



University of Reading
Department of Computer Science

Generating Tree Genus Classification and Change Maps to Assist Mitigate Climate Change

Pedro Junio

Supervisor: Muhammad Shahzad

A report submitted in partial fulfillment of the requirements of
the University of Reading for the degree of
Master of Science in *Data Science and Advanced Computing*

August 27, 2024

Declaration

I, Pedro Junio, of the Department of Computer Science, University of Reading, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalized accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning, and research.

Pedro Junio
August 27, 2024

Abstract

This study explores tree genus classification using Sentinel-2 imagery, focusing on seasonal and spectral band combinations and incorporating additional environmental data. The analysis revealed that the model performs best during summer and autumn, where clear spectral signatures enhance accuracy. The optimal spectral band combination was found to be B3, B6, B8, and B11, balancing classification performance with storage efficiency. Although integrating soil composition and elevation provided marginal improvements, Sentinel-2's high-resolution spectral data was most influential. Climate data from the ECMWF Reanalysis v5 (ERA5) showed minimal short-term impact on classification over five years, with the study's European focus possibly limiting broader applicability. Attempts to model climate change effects using a regression neural network yielded poor results, highlighting the complexity of capturing such relationships. Future research should prioritize long-term data analysis to better understand gradual climate effects, consider regions with more pronounced climate changes, and differentiate between commercial and natural forests to reveal distinct change patterns. Incorporating additional environmental variables and collaborating with ecologists could further enhance model accuracy and insights into tree genus distributions.

Keywords: Convolutional Neural Network (CNN), Sentinel-2, Copernicus, SoilGrid, EU-Forest, Forests, Tree Genus, Climate, ERA5

Report's total word count: 6000-7000

Contents

1	Introduction	1
1.1	Remote Sensing and Machine Learning in Forest Monitoring	1
1.2	Data Integration and Methodology	1
1.3	Challenges and Considerations	2
1.4	Tools and Software	2
2	Data Exploration	3
2.1	EU-Forest Labels	3
2.2	Sentinel-2 Features	5
2.3	Copernicus DEM GLO-30	9
2.4	SoilGrids	10
2.5	Summary	11
3	Feature Selection	12
3.1	Sentinel-2 Seasons	12
3.2	Sentinel-2 Bands	13
3.3	Soil and Elevation	15
3.4	Summary	15
4	Neural Network Configuration	16
4.1	Hyperparameter Optimization I	16
4.2	Hyperparameter Optimization II	17
4.3	Model Architecture	17
5	Climate Data	19
5.1	ERA5 Dataset	19
5.2	Dataset Exploration	20
6	Results and Validation	21
6.1	Classification Validation	21
6.2	Change Map Correlations	21
6.3	Relationship Modeling	22
7	Conclusions and Future Directions	24
7.1	Conclusions	24
7.2	Future Directions	25

List of Figures

2.1	Map of the most common tree genera in EU-Forest.	3
2.2	Distribution of genera (left) and species (right) in EU-Forest.	4
2.3	Distribution of genera (left) and species (right) per location.	4
2.4	Sentinel-2 mission infographic. It highlights important facts and achievements of the mission. Courtesy of ESA	5
2.5	Multiple sample locations overlaid on Google Earth (left) and Sentinel-2 (right) images.	6
2.6	Sample location overlaid on Google Earth (left) and Sentinel-2 (right) images.	6
2.7	Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalization (right) for RGB bands.	7
2.8	Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalization (right) for NIR bands.	7
2.9	Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalization (right) for SWIR bands.	8
2.10	Seasonal Sentinel-2 band correlations.	9
2.11	Distributions for elevation (left), including a sample means as a dotted line, and z-score normalization (right) for the Copernicus DEM GLO-30 dataset.	10
2.12	Distributions for soil properties (left), including a sample means as a dotted line, and z-score normalization (right) for the SoilGrids dataset.	10
3.1	Seasonal Sentinel-2 analysis for all season combinations using a 3D CNN.	12
3.2	Sentinel-2 analysis for all combinations within three band groups using a CNN. The horizontal black line represents the weighted f1-score for all bands.	14
3.3	Sentinel-2 analysis for selected combinations between each of the three groups using a CNN. The horizontal black line represents the weighted f1-score for all bands.	14
3.4	Analysis of SoilGrids and elevation data integration.	15
4.1	The top f2-scores out of nearly 700 Hyperband (Li et al. (2018)) trials. The best score is marked by a diamond shape.	16
4.2	The top f2-scores for the initial hyperparameter tuning process are highlighted, with the best score marked by a diamond shape.	17
4.3	The model's layered architecture is depicted. The top-right shows the input layers, the middle section displays the convolutional layers, and the bottom-right illustrates the fully-connected layers. The layered views were generated using VisualKeras (Gavrikov (2020)).	18
5.1	Correlations heatmap between various ERA5 variables. The data was downloaded from ECMWF (1950-Present), where detailed information is available.	19

LIST OF FIGURES

v

5.2	Median and quantiles of various monthly ERA5 variables by year.	20
6.1	Median and quantiles of Pearson correlations between changes in tree genera maps predicted by classification and differences in meteorological conditions. .	21
6.2	R ² score (top-left), MAE (top-right), and residuals (bottom) for a narrow, fully-connected regression neural network, using meteorological change maps as predictors for changes in tree genus maps derived from tree classification. .	23

Chapter 1

Introduction

Forests play a critical role in regulating the Earth's climate and supporting biodiversity, as emphasized by studies such as [Bonan \(2008\)](#) and [Watson et al. \(2018\)](#). However, they are increasingly threatened by human activities and environmental changes, including deforestation, land-use conversion, and climate change, which are well-documented in [Food and Agriculture Organization of the United Nations \(2020\)](#) and [Hansen et al. \(2013\)](#). Understanding the composition and dynamics of forest ecosystems is essential for effective conservation and management efforts, as discussed in [Turner et al. \(2003\)](#). Accurate mapping of tree species and monitoring changes in forest cover over time, as demonstrated by [Sexton et al. \(2013\)](#), are fundamental tasks in ecosystem management. Such maps provide valuable information for assessing biodiversity, tracking habitat loss, and understanding the impacts of climate change on forest ecosystems, as detailed by [Vose et al. \(2018\)](#).

1.1 Remote Sensing and Machine Learning in Forest Monitoring

Remote sensing technologies, including satellite imagery and LiDAR data, have revolutionized the ability to map and monitor forests at regional and global scales. Satellite imagery has provided comprehensive coverage and frequent updates, while LiDAR data offers precise measurements of forest structure, as shown in [Pettorelli et al. \(2016\)](#) and [Lefsky et al. \(2002\)](#). Recent advancements in machine learning algorithms, particularly Convolutional Neural Networks (CNNs) and Random Forests (RF), have significantly enhanced the accuracy of tree species predictions from remotely sensed data ([Zheng and Wu \(2019\)](#), [Breiman \(2001\)](#)). This study aims to develop and evaluate methods for generating tree genus predictions and change maps using these advanced remote sensing technologies. By leveraging these innovations, this study aims to advance our understanding of forest responses to environmental changes. Building on the methodologies and findings of recent studies such as [Hansen et al. \(2013\)](#), [Bolyn et al. \(2022\)](#), [Mehmood et al. \(2024\)](#), [Ahlsweide et al. \(2023\)](#), and [Wessel et al. \(2018\)](#), this research seeks to refine and enhance predictive models and change maps, offering new insights into how forests adapt to and are affected by shifting environmental conditions.

1.2 Data Integration and Methodology

In this research, high-resolution Sentinel-2 imagery has been integrated with detailed EU-Forest data ([Mauri et al. \(2017\)](#)), Copernicus DEM elevation information ([Copernicus \(2011-2015\)](#)), SoilGrids soil properties ([Poggio et al. \(2021\)](#)), and ECMWF Reanalysis v5 (ERA5) climate data ([ECMWF \(1950-Present\)](#)). This combination of datasets provides a robust foundation

for developing a CNN classifier. The classifier is designed to accurately identify tree genera across diverse European landscapes. Sentinel-2 imagery provides detailed multispectral data at a 10-meter and 20-meter resolutions, capturing the intricate spectral signatures of vegetation. The EU-Forest dataset offers essential ground-truth data across about 250,000 locations in Europe, facilitating the training and validation of machine learning models. Elevation data from the Copernicus DEM GLO-30 dataset introduces topographical context that influences vegetation distribution, while SoilGrids provides comprehensive soil composition information crucial for understanding the ecological niches of various tree genera.

Furthermore, climate variables, including temperature, precipitation, and soil moisture, have been integrated into the analysis to model the relationships between environmental factors and vegetation dynamics. This integration aims to provide deeper insights into how climate change is affecting forest ecosystems and to enhance the predictive power of the classification models. By combining these diverse datasets, the study endeavors to create a more nuanced and reliable model for tree genus classification.

1.3 Challenges and Considerations

Detecting changes in forest composition is essential for assessing the impacts of disturbances such as wildfire, climate change, and tree diseases. Time-series analysis of satellite imagery, combined with advanced algorithms for change detection, can enable researchers to quantify forest dynamics and identify areas undergoing significant ecological transitions. The inclusion of climate data further enriches this analysis by allowing for the exploration of how shifting environmental conditions are influencing forest ecosystems over time.

Despite these efforts, several challenges persist in remote sensing-based tree species or genus classification and change mapping. These include the need for improved methods to handle the complexity of forest ecosystems, increasing the availability of ground-truth data, incorporating uncertainty into mapping algorithms, and scaling up analyses to cover larger geographical areas. Additionally, the addition of elevation, soil, and climate data in this study resulted in only marginal improvements in model accuracy, prompting a critical evaluation of the relative importance of spectral versus environmental data in tree genus classification. This outcome suggests that while these additional datasets provide valuable context, their practical impact on classification accuracy within this specific framework may be limited.

This research highlights the complexities and trade-offs involved in integrating multi-source data into remote sensing models for forest monitoring. The findings underscore the importance of continued exploration and refinement of methodologies to optimize the use of diverse environmental datasets in understanding and conserving forest ecosystems in the face of global environmental change.

