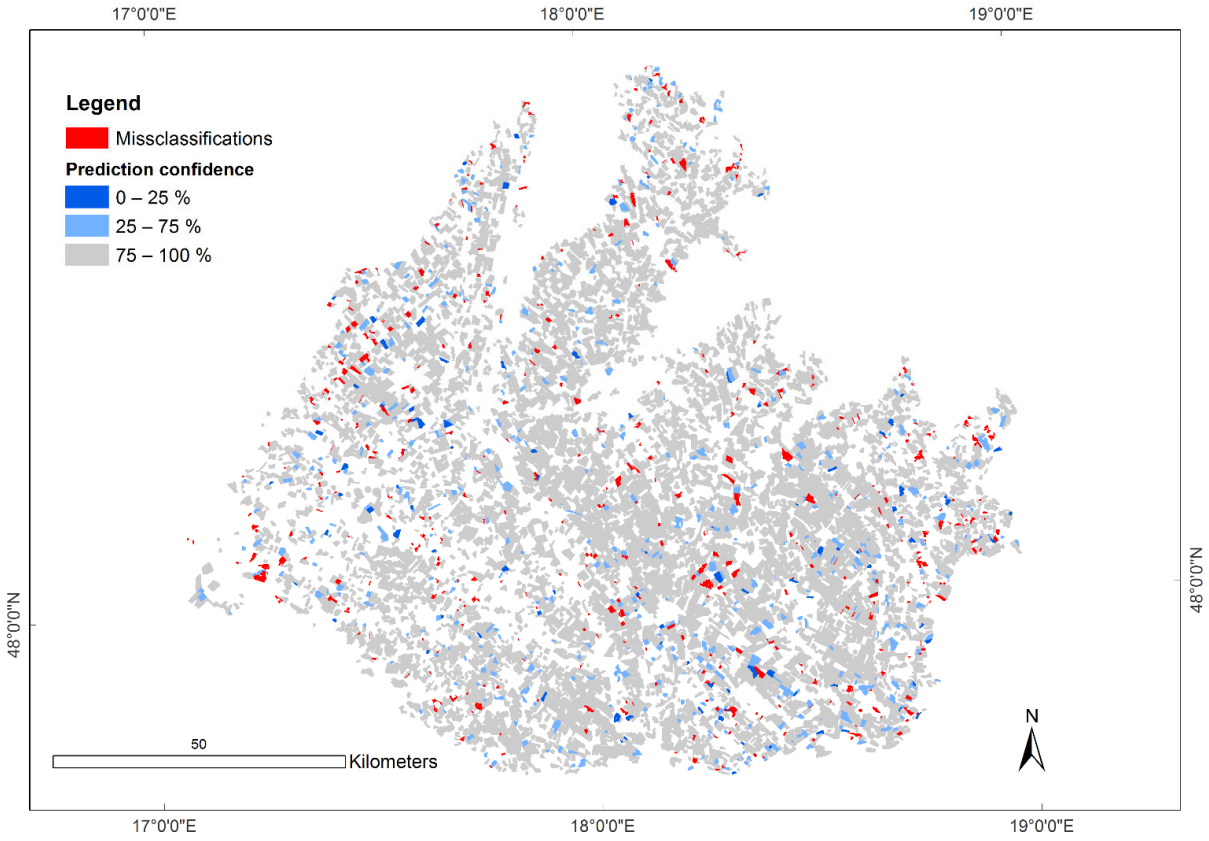

Figure 18. Spatial distribution of misclassification errors and parcel-based confidence levels for Scenario 1 in tested year 2022 in Danubian lowland region. The confidence levels range from lowest to highest, where the lowest confidence level indicates that only one model correctly classified the parcel, the medium confidence level indicates that two or three models correctly classified the parcel, and the highest confidence level indicates that all models correctly classified the parcel. The misclassification errors are displayed in parcels where none of the models correctly classified the given parcel. Spatial distribution of misclassification errors and parcel-based confidence levels for Scenario 1 in tested year 2022 in Danubian lowland region. The confidence levels range from lowest to highest, where the lowest confidence level indicates that only one model correctly classified the parcel, the medium confidence level indicates that two or three models correctly classified the parcel, and the highest confidence level indicates that all models correctly classified the parcel. The misclassification errors are displayed in parcels where none of the models correctly classified the given parcel.

4. Discussion

4.1. Spatiotemporal Generalization

The concept of spatiotemporal generalization emphasizes the need to balance both spatial and temporal aspects when developing models to ensure they can effectively handle variations in both dimensions. By carefully considering these factors, we can gain a deeper understanding of the limitations and opportunities for enhancing the generalization capabilities of crop classification models. Sykas et al. [27] conducted a study that concluded that the transfer of crop classification models across years or regions was not feasible. They trained their models using data from only two years. They then applied these models to large regions of France and Catalonia. These regions had distinct climates, agricultural practices, and crop growth patterns. The same issue might arise even in smaller regions. Therefore, a detailed investigation is essential to understand the generalization capabilities. These capabilities, associated with the spatiotemporal variability of domain distributions, are crucial for designing effective nationwide EO-based crop monitoring applications. We addressed this by training our models using data from multiple years and different regions. This approach may guarantee the inclusion of diverse spectrotemporal responses from

different crops. Our aim is to achieve the required generalization of the spectrotemporal feature space. We employed a simple empirical approach using accuracy measures as indicators of the classification models' ability to transfer in both spatial and temporal contexts. Although it is challenging to directly compare different studies, Table 1 presents accuracy measures from similar research to provide a rough benchmark for performance in crop classification transferability studies. Clearly, performance statistics differ substantially based on factors like scale, crop complexity, and the datasets employed, with performance estimates from our study aligning well with the identified benchmark of 90%. Accordingly, our results can be considered promising for the crop classification task, as they demonstrate the potential to effectively classify unseen labels from different regions and years. This suggests that our approach may offer valuable insights for nationwide agricultural monitoring. However, it is important to discuss certain considerations and limitations in more detail to provide guidance for applying similar workflows elsewhere.

4.2. Observation Quality Consideration

Firstly, the quality of the input temporal composites from Sentinel-2 data might cause a potential degradation in transferability performance. Factors such as cloud cover, atmospheric interference, and limited clear observations could impact the accuracy of median reflectance. This, in turn, could affect machine learning model performance across various regions or time periods. We used monthly median composites that appeared to strike an effective balance between monitoring unique crop phenology (Figures 5–8) and ensuring the availability of wall-to-wall data (Figure 4). However, median compositing is mainly effective when there are more than three observations per compositing period. With a higher number of observations, median composites can better represent the central tendency of the data, minimizing the influence of noise, clouds, and other disturbances. We made the simple assumption that the median composites could be degraded when fewer than three observations per pixel were used. However, we did not identify any coherent relationship between the patterns of observational quality and the year-to-year classification performance. For instance, June 2020 in the Danube region had one of the poorest observational qualities, yet the performance of Scenario 1 in 2020 was the best. This suggests that other factors may play a more significant role in influencing classification performance. Similarly, in Scenario 3, May 2019 for both regions exhibited substantially lower observational quality. However, the spatial transfer of 2019 still achieved acceptable performance. This suggests that factors other than the quality of median compositing may influence the spatiotemporal capabilities of the classification models. While median compositing has proven effective in similar geographical locations [28], it is worth considering alternative preprocessing methods. In future research, we should ensure that the most suitable approach is employed for preparing input features in transferable crop classification studies. For instance, other candidate methods exist for pixel based temporal compositing of Sentinel like time series, including medoid compositing [29], weighted score compositing [18], or phenology-adaptive compositing [30]. Furthermore, an alternative approach for representing the spectrotemporal feature space as inputs in classification models could involve using the modeled parameters of fitted spectral-temporal curves [31] or derived phenometrics [11]. These metrics have been suggested to maintain consistency in an inter-seasonal context (e.g., [4,9]), offering potentially valuable features for transferable crop classification. However, these approaches might also introduce uncertainty, depending on the raw data pattern, and could potentially degrade spectral information when spectral indices are used. Hence, conducting an in-depth assessment and comparison of these alternative methods could be beneficial in identifying the most effective approach for representing spectrotemporal features as inputs in transferable crop classification models.

4.3. Effect of Anomaly Seasons

Factors such as climate variability and extreme weather events can significantly influence the transferability of crop classification models. It is crucial to consider how these