1.4 Tools and Software

The analysis was conducted using a combination of Python files and Jupyter notebooks. Data for the study was obtained through the Google Earth Engine Python API. The pre-processing of this data involved several libraries, including NumPy, Scikit-Learn, Pandas, and GeoPandas. Data visualization was carried out using Plotly, Geemap, and Holoviews. For model development, TensorFlow and Keras were utilized to build and train machine learning models. Hyperparameter tuning was performed with Keras Tuner and the Hyperband algorithm to enhance model performance. The source code for this report and the code created during this study can be found at <https://github.com/pj097/ClimateForestModeling>.

Chapter 2

Data Exploration

2.1 EU-Forest Labels

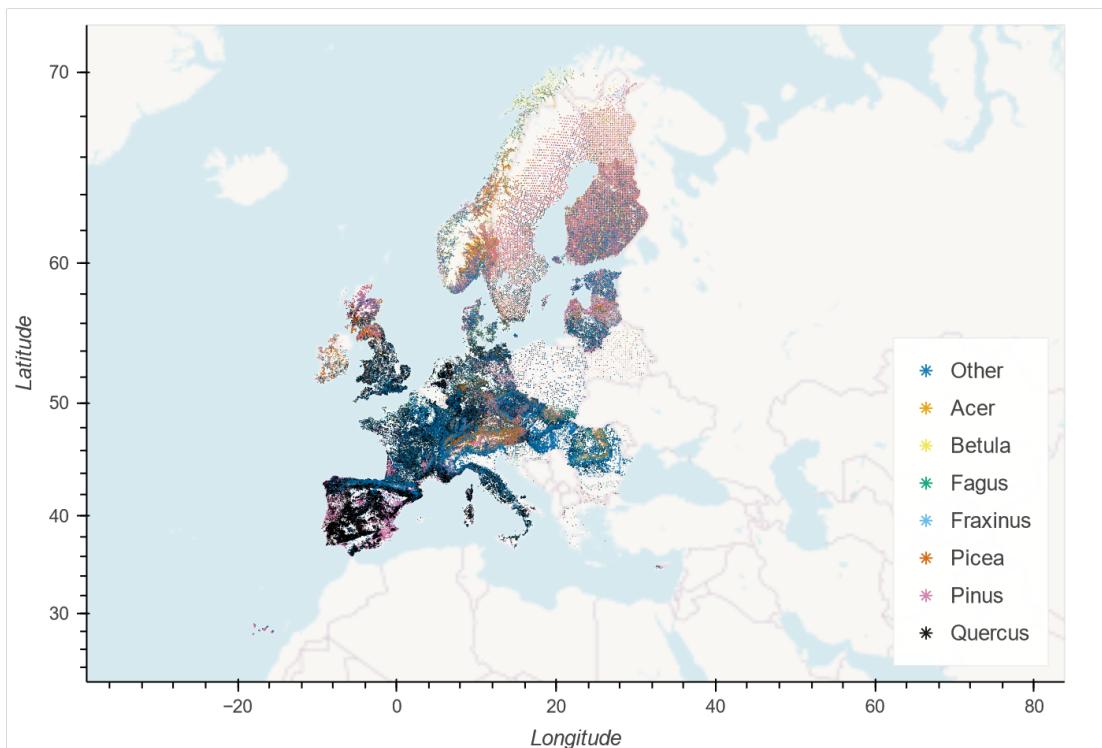


Figure 2.1: Map of the most common tree genera in EU-Forest.

EU-Forest is a dataset containing tree species and genera for nearly 250,000 locations across Europe. Each plot is $1\text{ km} \times 1\text{ km}$ and may contain multiple tree species and genera. Fig. 2.1 shows the distribution of tree genera in the EU-Forest dataset across 21 European countries. In this figure, the label 'Other' is an umbrella class for 70 tree genera with less than 20,000 occurrences each.

Using the EU-Forest dataset to train a CNN classifier of tree genera with Sentinel-2 data offers several significant advantages. Firstly, the dataset's high spatial resolution enables fine-grained analysis of tree species distribution, which enhances the accuracy of the classifier. Its comprehensive coverage across Europe, including diverse forest types and geographical areas, allows the model to learn from a wide variety of environments and tree genera. The rich

occurrence data provides detailed information on tree species, aiding in precise identification and classification.

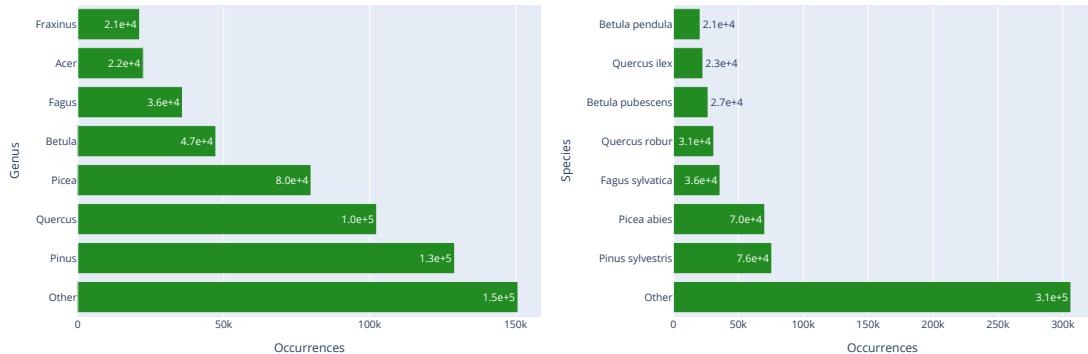


Figure 2.2: Distribution of genera (left) and species (right) in EU-Forest.

Integration with Sentinel-2 satellite data, which provides high-resolution multispectral images, allows for a robust model that leverages both ground-truth data and spectral information. The dataset, being relatively recent, offers a contemporary snapshot of forest conditions, ensuring that the trained model is relevant to current ecological and environmental conditions.

The plots in Fig. 2.2 underscore the prevalence of certain genera and species in European forests, providing valuable insights for training a CNN classifier. The dominance of specific genera and species in the dataset can enhance the classifier's ability to accurately identify and classify tree types when combined with Sentinel-2 data.

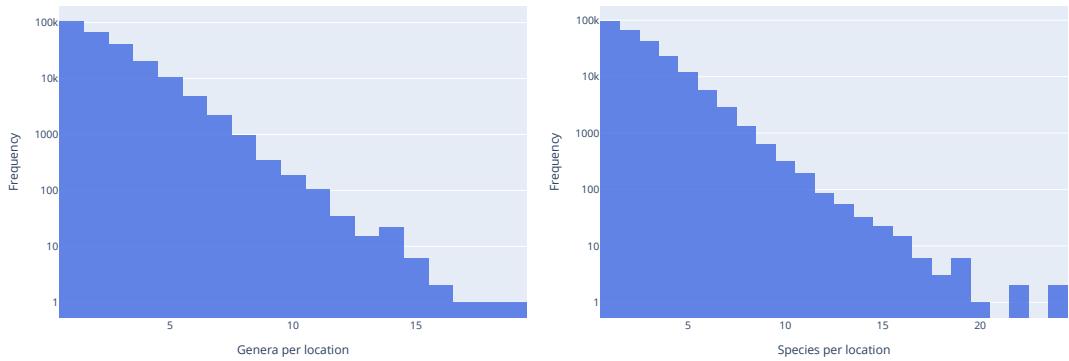


Figure 2.3: Distribution of genera (left) and species (right) per location.

The plots in Fig. 2.3 reveal a common pattern in biodiversity studies: most locations are characterized by a limited number of dominant genera and species, with a smaller number of locations exhibiting higher diversity. This pattern is important for training a CNN classifier, as it indicates that the classifier will often encounter locations with limited genera and species. However, it must also be capable of handling the less common, more diverse locations. The high-frequency, low-diversity areas will likely dominate the training process, influencing the classifier's ability to generalize across different forest types.

2.2 Sentinel-2 Features

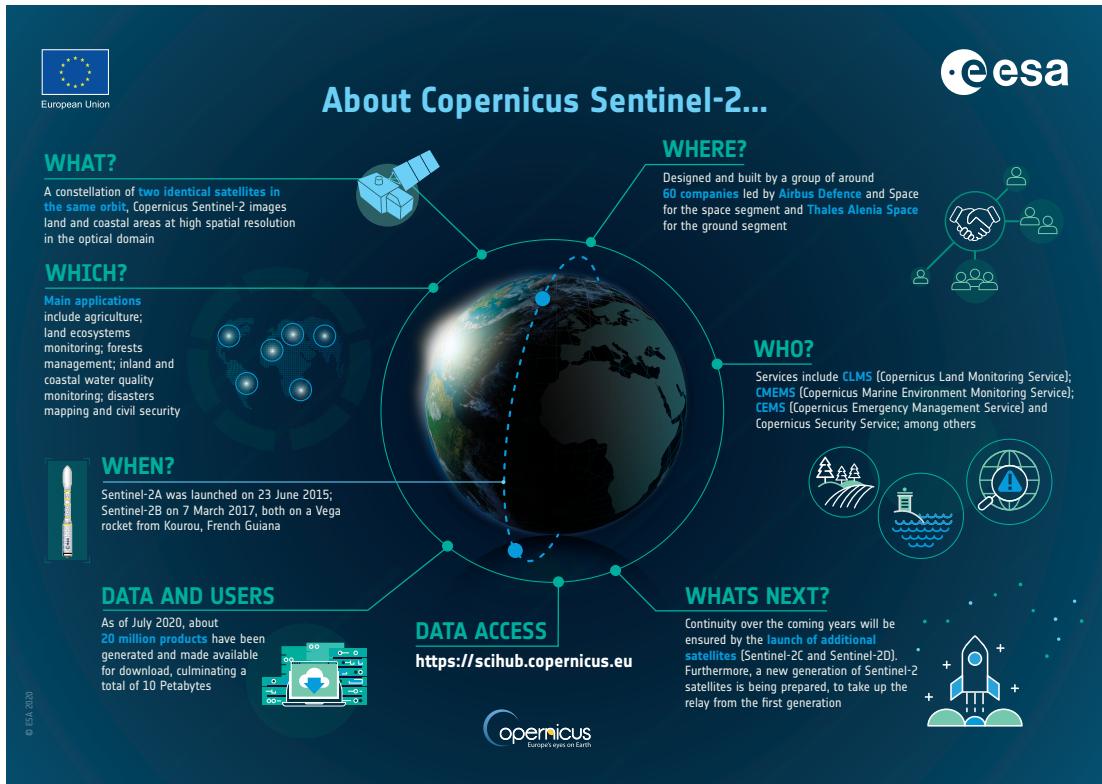


Figure 2.4: Sentinel-2 mission infographic. It highlights important facts and achievements of the mission. Courtesy of [ESA](#).

Using Sentinel-2, Fig. 2.4, specifically the 10-meter and 20-meter resolution bands, for training a CNN classifier of tree genera offers several significant advantages. Sentinel-2 provides high-resolution imagery with these bands capturing detailed spatial information essential for precise classification tasks.

The 10-meter resolution bands include visible red, green, and blue (RGB) wavelengths, as well as near-infrared (NIR) wavelengths, which are essential for assessing vegetation health and differentiating tree genera based on their reflectance properties. The 20-meter resolution bands encompass the red-edge, shortwave infrared (SWIR), and additional near-infrared regions, which enhance the classifier's ability to distinguish between tree genera by capturing subtle variations in spectral signatures. For subsequent analysis, the 20-meter bands were resampled to match the 10-meter resolution.

The multispectral imaging capability of Sentinel-2, with these selected bands, allows for detailed analysis and precise classification of tree genera. Each genus reflects and absorbs light differently across these wavelengths, providing rich data for the classifier to learn from and accurately identify tree types.

Moreover, Sentinel-2 has a frequent revisit time, with satellites passing over the same area every 5 days at the equator. This frequent update cycle is crucial for handling cloud cover, as it increases the likelihood of acquiring cloud-free images, ensuring that the classifier is trained on clear and usable data.

Sentinel-2 also offers extensive geographical coverage, capturing large areas in each image. This comprehensive coverage is essential for training classifiers intended for wide-ranging applications across different forest types and regions, and it supports the development of

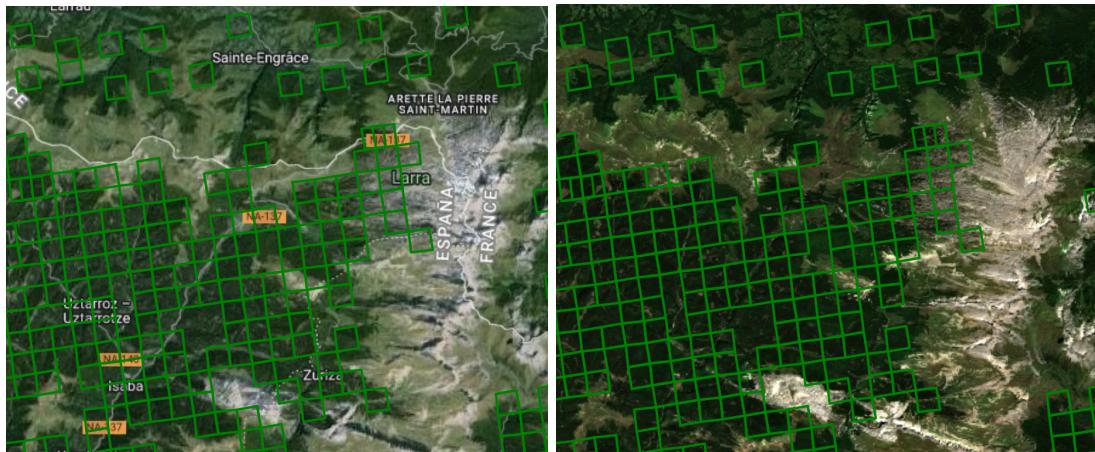


Figure 2.5: Multiple sample locations overlaid on Google Earth (left) and Sentinel-2 (right) images.

global models for tree genus classification.

Additionally, Sentinel-2 data is freely available through the European Space Agency's Copernicus program and Google Earth Engine. This open access eliminates budget constraints, ensuring high-quality satellite data is accessible for research and operational purposes.



Figure 2.6: Sample location overlaid on Google Earth (left) and Sentinel-2 (right) images.

Figs. 2.5 and 2.6 illustrate the integration of high-resolution geographical context with Sentinel-2 satellite data for detailed environmental analysis. The grid overlay in the images represents areas for data collection and analysis. The left images provide a more detailed geographical context, while the right images show how Sentinel-2 bands B2, B3, and B4 (blue, green, and red respectively) are structured and utilized for spectral analysis.

Figs. 2.7, 2.8, and 2.9 provide a detailed look at the distribution of surface reflectance values and their z-scores for various Sentinel-2 spectral bands, a normalization method which is crucial for many classification tasks using CNNs. These figures use medians taken over the summer months (June, July, and August) between 2017, 2018, and 2019.

Fig. 2.7 shows that the reflectance values for the bands B2, B3, and B4 generally follow a log-normal distribution. The z-scores for these bands peak near zero, demonstrating that the reflectance values have been normalized effectively. This normalization is essential for ensuring

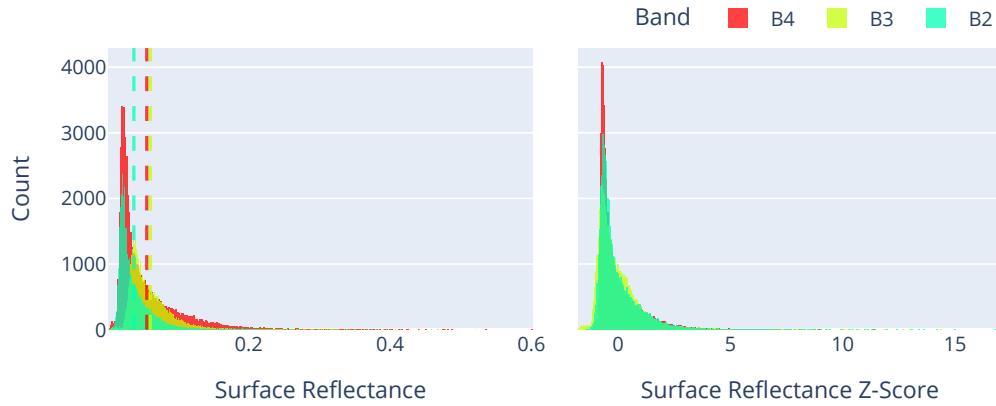


Figure 2.7: Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalization (right) for RGB bands.

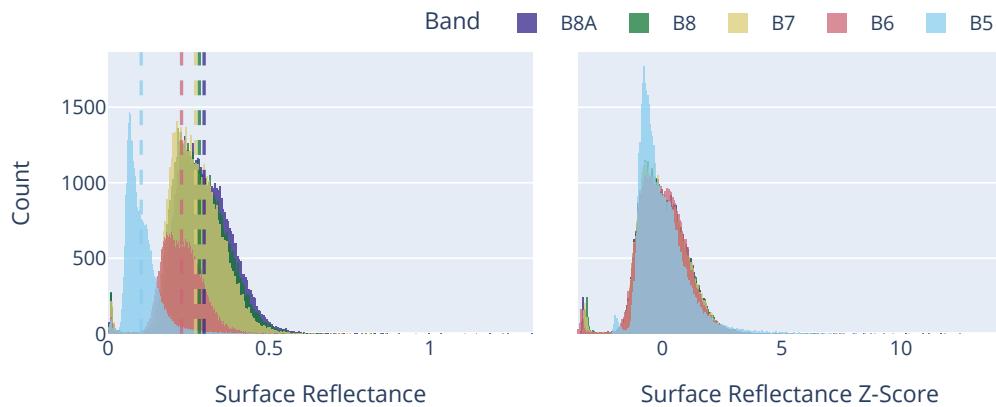


Figure 2.8: Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalization (right) for NIR bands.

that the values from different bands are on the same scale, which is particularly important when using CNN for classification.

Fig. 2.8 presents the distribution for the NIR bands B5, B6, B7, B8, and B8A. The reflectance values for these bands display a central tendency, with a significant number of pixels having values around the mean. The z-score distributions, peaking near zero, indicate successful normalization of the reflectance values. Normalizing the data in this way ensures that the CNN can process and compare these values more effectively, enhancing the model's ability to classify different types of tree genera accurately.

Fig. 2.9 shows histograms for the SWIR bands B11 and B12. The bands display a similar pattern, with the majority of reflectance values clustering on the left of the mean and tailing off to the right of the mean. The z-scores for these bands also peak at zero, confirming that the data has been normalized successfully.

The correlations among different spectral bands of Sentinel-2 data across seasons, as depicted in the correlation matrices in Fig. 2.10, highlight significant patterns that are useful for tree genus classification. All plots show the Pearson's correlation coefficients between various bands, providing insights into how these relationships change with the seasons.

During winter, all bands in NIR and RGB groups show very strong correlations with each other. These high correlations, often close to 1, indicate that the spectral responses of these bands are highly similar during this season. This consistency is likely due to the uniform

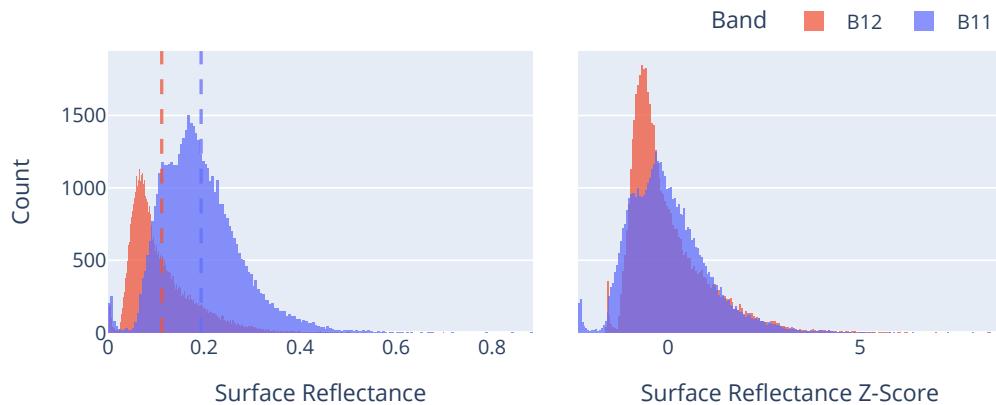


Figure 2.9: Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalization (right) for SWIR bands.

reflectance properties of vegetation and the ground cover in winter. This strong correlation is beneficial for tree genus classification as it suggests that the data from these bands can be reliably used to distinguish tree genera based on their reflectance characteristics, or lack thereof, in winter.