changes might affect model performance. Predicting inter-seasonal variations and extreme events is challenging. Longer-term analyses could be advantageous for understanding complex interannual patterns of crop phenology and extreme weather effects. So far, the most pertinent domain generalization study has been carried out by Cai et al. [14], who utilized a 15-year Landsat series dataset. Their experiment design includes an assessment of accuracies for independent testing data in 2015, with the model trained on data from all years before 2014. The study showed that a greater amount of training data can enhance classification performance. Including more years of data can lead to higher performance in classifying crop types for the next year. This can even reach best-case overall accuracies of 96%. Their findings demonstrated that incorporating one to five seasons in training significantly improved classification performance on unseen years, but adding more years beyond that range caused the increasing trend to decelerate and eventually saturate. However, they also observed that outliers in performance occurred during drought years or when data availability was limited. In our study, we found that if a model is trained on data covering four seasons within the same region, it can effectively generalize across different years while maintaining an acceptable level of performance. It appears that the "com-mon" interannual variability of climate, which affects crop phenology, can be effectively tracked by the models. This outcome may be attributed to the effective aggregation of gradual crop development in monthly median composites, which could help minimize abrupt year-to-year differences in phenology. The most challenging situation occurred in 2022, when an unusual drought emerged during the summer. Naturally, we assume that with denser Sentinel-2 data series, the model can learn from a broader data space and become more efficient in handling such complex situations. We anticipate that this will be the case over the upcoming years as the Sentinel-2 data series continues to expand and become denser. In any case, the across-year transferability performance varied substantially across different regions, with notably lower performance in the eastern Slovakian lowlands. Similarly, Johnson et al. [13], employing four years of Sentinel-2 seasonal composites (2016–2019) in 16 region-specific multiseasonal transferability tasks, reported a wide range of overall accuracies, from 52% to 88%. Hence, we further demonstrate that in a multi-seasonal transfer scenario for crop classification, making it region-specific can be beneficial to minimize the impact of factors other than interannual variability. By tailoring the classification model to a specific region, the model can better account for unique regional characteristics, such as soil conditions or agronomic practices. The complexity of the crop classification task represents another factor that needs to be considered when applying transferable crop classification models. Clearly, the complexity of crop classification tasks increases when applied to more intricate crop nomenclature, as opposed to simple crop configurations involving only two or three crops [14,32]. In our study, we employed a challenging 8-class crop configuration, which necessitated addressing increased complexity within the spectrotemporal feature space. The increased number of classes means that the model must discern and learn more intricate relationships between spectral and temporal features, as well as the unique patterns and characteristics associated with each crop type. In any case, in the context of model transferability, it is important to explore whether these spectrotemporal differences among crops are independent of season and region. Understanding the interplay between crop spectral variance, seasons, and regional factors is crucial for developing reliable transfer learning models for crop classification. We observed that different crops respond differently to interannual climate variability. For example, rapeseed and sugar beet are consistently classified more accurately than maize or soybean. This can be attributed to various factors. For rapeseed, its spectral uniqueness, which reflects the early spring phenophase, may contribute to better classification performance regardless of interannual variation during this period, and it remains unaffected by other subsequent periods. For sugar beet, its distinct spectral signature might be less sensitive to climate-related factors like drought, which in turn results in a more consistent classification accuracy over time despite variations in environmental conditions. On the other hand, summer crops like maize, sunflower, and soybean exhibit greater interannual spectral variations, which may

explain their higher misclassification rates when these complex responses are not effectively captured or represented in the model structure. This highlights the importance of considering the unique spectral and temporal characteristics of each crop type when developing classification models, especially in the context of climate variability and its potential impacts on crop spectral signatures.

4.4. Regional Variability Consideration

Luo et al. [7] proposed and implemented an alternative approach to address interseasonal climate variability by including larger regions with highly diversified climatic characteristics, driven in part by altitudinal gradients, such as in France. They conducted purpose-specific sampling from French crop types, which covered most climate types found in their study area. To do so, they aimed to capture phenological differences in other regions, like Italy and Germany. However, this approach may introduce additional variability due to site-specific conditions, such as soil types, different crop configurations, or agronomic practices. In our study, we found transferability across regions to be less effective than across seasons for multicrop classification tasks. Furthermore, the observed spatial generalization exhibited noticeable year-to-year variability. This could be attributed to various factors, such as the longitudinal effect, which includes the influence of a more continental climate, site-specific factors like soil conditions and agronomic practices, or a combination of all these factors. For example, specific site conditions in the eastern Slovakian lowlands, characterized by a high proportion of heavy hydromorphic soils, can lead to typical early spring waterlogging, causing delays in sowing and increasing scattered patterns of crop development. Indeed, a more challenging situation arises when both spatial and temporal variability must be accounted for simultaneously. In contrast to Luo et al. [7], we tested an opposing approach in scenario 4. We attempted to account for complex variability across regions. Our method involved training the model on several seasons from the source region, the Danubian lowland. We assumed this would more effectively capture complex crop spectral responses under various climatic and site conditions in the target region, the eastern Slovakia lowland. This approach was different from using a single-year spatial transfer. We obtained a slight improvement in its overall performance in predicting crop classifications in an unseen region. This particular scenario could be relevant in situations where dense multi-year monitoring of crop reference data exists for one region but is unavailable in another region. Similarly, in the across-season transferability scenario, performance across regions was found to be crop-specific. For example, rapeseed and sugar beet were effectively classified across regions. By implementing scenario 4, cereals also appeared to be classified more effectively. However, similar to the across-year scenario, summer crops proved to be the most challenging to classify using across-region transfer. This highlights the importance of considering crop-specific factors when evaluating the transferability performance of classification models in different regions.

Drawing from our study and related literature, we propose a two-step approach for spatiotemporal generalization of crop classification models: First, determine the optimal spatial domain where the model can accurately represent crop type distribution; then, apply interannual training to capture and account for variability across years. The identification of the optimal spatial domain should be carried out using data from multiple seasons. Johnson et al. [13] demonstrated a considerable degree of variability in the interseasonal transferability of models across 16 counties, highlighting the complexity and challenges associated with the application of spatiotemporal transfer learning. This approach helps account for various factors, such as different crop rotation practices and local climate-driven patterns, at a detailed scale within the region (e.g., waterlogged soils). Indeed, it is not solely the extent but also the spatial heterogeneity of the agricultural landscape that plays a significant role. For example, we demonstrated differences in across-year performance in scenario 2, where the transfer of classification models performed well only in 2 out of 5 years, even though the eastern Slovakia region was smaller than the Danubian lowland. By incorporating data from multiple seasons, researchers can gain a deeper understanding

of the nuances of the region and, more importantly, differentiate between spatial and interseasonal variability drivers. Our findings emphasize the importance of considering the unique characteristics of agricultural landscapes. Such characteristics might include crop configurations, soil types, and the local climate. These factors should be taken into account when assessing the transferability of crop classification models across different regions and time periods.