In spring, the correlations remain high within the NIR and RGB groups but to a slightly lesser degree compared to winter. The correlation between these two groups starts to decrease, reflecting the changes in vegetation as new growth begins. This seasonal variation provides additional information that can be leveraged to improve the classification models, as the differences in reflectance between the bands become more pronounced.

During summer and autumn, while there are still strong correlations within the NIR and RGB groups, the correlation between these groups is notably lower. This reduced correlation can be attributed to the varying phenological stages of the trees, including differences in leaf development and moisture content. These seasonal changes affect the reflectance properties differently in the NIR and RGB bands. The distinct spectral responses during these seasons offer complementary information that can enhance the classification of tree genera by providing a more diverse set of data points that capture the variability in tree characteristics.

Overall, the varying correlations across seasons underscore the robustness of using Sentinel-2 data for tree genus classification. Normalizing the data using z-scores is crucial as it standardizes the values across different bands, ensuring they are on the same scale. This is particularly important for CNNs, which are sensitive to the relative magnitudes of input data. By bringing the bands to a comparable scale, z-score normalization enhances the model's ability to process and accurately compare the multi-band data, leading to more reliable and precise classification results.

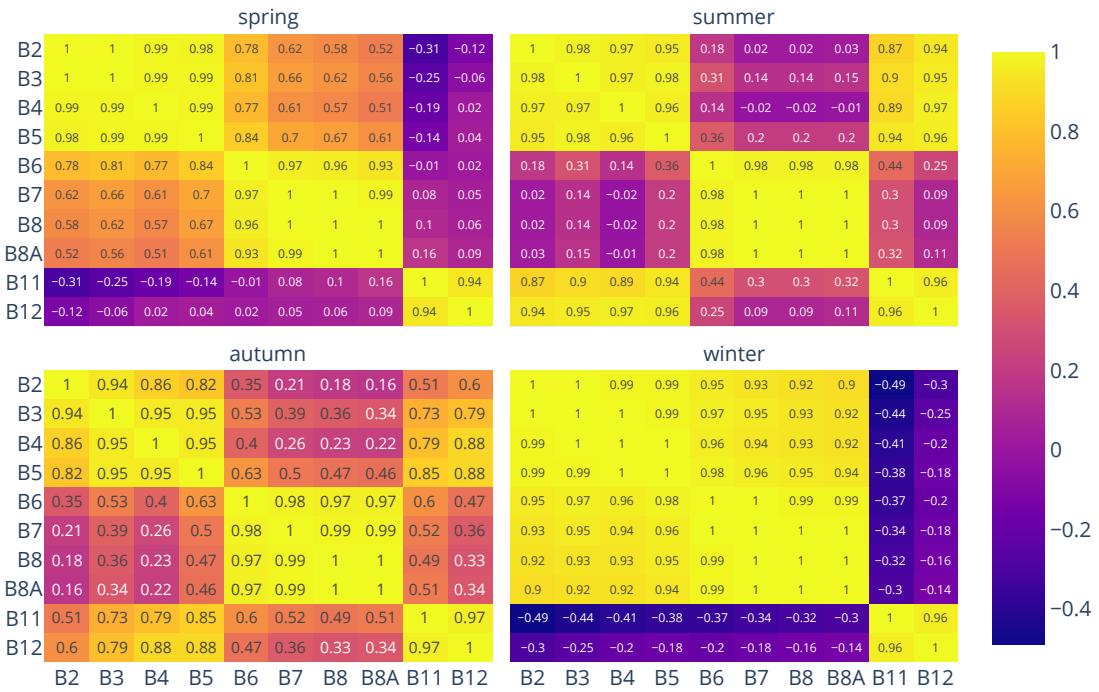


Figure 2.10: Seasonal Sentinel-2 band correlations.

2.3 Copernicus DEM GLO-30

Integrating the Copernicus Global 30 m Digital Elevation Model (DEM) data into the classification model can significantly enhance its performance for tree genus classification. Elevation data provides critical information about the terrain, which influences various ecological factors such as temperature, moisture availability, and soil type. These factors, in turn, affect vegetation types and distribution. The 30 m resolution of the DEM is particularly well-suited for this purpose, as it offers a fine enough scale to capture relevant topographic features while being broad enough to complement the 10 m and 20 m resolutions of Sentinel-2 data. This synergy allows the model to better understand the environmental context of each location, leading to more accurate predictions. For example, certain tree genera may be more prevalent at specific elevation ranges due to their adaptation to particular climatic conditions or soil properties. Therefore, integrating elevation data can help in distinguishing between tree genera that occupy different ecological niches.

Fig. 2.11 presents the distribution of elevation values in meters, as well as the z-scores, which standardize these values. The elevation values range from 0 to around 2500 meters, with most data points clustered below 500 meters. This suggests that the majority of the study area is relatively low-lying. Standardizing elevation values using z-scores is beneficial for integrating elevation data with spectral data, as it ensures that elevation values are on a comparable scale to the reflectance values from Sentinel-2 bands.

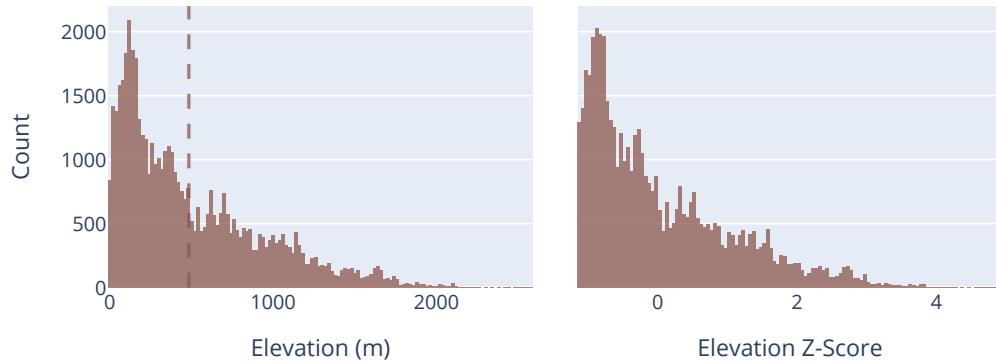


Figure 2.11: Distributions for elevation (left), including a sample means as a dotted line, and z-score normalization (right) for the Copernicus DEM GLO-30 dataset.

2.4 SoilGrids

The SoilGrids dataset provides global soil information at a spatial resolution of 250 meters and incorporates quantified spatial uncertainty. It includes key soil properties such as organic carbon content, pH, sand, silt, and clay percentages, bulk density, cation exchange capacity, and more. These data are derived from machine learning models trained on extensive soil sample databases and environmental covariates.

Integrating SoilGrids data into the tree genus classification model can significantly enhance its performance. Soil properties profoundly influence vegetation types and distribution, as different tree genera have specific soil requirements and preferences. For instance, soil pH, nutrient content, and texture can determine the suitability of a habitat for particular tree species. By incorporating detailed soil composition data, the model can better understand the environmental context, leading to more accurate predictions of tree genera. This integration helps to account for the ecological niche of each genus, improving the robustness and precision of the classification.

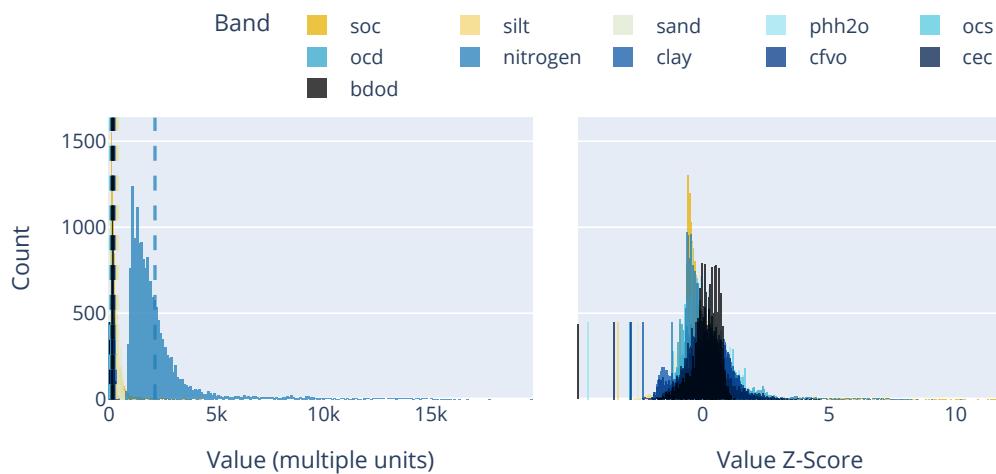


Figure 2.12: Distributions for soil properties (left), including a sample means as a dotted line, and z-score normalization (right) for the SoilGrids dataset.

Fig. 2.12 displays the distribution of various soil properties present in the SoilGrids dataset.

The x-axis represents the value of these properties in multiple units, while the y-axis represents the count of observations. Additionally, the z-score distribution is shown to standardize these values.

The distribution of soil properties varies, reflecting the diversity of soil types across the study area. The z-score distribution standardizes these values, bringing them onto a common scale, which is essential for integrating soil data with spectral and elevation data in the classification model.

2.5 Summary

While the integration of high-resolution Sentinel-2 imagery with EU-Forest data, Copernicus DEM elevation information, and SoilGrids soil properties provided a comprehensive dataset, the improvement in model accuracy was marginal. This suggests that, for the task of tree genus classification, the spectral information from Sentinel-2 may be sufficiently robust on its own. However, these additional datasets might still offer value in specific contexts or could enhance model performance in more complex classification tasks or in areas with greater environmental variability.