4.5. Feature Importance

The variable importance analysis revealed the added value of incorporating multitemporal information in the form of seasonal monthly series. This was evidenced by the inclusion of all monthly periods in the classification models. This finding is intuitive and aligns with previous research, as documented in studies such as those by Vuolo et al. [1]. The use of multitemporal data allows for a more comprehensive understanding of crop phenology and growth patterns, ultimately leading to improved crop classification models. However, we found that in certain situations, such as during prolonged summer droughts, including these months in the classification models might not be beneficial if the training data does not reflect these conditions. This highlights the importance of considering the impact of extreme weather events on the transferability capabilities of crop classification models. Our specific analysis of the sequential inclusion of monthly composites as features in classification models aimed to provide insights into the potential implementation of a within-season crop mapping concept. Some crops, such as rapeseed, can be classified relatively early in the season; however, a tradeoff between accuracy and timeliness must be considered for other crop types, which ultimately depends on the specific agricultural application (e.g., yield prediction or crop status monitoring). In the case of complex crop classification tasks, it appears that the end of June is optimal for the earliest full crop classification with acceptable levels of accuracy. Add examples. This finding helps inform the development of crop classification models that balance both accuracy and timeliness in various agricultural monitoring contexts. In any case, this might be valid in our geographical conditions, and this timeliness might vary in other regions [5,33]. Furthermore, the variable importance analysis (Figures 9 and 10) highlights the well-reported added value of Sentinel-2 red edge spectral bands for crop classification [34,35]. In particular, spectral bands 6, 7, and 8A have been identified as important for creating transferable classification models. This can be explained by the unique properties of red edge bands, particularly their narrow spectral width, which makes them highly sensitive to vegetation characteristics. Additionally, compared to spectral bands from the visible range (such as B2, B3, and B4), which were ranked as less important for classification models, red edge bands might be less sensitive to atmospheric conditions inferred in input composites.

4.6. Future Prospect: Multi-Sensor Synergies

This finding becomes especially significant when considering the combination of Sentinel-2 data with Landsat data, which lacks the red edge spectral information. Previous research, including the study by Johnson et al. [13], has highlighted the potential improvements of Sentinel 2 over Landsat data in transferable crop classification due to its increased spatial and temporal resolution. The study suggests that Sentinel 2 data is more practical for midseason prediction. In addition, for smaller fields or those with complex boundaries, Sentinel-2 data shows better performance. This advantage is particularly evident when the data is evaluated at a 10-m resolution rather than a 30-m resolution. The combination of the two datasets, e.g., in the harmonized Landsat sentinel product—HLS30 [36] or Sen2Like dataset [37], could benefit from the increased observational capacity, as has been demonstrated by Griffiths et al. [38]. However, harmonizing approaches may introduce some uncertainty regarding the different spectral configurations of the sensors. Therefore, it is essential to carefully evaluate the trade-offs when integrating data from different sources for a transferable crop classification task. As HLS30 data was not available for our study region at the time of our analysis, assessing the added value and trade-offs associated with

using harmonized datasets will be explored in future research. In addition to the optical synergies, many studies have explored multisensory approaches that involve different sensor domains, such as synthetic aperture radar (SAR) for crop classification [39]. However, in the context of the transferability of crop classification models, there is still a notable gap in the literature [11]. In our initial study, we chose to focus on the generalization capabilities of Sentinel-2 data. This decision was made to reduce potential input uncertainty and to provide a comprehensive assessment of Sentinel-2 as a primary data source for crop classification. The potential benefits of multisensory approaches for crop classification and transferability remain an open area for future exploration. Combining optical and radar sensor data can improve the performance and transferability of crop classification models, particularly in regions with limited optical data availability due to persistent cloud cover, by accounting for wider spatial and temporal variability.

It is important to mention another possible consideration for future studies. In our research, we focused on a spectral-only approach that relied solely on earth observation data. However, recent studies employed prediction models that incorporate historical crop rotation information, enabling direct crop classification in unseen years [13] or the generation of so-called trusted labels in unseen years [32]. Johnson et al. [13] demonstrated a significant improvement by incorporating historical crop rotation administrative information into the classifier training. As more Sentinel-2 seasons become available, incorporating historical crop rotation information derived from annual single-year classifications becomes possible. By integrating these practices into classifier training, models can better account for spatial variability in agricultural practices, ultimately leading to more accurate and region-specific crop classification outcomes. Furthermore, incorporating additional ancillary data that can further explain site-specific variable conditions should be explored in future research. Although site conditions and environmental factors may be better suited for explaining natural vegetation development in areas without abrupt agronomic interventions, their inclusion in crop classification models can still provide valuable insights. For instance, some studies have examined the role of environmental similarity as a factor for label generation under unseen conditions [40]. Indeed, meaningful environmental variables like digital elevation models (DEM) can shed light on altitudinal phenology variations. In addition, specific weather parameters can offer further explanation. For instance, a delay in sowing might be attributed to prolonged rain in the early spring. More importantly, accumulated temperature or growing degree days (GDD) have been found to be an effective predictor that can capture the natural variability in crop growth [8,32]. Incorporating such variables in crop classification models may help improve model performance and transferability across different regions and time periods, leading to more accurate and robust agricultural monitoring systems.

4.7. Future Prospect: Reference Datasets

We used the LPIS (land parcel information system). It is a valuable dataset for training crop classification models across Europe. However, standardization between countries is lacking. This limitation restricts its usage across European Union (EU) countries [7]. For instance, Sykas et al. [27] reported that harmonizing contextual information across different LPIS systems in different countries was a major challenge due to the variation in crop type taxonomy structure. They resolved this issue by adapting a crop type classification structure based on the Food and Agriculture Organization (FAO) system. Nonetheless, LPIS remains a valuable resource for nationwide crop classification applications. However, since this dataset is based on farmer declarations, there may be some uncertainties that are not consistently assessed. In Slovakia, some issues may arise during classification model training due to the absence of a declaration of intercropping and two cropping systems. Intercropping has become increasingly popular in agricultural practice. Additionally, farmers do not declare whether cereals are winter or summer, and there is no declaration for maize on whether it is sown for silage or grain. These uncertainties can affect classification accuracy, and researchers should take them into account when using LPIS data for crop

classification. In any case, machine learning classification algorithms have a data-hungry nature. In order to facilitate the application of our workflow in other regions, we have made our crop-labeled data for each year and region publicly available. This dataset can be used for future research or integrated into larger datasets, contributing to recent initiatives such as Breizhcrops [41], Sen4AgriNet [27], or TimeSen2Crop [28], which aim to develop large standardized datasets for advanced classification algorithms in agriculture monitoring.

4.8. Future Prospect: Multi-Model Consideration

All classification algorithms showed good performance when trained and validated in the given year (data not shown). However, our main focus was on assessing their ability to generalize to unseen temporal and regional domains. We found that quadratic discriminant analysis, support vector machines, and neural networks consistently performed well across different scenarios, indicating their potential for generalization in crop classification models. In contrast, the random forest algorithm appeared to be less robust in its ability to generalize across different scenarios, particularly when the distribution of the unseen domains deviated significantly from that used for training. This was observed in the transfer across years in the anomaly year 2022 as well as in the transfer across regions in scenario 3 for 2021. These findings are important because many transferability studies, such as those conducted by Johnson et al. [13], have exclusively used random forest algorithms; hence, the need for multi-algorithm comparison studies is emphasized. In this regard, the development of openly accessible toolboxes, such as the one published by Aghababaei et al. [42], is advantageous for facilitating multi-model comparisons. The lower robustness of the random forest algorithm in our study may be due to its sensitivity to overfitting in complex multidimensional feature spaces [43]. This might be indicated by the fact that RF was the only algorithm that showed improvement on the feature reduction test when PCA was applied (data not shown) or when shrinking of the inputs in 2022 was applied (Figure 16). The relatively good performance of simpler algorithms, such as QDA, compared to more advanced ones like SVM or NN, indicates that the spectrotemporal feature space of the crop types used in this study is well distinguishable. This may be further due to thorough data processing and the pixel-based nature of the data space, resulting in a relatively easy classification task. Considering accuracy and computational cost, advanced algorithms like SVM and NN need more resources and time than simpler ones like QDA. This is significant because controlling computational costs is crucial. Using multiple algorithms offers the delivery of so-called confidence maps derived from a decision-level ensemble of multi-output results per pixel [26]. These confidence maps provide insight into the uncertainty of the classification results, which is valuable for decision-making processes in agriculture and land-use management. Furthermore, identifying potential spatial trends in error pixels across all classifiers can be useful in detecting undersampled areas (which were not observed in our study) or incorrectly reported crop types by farmers. We need to note that the crop map products were generated by aggregating the results of individual pixels with the existing parcel boundaries. It should be acknowledged that this may not be possible in other regions or countries where parcel boundaries are not available or are not accurately defined. In such cases, alternative methods such as segmentation-based approaches or object-based image analysis [44] may need to be applied to generate parcel-based mapping outputs. Furthermore, alternative workflows can utilize image-based approaches that employ deep learning methods to fully capitalize on contextual information within image series [45]. There have been considerable research efforts dedicated to the development of pre-trained deep networks, including transformer-based architectures specifically tailored for crop mapping [46,47]. These networks aim to provide a solid foundation for transfer learning, enabling them to be fine-tuned with limited data and effectively applied to new spatial or temporal domains in crop classification tasks. Although there have been recent advancements in this field, challenges persist, such as higher computational costs and the limited availability of suitable training datasets for fully harnessing transfer learning in nationwide satellite-based

crop mapping. Furthermore, the implementation of innovative verification workflows, such as formal methods [48,49], could contribute to enhancing the reliability and interpretability of models. These issues present opportunities for upcoming research aimed at developing transferable crop classification models.