Chapter 3

Feature Selection

3.1 Sentinel-2 Seasons

Fig. 3.1 shows the performance of the CNN model across different seasons, with metrics including recall, precision, weighted f1-score, precision-recall curve area under the curve (PRC), and receiver operating characteristic area under the curve (AUC). Filled boxes indicate the combination of seasons used to train and validate the model. Blue boxes indicate a combination of 3D CNN and fully-connected layers and the orange box indicates the use of a similar model but with the introduction Long Short-Term Memory (LSTM) alongside 3D convolutions and fully-connected layers.

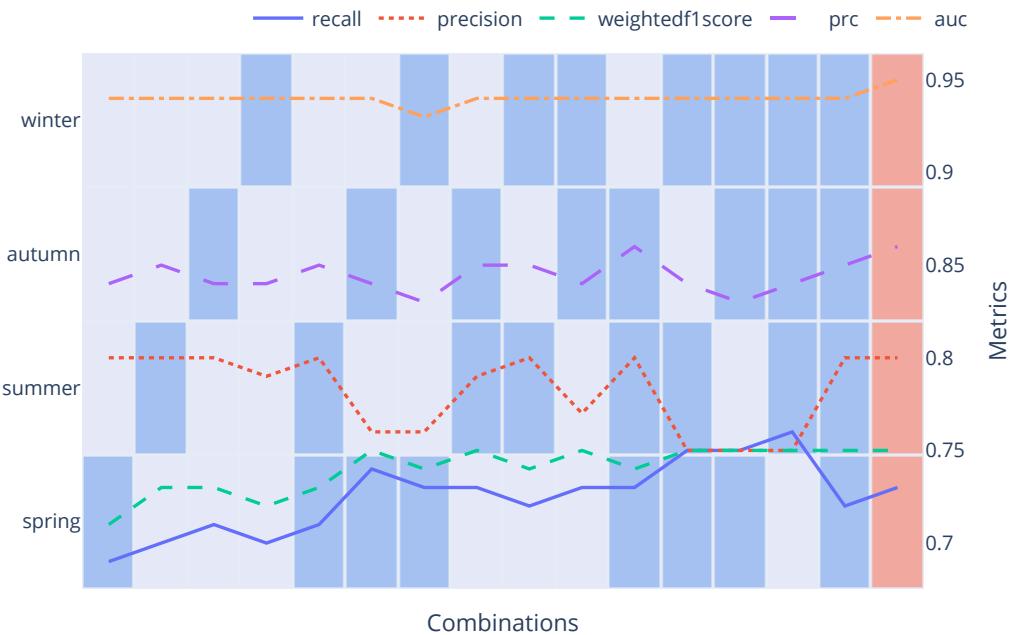


Figure 3.1: Seasonal Sentinel-2 analysis for all season combinations using a 3D CNN.

3D convolutions were selected for this model due to their suitability for this problem as they can effectively capture the spatial and spectral dependencies in the multi-temporal Sentinel-2 data. By considering the additional temporal dimension, 3D convolutions can leverage seasonal variations and changes in vegetation phenology, which are crucial for accurate tree

genus classification.

LSTM was introduced to the model alongside 2D convolutions because they process the spatial structure of the Sentinel-2 images through convolution operations while simultaneously capturing temporal sequences. This dual capability allows the model to learn intricate spatial patterns within each image and understand how these patterns evolve over time.

The selected metrics used to create Fig. 3.1 are tailored to address class imbalance. Recall measures the ability to capture instances of minority classes, while precision assesses the accuracy of positive predictions. The weighted f1-score combines precision and recall, considering class support. PRC emphasizes the precision-recall trade-off, particularly for minority classes. AUC evaluates overall model performance across all thresholds, indicating its discriminative ability.

For individual seasons, the model performs best in summer and autumn overall across the metrics. This suggests that the CNN model is more effective at classifying tree genera during these times, likely due to clearer and more distinct spectral signatures in the data collected during summer and autumn. The lower performance in spring and winter might be attributed to less distinct spectral signatures or more challenging environmental conditions, such as cloud cover and snow, which can affect data quality.

Based on the weighted f1-scores shown in Fig. 3.1, adding more seasons does not seem to offer significant benefits. For instance, some two-season combinations, such as summer and autumn, performed on par with the more complex four-season models. Additionally, single-season models were only a few percentage points below the top-performing models.

Based on these results, further analysis focused solely on summer seasons. This approach benefits from faster model training and reduced storage requirements, as adding an extra season nearly doubles the storage needs, a challenge that intensifies with the extension of the analysis over additional years. Despite these adjustments, a complete Sentinel-2 dataset for a single season still requires nearly 200 GB, or approximately 1 MB per location.

3.2 Sentinel-2 Bands

In addition to the seasonal analysis in Section 3.1, another analysis was conducted to identify the most effective band combinations. Due to the large number of possible combinations, a direct analysis was impractical. Instead, Sentinel-2 bands were divided into three groups based on summer correlation groupings shown in Fig. 2.10: B2, B3, B4, and B5; B6, B7, B8, and B8A; and B11 and B12.

The resulting analysis, shown in Fig. 3.2, indicates that NIR and SWIR bands perform slightly better overall. Based on these results, another group was selected: B2, B3, B6, B8, and B11. These bands represented the best combinations that resulted in a practical analysis within the available timeframe.

Fig. 3.3 shows that most selected combinations performed relatively well, as indicated by their proximity to the weighted f1-score for all bands. Based on these results, the most reasonable choices appear to be B3, B8, and B11, or the same but with B6 in addition. As the B6 combination displays better recall, precision, and weighted f1-score, as well as taking a fraction of storage compared to 10 m bands due to being 20 m. For the 250,000 samples, this combination should take roughly 50 GB of storage. As such, the combination B3, B6, B8, and B11, was used for the remainder of this study.



Figure 3.2: Sentinel-2 analysis for all combinations within three band groups using a CNN. The horizontal black line represents the weighted f1-score for all bands.

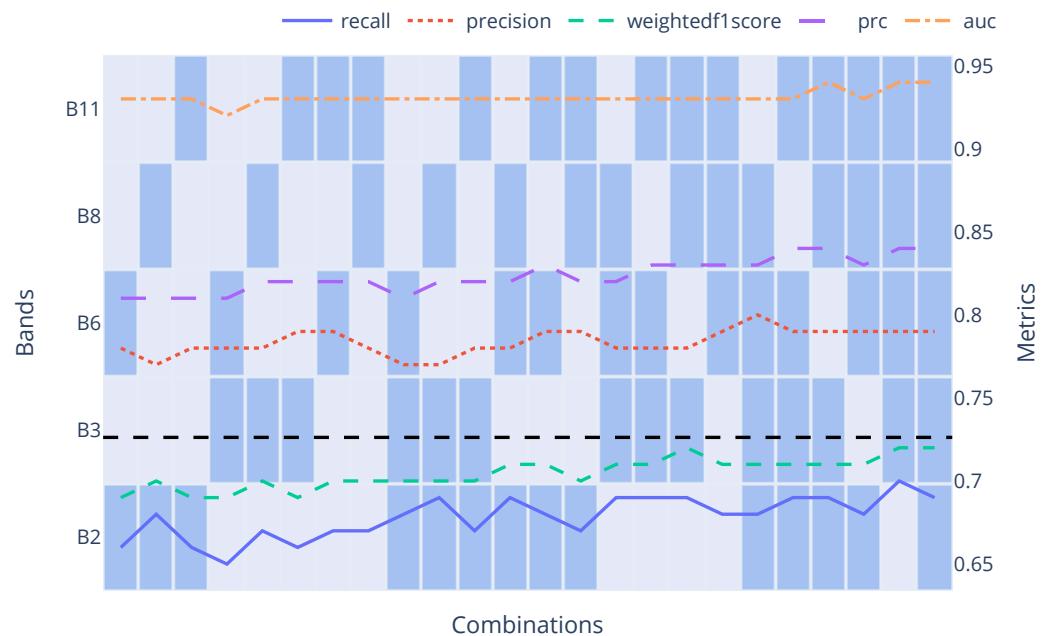


Figure 3.3: Sentinel-2 analysis for selected combinations between each of the three groups using a CNN. The horizontal black line represents the weighted f1-score for all bands.

3.3 Soil and Elevation

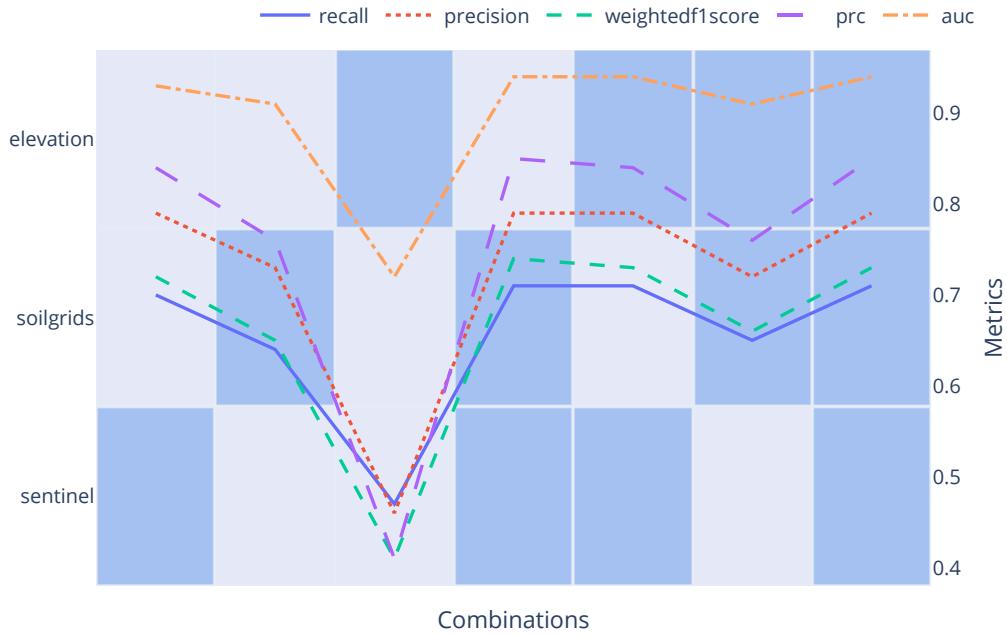


Figure 3.4: Analysis of SoilGrids and elevation data integration.