5. Conclusions

Our study employs a domain generalization approach using a five-year Sentinel-2 dataset across various agricultural regions in Slovakia. Our results show promise for the crop classification task, as they illustrate the potential to effectively classify unseen labels from different years and regions. This may provide valuable insights for nationwide agricultural monitoring. The study emphasizes the importance of considering factors such as interseasonal climate variability, extreme weather events, diverse crop rotation practices, observational quality, and local site-specific factors. To tackle these challenges, we suggest a two-step approach: (1) determining the optimal spatial domain using data from multiple seasons; and (2) applying interannual training to capture and account for variability across years. By implementing these steps, we can develop more accurate and reliable crop classification models for nationwide agricultural monitoring. The growing availability of multi-seasonal Sentinel-2 data will enhance crop classification models by addressing challenges related to interseasonal variations and extreme weather events. Our findings show that transferability performance varies somewhat among machine learning classifiers, with quadratic discriminant analysis, support vector machines, and neural networks exhibiting better generalization potential compared to random forests. Nevertheless, it is crucial to acknowledge that future research could advance the reported workflow and refine its application to other regions. Some directions for future research include: 1. investigating the benefits of multisensor approaches, such as integrating Sentinel-2, Landsat and Sentinel-1 data for crop classification transferability; 2. exploring the incorporation of historical crop rotation information and additional ancillary data, such as environmental variables, to improve model performance and transferability; 3. examining alternative methods for generating parcel-based mapping outputs in regions without accurate parcel boundaries, such as object-based approaches; and including image-based approaches using deep learning methods for transferable crop classification models; and 4. implementation of complex formal verification workflows in the ML based crop model development.

Supplementary Materials: The following supporting information can be downloaded at https://doi.org/10.6084/m9.figshare.23309198 (accessed on 15 May 2023). Supplementary Materials contain Figures S1–S18; S1 Excel file with hyperparameter tuning for each scenario, and S2 Excel file containing classification accuracy for scenarios.

Author Contributions: Conceptualization: T.R., T.K., P.M., J.M., J.Z., D.A., M.S. and A.H.; methodology: T.R. and A.H.; software: T.R. and A.H.; validation: T.R. and A.H.; data collection: T.R. and A.H.; visualization: T.R.; writing—original draft preparation: T.R., T.K., P.M., J.M., J.Z., D.A., M.S. and A.H.; writing—review and editing: T.R., T.K., P.M., J.M., J.Z., D.A., M.S. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Integrated Infrastructure Operational Programme funded by the ERDF, project ITMS2014+ 313011W580, "Scientific support of climate change adaptation in agriculture and mitigation of soil degradation".

Data Availability Statement: The reference dataset is available at: https://doi.org/10.5281/zenodo.7912749 (accessed on 15 May 2023).