Fig. 3.4 demonstrates the performance metrics for different combinations of Sentinel-2, SoilGrids, and elevation data in tree genus classification. Sentinel-2 data alone provides a solid baseline, attributed to its high resolution and rich spectral information. When considering SoilGrids data alone, the metrics are lower than those for Sentinel-2, reflecting the coarse resolution and potentially limited predictive power for this specific task. Similarly, elevation data alone shows lower performance metrics, indicating that while elevation is a relevant feature, it does not provide sufficient information by itself for accurate classification of tree genera. The combined use of these datasets shows marginal improvements, suggesting that while additional data may contribute some value, Sentinel-2's high-resolution spectral data is the most significant factor in the model's performance.

Given that Sentinel-2 data alone provides strong results, further enhancements and optimizations were solely focused on this dataset.

3.4 Summary

The analysis highlights the influence of various seasonal combinations, Sentinel-2 bands, and additional datasets such as soil composition and elevation data on the performance of the CNN model for tree genus classification. Summer and autumn seasons were found to be the most effective for accurate classification, with minimal gains observed from including additional seasons. In terms of band selection, NIR and SWIR bands demonstrated the best results, leading to the selection of the B3, B6, B8, and B11 combination for further analysis due to its optimal balance between performance and storage efficiency. While soil composition and elevation data provided marginal improvements, Sentinel-2 data remained the most critical factor in model accuracy. Consequently, further optimizations focused primarily on the B3, B6, B8, and B11 bands of Sentinel-2.

Chapter 4

Neural Network Configuration

4.1 Hyperparameter Optimization I

Fig. 4.1 provides a summary of the search space for the initial hyperparameter tuning process and its results. These parameters are critical as they influence the model's learning process, generalization ability, and susceptibility to issues such as overfitting or bias. The chosen values in the search space represent a balance between exploring a wide range of options and focusing on promising areas based on prior knowledge or domain-specific considerations.

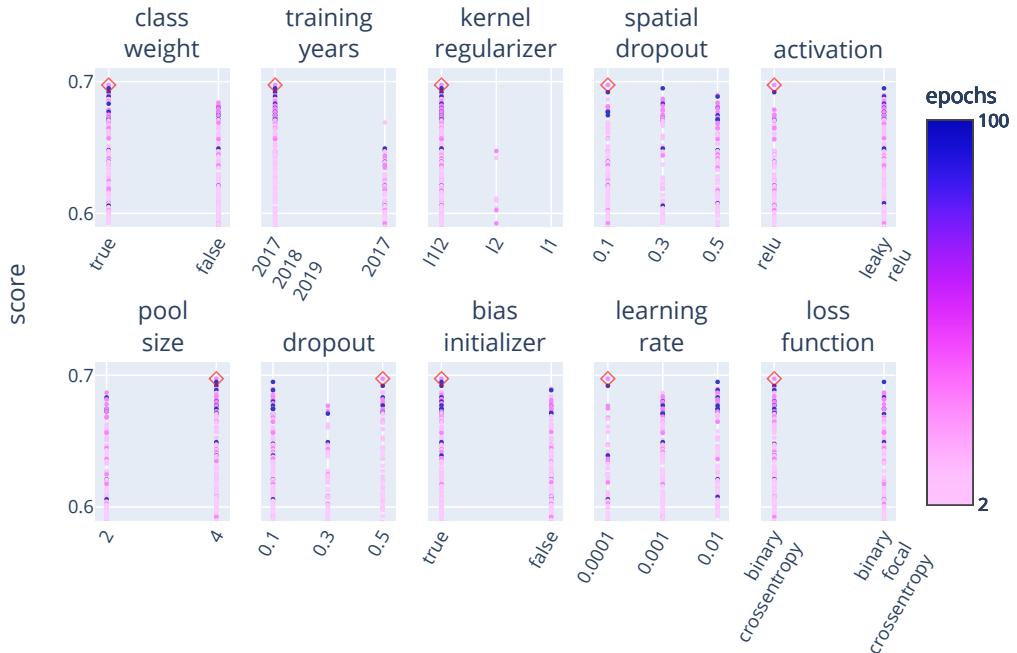


Figure 4.1: The top f2-scores out of nearly 700 Hyperband (Li et al. (2018)) trials. The best score is marked by a diamond shape.

The most successful trial values are marked by a diamond shape in Fig. 4.1. While some of the initial trial's hyperparameters do not show significant overall improvement, some parameters such as class weight, training years, and kernel regularizer do stand out.

4.2 Hyperparameter Optimization II

A second hyperparameter tuning process was performed using the best values from the most successful trial. This second set of trials looked into the number of layers, the filter size of the convolutional layers, and the number of units in the fully-connected (dense) layers.

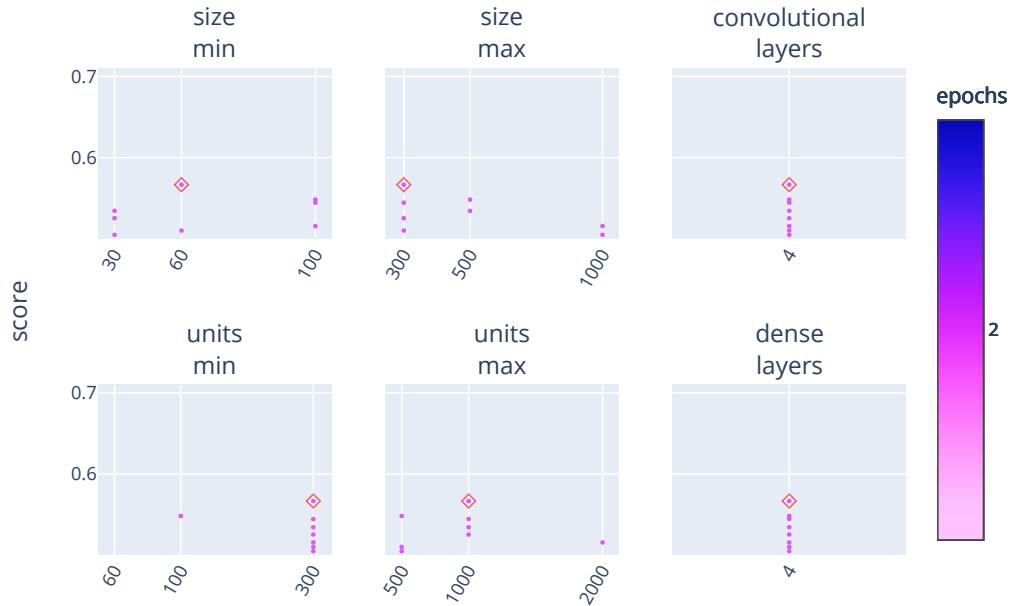


Figure 4.2: The top f2-scores for the initial hyperparameter tuning process are highlighted, with the best score marked by a diamond shape.

4.3 Model Architecture

Fig. 4.3 illustrates the detailed architecture of the deep learning model built using the results of the hyperparameter tuning processes. The figure highlights the flow of data through its various layers. The model starts with input layers, which include upsampling layers to handle inputs with different dimensions. This ensures that all input sources are properly aligned for processing within the model. The network then progresses through multiple convolutional layers, with filter sizes generally increasing as the layers deepen. Batch normalization and dropout layers are strategically placed to regularize the model and prevent overfitting.

The architecture incorporates residual connections for maintaining gradient flow during backpropagation. These connections enable the network to be deep without suffering from issues like vanishing gradients, as highlighted in He et al. (2015). As the network advances, spatial dimensions are reduced through pooling and striding operations, leading to a flattening layer that prepares the data for the dense layers. These dense layers further reduce the dimensionality, eventually outputting a 7-dimensional vector corresponding to the classification into 7 different tree genera, aligning with the model's goal of predicting tree types.

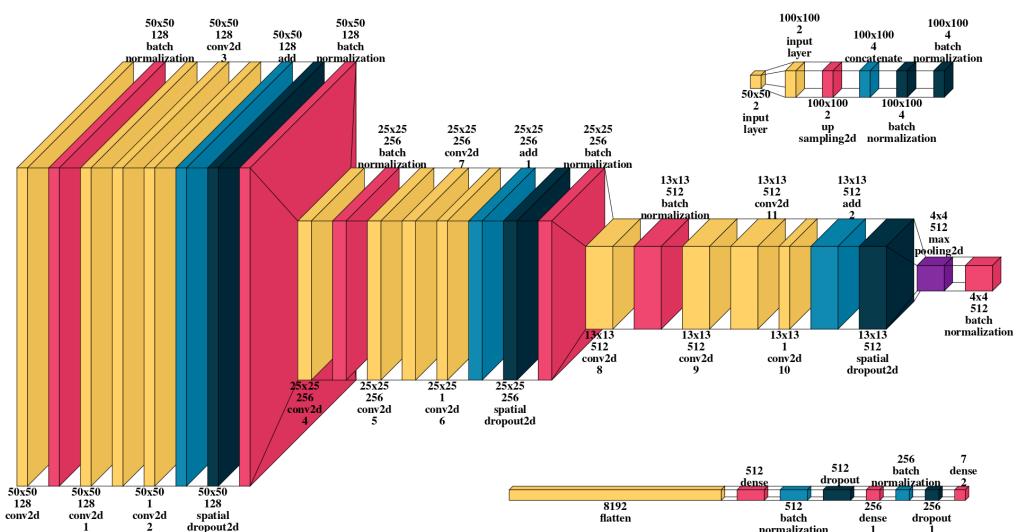


Figure 4.3: The model’s layered architecture is depicted. The top-right shows the input layers, the middle section displays the convolutional layers, and the bottom-right illustrates the fully-connected layers. The layered views were generated using VisualKeras ([Gavrikov \(2020\)](#)).

Chapter 5

Climate Data

5.1 ERA5 Dataset

The ECMWF Reanalysis v5 (ERA5) dataset ([Hersbach et al. \(2020\)](#)) was selected for correlating with tree genus classification due to its extensive and high-resolution climate data, which includes variables such as temperature, precipitation, and soil moisture. This dataset offers detailed historical weather information, capturing both spatial and temporal variations essential for understanding the environmental conditions that impact tree growth and distribution. By leveraging ERA5 data, this study aimed to explore how different climate factors might influence the abundance of various tree genera.