Acknowledgments: This research was supported by the Integrated Infrastructure Operational Programme funded by the ERDF, project ITMS2014+ 313011W580, "Scientific support of climate change adaptation in agriculture and mitigation of soil degradation".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vuolo, F.; Neuwirth, M.; Immitzer, M.; Atzberger, C.; Ng, W.-T. How Much Does Multi-Temporal Sentinel-2 Data Improve Crop Type Classification? *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *72*, 122–130. [CrossRef]
2. Tuia, D.; Persello, C.; Bruzzone, L. Recent Advances in Domain Adaptation for the Classification of Remote Sensing Data. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
3. Belgiu, M.; Bijker, W.; Csillik, O.; Stein, A. Phenology-Based Sample Generation for Supervised Crop Type Classification. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *95*, 102264. [CrossRef]
4. Yang, Z.; Diao, C.; Gao, F. Towards Scalable within-Season Crop Mapping with Phenology Normalization and Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1390–1402. [CrossRef]
5. Lin, C.; Zhong, L.; Song, X.-P.; Dong, J.; Lobell, D.B.; Jin, Z. Early- and in-Season Crop Type Mapping without Current-Year Ground Truth: Generating Labels from Historical Information via a Topology-Based Approach. *Remote Sens. Environ.* **2022**, *274*, 112994. [CrossRef]
6. Sexton, J.O.; Urban, D.L.; Donohue, M.J.; Song, C. Long-Term Land Cover Dynamics by Multi-Temporal Classification across the Landsat-5 Record. *Remote Sens. Environ.* **2013**, *128*, 246–258. [CrossRef]
7. Luo, Y.; Zhang, Z.; Zhang, L.; Han, J.; Cao, J.; Zhang, J. Developing High-Resolution Crop Maps for Major Crops in the European Union Based on Transductive Transfer Learning and Limited Ground Data. *Remote Sens.* **2022**, *14*, 1809. [CrossRef]
8. Wang, S.; Azzari, G.; Lobell, D.B. Crop Type Mapping without Field-Level Labels: Random Forest Transfer and Unsupervised Clustering Techniques. *Remote Sens. Environ.* **2019**, *222*, 303–317. [CrossRef]
9. Zhong, L.; Gong, P.; Biging, G.S. Efficient Corn and Soybean Mapping with Temporal Extendability: A Multi-Year Experiment Using Landsat Imagery. *Remote Sens. Environ.* **2014**, *140*, 1–13. [CrossRef]
10. Qin, R.; Liu, T. A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability. *Remote Sens.* **2022**, *14*, 646. [CrossRef]
11. Hu, Y.; Zeng, H.; Tian, F.; Zhang, M.; Wu, B.; Gilliams, S.; Li, S.; Li, Y.; Lu, Y.; Yang, H. An Interannual Transfer Learning Approach for Crop Classification in the Hetao Irrigation District, China. *Remote Sens.* **2022**, *14*, 1208. [CrossRef]
12. You, N.; Dong, J. Examining Earliest Identifiable Timing of Crops Using All Available Sentinel 1/2 Imagery and Google Earth Engine. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 109–123. [CrossRef]
13. Johnson, D.M.; Mueller, R. Pre- and within-Season Crop Type Classification Trained with Archival Land Cover Information. *Remote Sens. Environ.* **2021**, *264*, 112576. [CrossRef]
14. Cai, Y.; Guan, K.; Peng, J.; Wang, S.; Seifert, C.; Wardlow, B.; Li, Z. A High-Performance and in-Season Classification System of Field-Level Crop Types Using Time-Series Landsat Data and a Machine Learning Approach. *Remote Sens. Environ.* **2018**, *210*, 35–47. [CrossRef]
15. Lapin, M.; Faako, P.; Melo, M.; Stastny, P.; Tomlain, J. *Climatic Regions; 1:1,000,000; 27. Klimaticke Oblasti; 1:1,000,000*; Ministry of Environment of the Slovak Republic Bratislava: Bratislava, Slovakia, 2002; ISBN 80-88833-27-2.
16. Michaeli, E.; Vilček, J.; Ivanová, M. Characteristics of Agricultural Soils in the East-Slovak Lowland and the Possibilities of Improving of Their Productive Potential. *Zivotn. Prostr.* **2013**, *47*, 242–246.
17. Miklós, L.; Izakovičová, Z. *Atlas of Representative Geoecosystems of Slovakia*; Slovak Academy of Sciences, Ministry of Environment and Ministry of Education of the Slovak Republik: Bratislava, Slovakia, 2006; pp. 95–123. ISBN 80-969272-4-8.
18. Griffiths, P.; Van Der Linden, S.; Kuemmerle, T.; Hostert, P. A Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2088–2101. [CrossRef]
19. Bull, A.D. Convergence Rates of Efficient Global Optimization Algorithms. *J. Mach. Learn. Res.* **2011**, *12*, 2879–2904.
20. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
21. Phiri, D.; Morgenroth, J. Developments in Landsat Land Cover Classification Methods: A Review. *Remote Sens.* **2017**, *9*, 967. [CrossRef]
22. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
23. Mas, J.F.; Flores, J.J. The Application of Artificial Neural Networks to the Analysis of Remotely Sensed Data. *Int. J. Remote Sens.* **2008**, *29*, 617–663. [CrossRef]
24. Foody, G.M. Status of Land Cover Classification Accuracy Assessment. *Remote Sens. Environ.* **2002**, *80*, 185–201. [CrossRef]
25. Fisher, A.; Rudin, C.; Dominici, F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81.
26. Liu, W.; Gopal, S.; Woodcock, C.E. Uncertainty and Confidence in Land Cover Classification Using a Hybrid Classifier Approach. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 963–971. [CrossRef]
27. Sykas, D.; Sdraka, M.; Zografakis, D.; Papoutsis, I. A Sentinel-2 Multiyear, Multicountry Benchmark Dataset for Crop Classification and Segmentation with Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3323–3339. [CrossRef]
28. Weikmann, G.; Paris, C.; Bruzzone, L. TimeSen2Crop: A Million Labeled Samples Dataset of Sentinel 2 Image Time Series for Crop-Type Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4699–4708. [CrossRef]
29. Flood, N. Seasonal Composite Landsat TM/ETM+ Images Using the Medoid (a Multi-Dimensional Median). *Remote Sens.* **2013**, *5*, 6481–6500. [CrossRef]