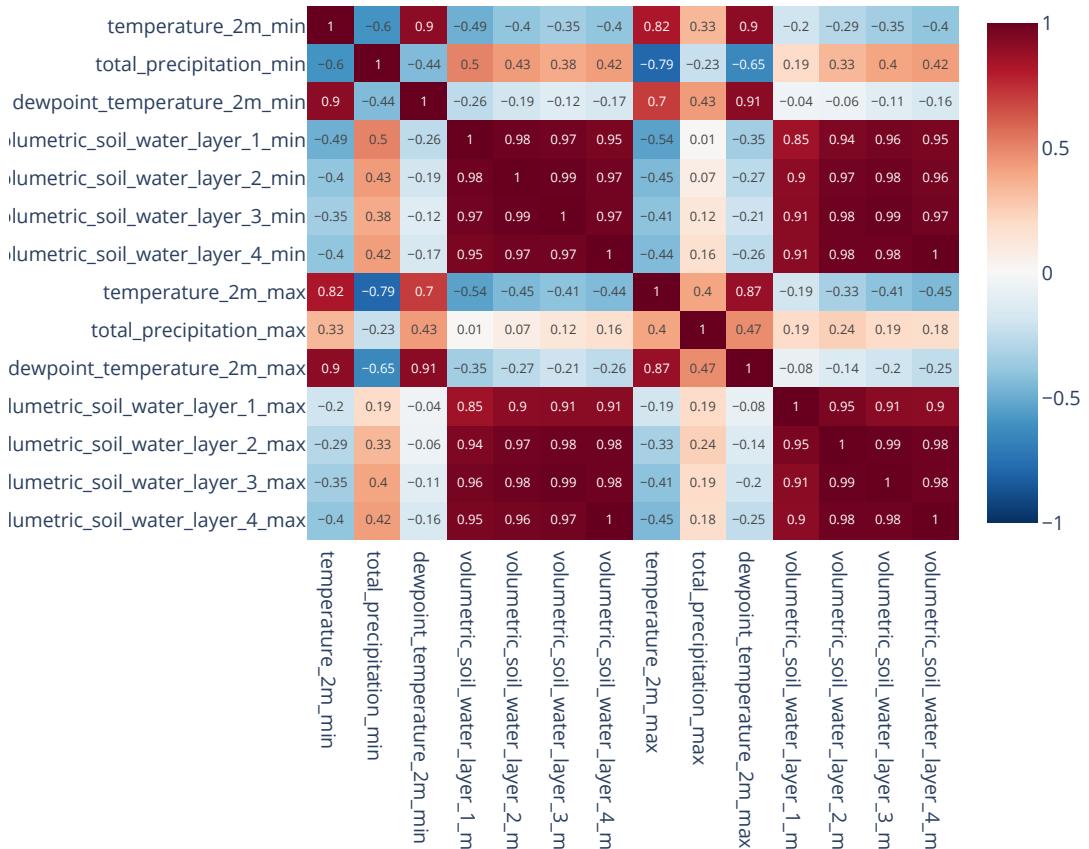


Figure 5.1: Correlations heatmap between various ERA5 variables. The data was downloaded from [ECMWF \(1950-Present\)](#), where detailed information is available.

5.2 Dataset Exploration

As the ERA5 dataset contains numerous meteorological variables, a selection was made based on domain knowledge, correlations (Fig. 5.1), and studies such as Toledo et al. (2011). Fig. 5.2 summarizes this selection, showing the median and quantiles for the aggregated dataset across all study locations. The figure illustrates the significant variability within Europe, with the top-left plot indicating a difference of approximately 50 K between the 1.0 quantile of minimum and maximum temperature measurements. This suggests that, although limited to Europe, this study explored a significant climate gradient.

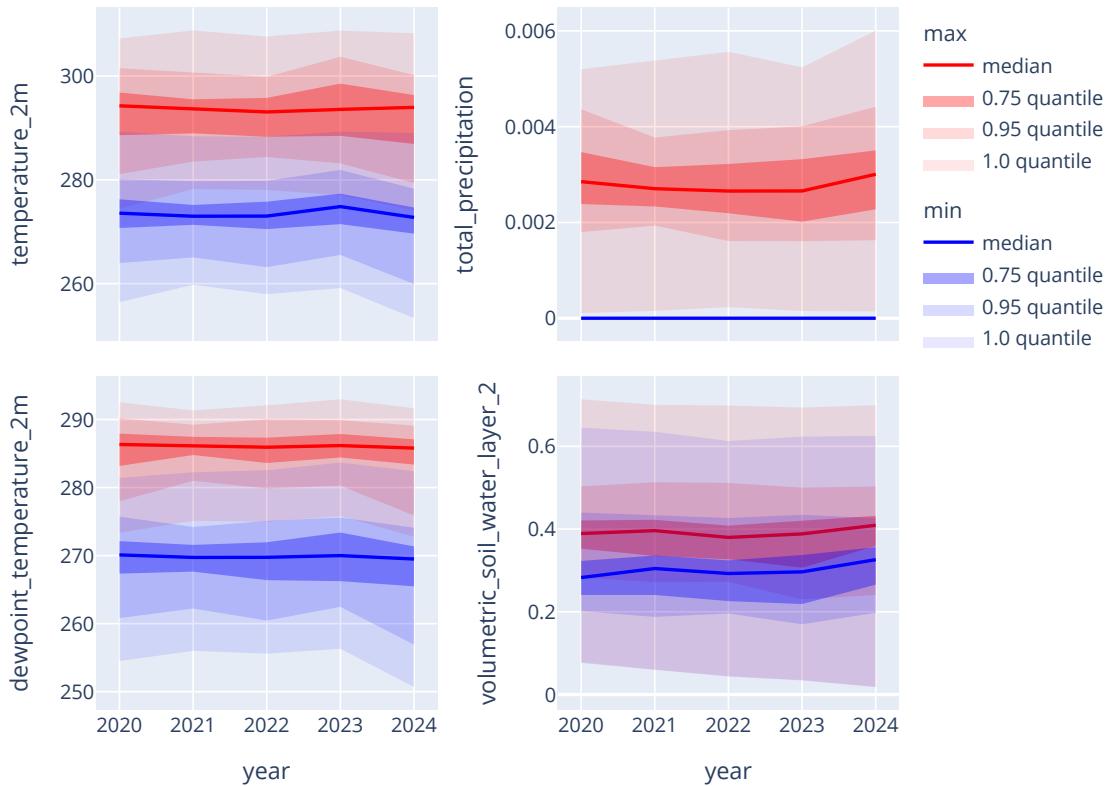


Figure 5.2: Median and quantiles of various monthly ERA5 variables by year.

Chapter 6

Results and Validation

6.1 Classification Validation

6.2 Change Map Correlations

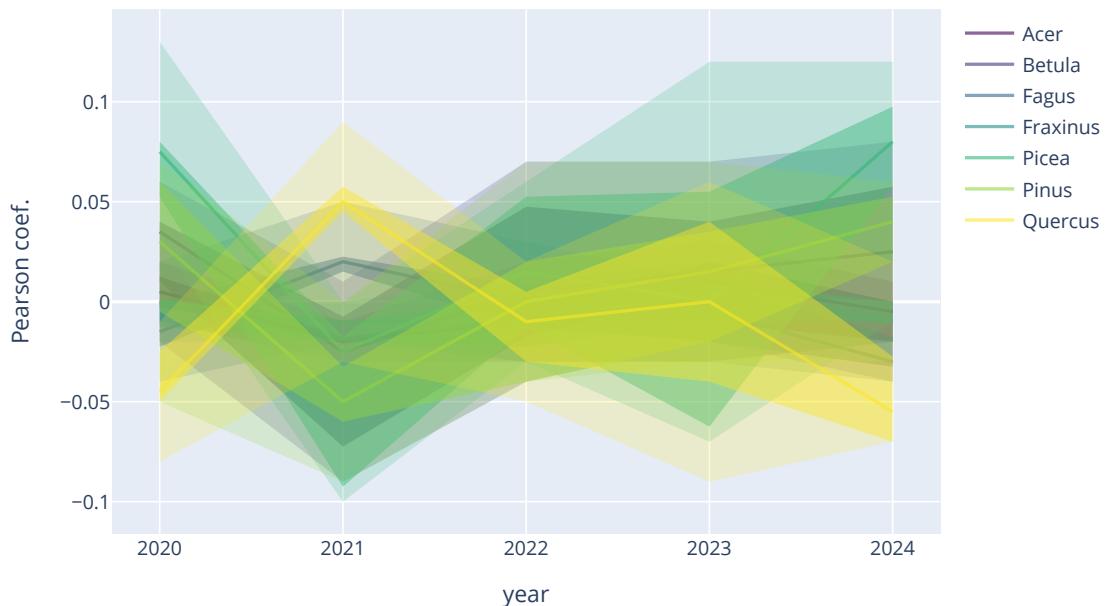


Figure 6.1: Median and quantiles of Pearson correlations between changes in tree genera maps predicted by classification and differences in meteorological conditions.

The representative medians of the variables shown in Fig. 5.2 were calculated for the years 2010 to 2017. The differences between these medians and those for each individual year from 2020 to 2024 were analyzed and correlated with the discrepancies between classification predictions and actual values. These results are summarized in Fig. 6.1. This figure suggests that, over a short-term period of 5 years, the 7 most prominent European tree genera were not significantly affected by climate change. However, this does not account for potential long-term impacts, where even small changes could have substantial effects on ecological dynamics. Additionally, this analysis is limited to Europe, a region that may experience less pronounced

climate changes compared to other parts of the world, as discussed in Hoegh-Guldberg et al. (2018). Furthermore, the study does not consider less frequent tree genera, which might be more sensitive to climate change and could exhibit different patterns of response.

6.3 Relationship Modeling

Unlike correlation, which measures the strength and direction of linear relationships between variables, relationship modeling with neural networks aims to uncover complex, non-linear interactions between variables. To further explore the relationships illustrated in Fig. 6.1, a narrow, fully-connected regression neural network was developed. This network used changes in meteorological variables as predictors to analyze how these changes relate to variations in tree genera maps.