30. Frantz, D.; Röder, A.; Stellmes, M.; Hill, J. Phenology-Adaptive Pixel-Based Compositing Using Optical Earth Observation Imagery. *Remote Sens. Environ.* **2017**, *190*, 331–347. [CrossRef]
31. Roy, D.P.; Yan, L. Robust Landsat-Based Crop Time Series Modelling. *Remote Sens. Environ.* **2020**, *238*, 110810. [CrossRef]
32. Zhang, L.; Gao, L.; Huang, C.; Wang, N.; Wang, S.; Peng, M.; Zhang, X.; Tong, Q. Crop Classification Based on the Spectrotemporal Signature Derived from Vegetation Indices and Accumulated Temperature. *Int. J. Digit. Earth* **2022**, *15*, 626–652. [CrossRef]
33. Wei, M.; Wang, H.; Zhang, Y.; Li, Q.; Du, X.; Shi, G.; Ren, Y. Investigating the Potential of Crop Discrimination in Early Growing Stage of Change Analysis in Remote Sensing Crop Profiles. *Remote Sens.* **2023**, *15*, 853. [CrossRef]
34. Immitzer, M.; Vuolo, F.; Atzberger, C. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
35. Yi, Z.; Jia, L.; Chen, Q. Crop Classification Using Multi-Temporal Sentinel-2 Data in the Shiyang River Basin of China. *Remote Sens.* **2020**, *12*, 4052. [CrossRef]
36. Claverie, M.; Ju, J.; Masek, J.G.; Dungan, J.L.; Vermote, E.F.; Roger, J.-C.; Skakun, S.V.; Justice, C. The Harmonized Landsat and Sentinel-2 Surface Reflectance Data Set. *Remote Sens. Environ.* **2018**, *219*, 145–161. [CrossRef]
37. Saunier, S.; Pflug, B.; Lobos, I.M.; Franch, B.; Louis, J.; De Los Reyes, R.; Debaecker, V.; Cadau, E.G.; Boccia, V.; Gascon, F.; et al. Sen2Like: Paving the Way towards Harmonization and Fusion of Optical Data. *Remote Sens.* **2022**, *14*, 3855. [CrossRef]
38. Griffiths, P.; Nendel, C.; Hostert, P. Intra-Annual Reflectance Composites from Sentinel-2 and Landsat for National-Scale Crop and Land Cover Mapping. *Remote Sens. Environ.* **2019**, *220*, 135–151. [CrossRef]
39. Kussul, N.; Mykola, L.; Shelestov, A.; Skakun, S. Crop Inventory at Regional Scale in Ukraine: Developing in Season and End of Season Crop Maps with Multi-Temporal Optical and SAR Satellite Imagery. *Eur. J. Remote Sens.* **2018**, *51*, 627–636. [CrossRef]
40. Liu, Z.; Zhang, L.; Yu, Y.; Xi, X.; Ren, T.; Zhao, Y.; Zhu, D.; Zhu, A. Cross-Year Reuse of Historical Samples for Crop Mapping Based on Environmental Similarity. *Front. Plant Sci.* **2022**, *12*, 761148. [CrossRef]
41. Rußwurm, M.; Pelletier, C.; Zollner, M.; Lefèvre, S.; Körner, M. BreizhCrops: A Time Series Dataset for Crop Type Mapping. *arXiv* **2020**, arXiv:1905.11893. [CrossRef]
42. Aghababaei, M.; Ebrahimi, A.; Naghipour, A.A.; Asadi, E.; Pérez-Suay, A.; Morata, M.; Garcia, J.L.; Rivera Caicedo, J.P.; Verrelst, J. Introducing ARTMO's Machine-Learning Classification Algorithms Toolbox: Application to Plant-Type Detection in a Semi-Steppe Iranian Landscape. *Remote Sens.* **2022**, *14*, 4452. [CrossRef]
43. Tang, B.; Bi, Y.; Li, Z.-L.; Xia, J. Generalized Split-Window Algorithm for Estimate of Land Surface Temperature from Chinese Geostationary FengYun Meteorological Satellite (FY-2C) Data. *Sensors* **2008**, *8*, 933–951. [CrossRef] [PubMed]
44. Peña-Barragán, J.M.; Ngugi, M.K.; Plant, R.E.; Six, J. Object-Based Crop Identification Using Multiple Vegetation Indices, Textural Features and Crop Phenology. *Remote Sens. Environ.* **2011**, *115*, 1301–1316. [CrossRef]
45. Gilcher, M.; Udelhoven, T. Field Geometry and the Spatial and Temporal Generalization of Crop Classification Algorithms—A Randomized Approach to Compare Pixel Based and Convolution Based Methods. *Remote Sens.* **2021**, *13*, 775. [CrossRef]
46. Yuan, Y.; Lin, L. Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 474–487. [CrossRef]
47. Nowakowski, A.; Mrziglod, J.; Spiller, D.; Bonifacio, R.; Ferrari, I.; Mathieu, P.P.; Garcia-Herranz, M.; Kim, D.-H. Crop Type Mapping by Using Transfer Learning. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *98*, 102313. [CrossRef]
48. Krichen, M.; Mihoub, A.; Alzahrani, M.Y.; Adoni, W.Y.H.; Nahhal, T. Are Formal Methods Applicable to Machine Learning and Artificial Intelligence? In Proceedings of the 2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 9–11 May 2022; pp. 48–53.
49. Raman, R.; Gupta, N.; Jeppu, Y. Framework for Formal Verification of Machine Learning Based Complex System-of-Systems. *Insight* **2023**, *26*, 91–102. [CrossRef]