The model was trained using the Huber loss function, which combines the advantages of Mean Absolute Error (MAE) and Mean Squared Error (MSE). Huber loss is particularly effective in scenarios with noisy data and outliers, as it provides a balance between the robustness of MAE and the sensitivity to errors of MSE.

R^2 was used as a primary metric because it directly measures the proportion of variance in the dependent variable that can be explained by the independent variables. This metric is particularly valuable for assessing how well the model captures the underlying relationships in the data. A higher R^2 value indicates a stronger relationship between the predictors and the target variable, offering insights into the effectiveness of the model in explaining variability.

MAE was chosen to evaluate the accuracy of the model's predictions. MAE provides an average of the absolute differences between the predicted and actual values, giving a clear indication of the magnitude of prediction errors. While MAE does not measure the strength of relationships directly, it helps assess how closely the model's predictions align with actual outcomes.

As illustrated in Fig. 6.2, the model exhibits a very low, and at times negative, R^2 value along with a high MAE. Additionally, a random sample of 50,000 residuals for the year 2024 do not display a discernible pattern, aside from clustering that reflects the original data being naturally grouped into true and false values. These results collectively indicate that the model fails to capture any significant relationship between the climate change maps and the changes in tree genus distributions. The poor performance metrics suggest that the predictors used do not meaningfully account for the variability in tree genus changes.

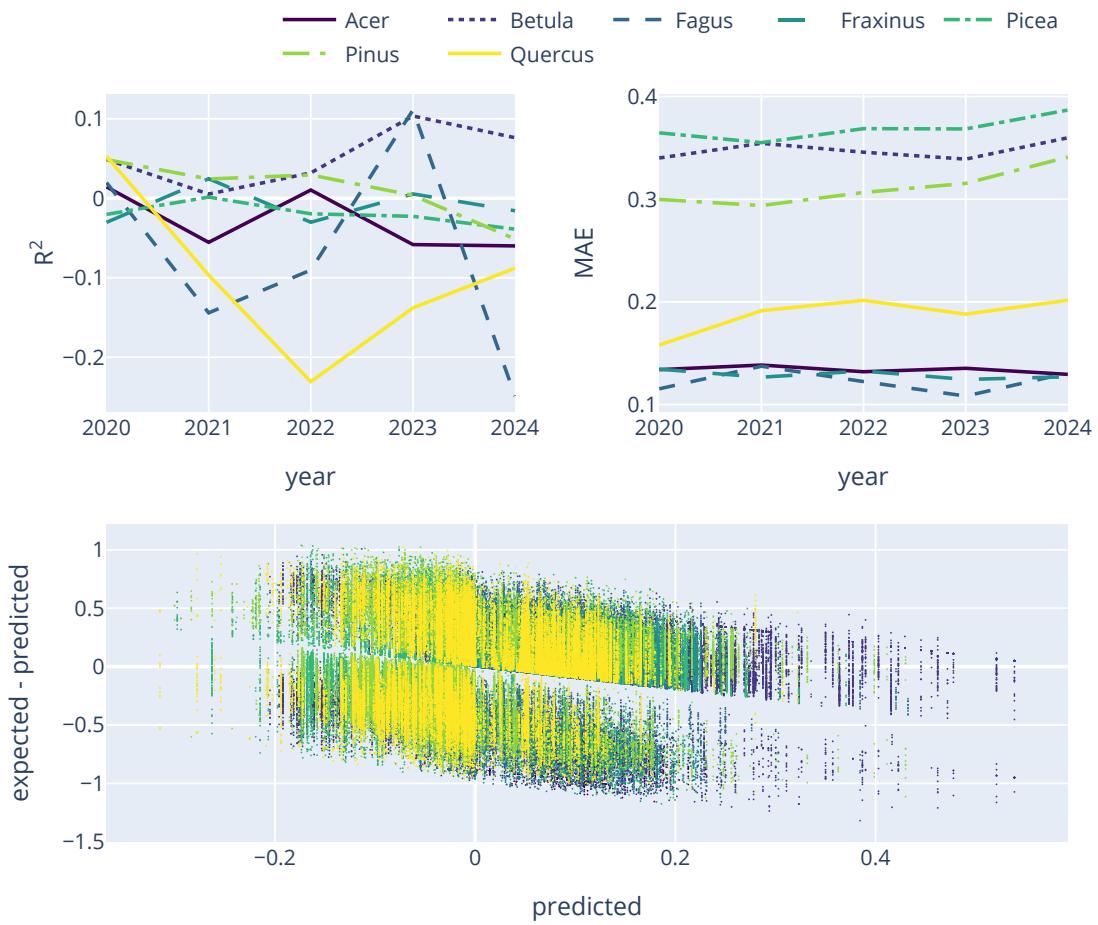


Figure 6.2: R^2 score (top-left), MAE (top-right), and residuals (bottom) for a narrow, fully-connected regression neural network, using meteorological change maps as predictors for changes in tree genus maps derived from tree classification.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

This study explored how various factors influence tree genus classification using Sentinel-2 imagery, different seasonal and spectral band combinations, and additional environmental datasets.

The analysis revealed that the classification model performed most effectively during summer and autumn, when the spectral signatures for tree genera were the clearest and most distinct. This is further supported by studies such as [Bolyn et al. \(2022\)](#), which focused on summer months. In contrast, performance dropped during spring and winter, likely due to less distinct spectral features and complications from environmental factors such as cloud cover and snow.

Among the Sentinel-2 bands, those in the NIR and SWIR ranges yielded better results. The combination of bands B3, B6, B8, and B11 emerged as the optimal choice, balancing classification accuracy with storage efficiency. This blend provided an effective trade-off between data richness and computational demands. Other studies have concluded differently: [He et al. \(2023\)](#) found that the RGB bands performed best, whereas [Xi et al. \(2021\)](#) found correlation between the best performing bands and the months used in the analysis.

Incorporating soil composition and elevation data offered only marginal improvements in model performance compared to using Sentinel-2 data alone despite soil composition alone showing a strong predictor value. The high-resolution spectral data from Sentinel-2 proved to be the most critical factor in achieving accurate classification results.

Further analysis using the ERA5 dataset revealed that short-term climate variations had minimal impact on tree genus classification over a 5-year period. This brief perspective may not fully capture long-term climate effects, which could significantly influence ecological dynamics. Additionally, focusing on Europe, a region currently experiencing relatively moderate climate changes, might limit the broader applicability of these findings.

Efforts to model the relationship between climate change and tree genus distributions using a regression neural network were unsuccessful. The model demonstrated poor performance, with low R^2 values and high MAE, indicating that the climate predictors did not effectively explain the variations in tree genus changes.

In conclusion, this research underscores the critical role of seasonal timing and spectral band selection in tree genus classification. It also highlights the limited impact of additional environmental data and short-term climate changes on model performance. While the current models can achieve high accuracy with well-chosen data and parameters, predicting the long-term impacts of climate change on tree distributions remains a complex and challenging task.

7.2 Future Directions

To achieve a more comprehensive understanding of the effects of climate change on tree genera, future research should prioritize long-term data analysis. For instance, the approach utilized in [Mehmood et al. \(2024\)](#) emphasizes the importance of examining extended time periods. This method involves incorporating broader climate datasets, enabling the capture of gradual shifts and potential delayed effects on tree distributions. While Sentinel-2 data offers high-resolution imagery suitable for short-term studies, datasets such as Landsat, which provide decades of historical data, are invaluable for contextualizing these changes over a longer timeline.

Expanding the geographic scope beyond Europe, as applied in [Pazos-Outón et al. \(2024\)](#), could yield more generalized insights. Studying regions experiencing more pronounced climate changes could reveal diverse patterns of interaction between climate variables and tree distributions at both the genus and species levels. This broader approach may uncover regional differences that contribute to a more nuanced understanding of climate impacts globally.

Moreover, future research should differentiate between commercial forests and long-term, natural forests. The change maps of these forest types may exhibit distinct patterns due to varying management practices and ecological dynamics. As highlighted by [Chauvier et al. \(2021\)](#), commercial forests, influenced by planned harvesting, replanting, and other management activities, often display predictable changes. In contrast, long-term, natural forests are subject to complex ecological processes such as natural disturbances, successional dynamics, and gradual environmental shifts. A comparative analysis of these forest types could reveal how different management strategies and natural processes influence tree genus distributions and their responses to climate change, thereby providing more nuanced and actionable findings.

Additionally, incorporating a wider range of environmental variables beyond soil composition and elevation could enhance the robustness of future models. Factors such as land cover changes, pest outbreaks, and other biotic interactions may significantly influence tree genus distributions. For example, [Krist et al. \(2014\)](#) illustrates the importance of considering variables that may not be immediately apparent, such as surface slope and frost periods, which could play critical roles in shaping tree distributions.

Finally, interdisciplinary collaboration with ecologists and botanists could greatly enrich future studies. Integrating ecological insights, such as those on competition and species-specific responses to environmental changes discussed in [Magalhães et al. \(2021\)](#), can refine model parameters and improve prediction accuracy. This holistic approach would allow for the incorporation of biological factors that directly affect tree distributions, ultimately leading to more precise and reliable outcomes.

References

- Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B. and Kleinschmit, B. (2023), 'TreeSatAI Benchmark Archive: a multi-sensor, multi-label dataset for tree species classification in remote sensing', *Earth System Science Data* **15**(2), 681–695.
URL: <https://essd.copernicus.org/articles/15/681/2023/>
- Bolyn, C., Lejeune, P., Michez, A. and Latte, N. (2022), 'Mapping tree species proportions from satellite imagery using spectral–spatial deep learning', *Remote Sensing of Environment* **280**, 113205.
URL: <https://www.sciencedirect.com/science/article/pii/S0034425722003145>
- Bonan, G. B. (2008), 'Forests and climate change: Forcings, feedbacks, and the climate benefits of forests', *Science* **320**(5882), 1444–1449.
- Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.
- Chauvier, Y., Thuiller, W., Brun, P., Lavergne, S., Descombes, P., Karger, D. N., Renaud, J. and Zimmermann, N. E. (2021), 'Influence of climate, soil, and land cover on plant species distribution in the european alps', *Ecological Monographs* **91**(2), e01433.
- Copernicus (2011-2015), 'Copernicus DEM Global and European Digital Elevation Model', <https://doi.org/10.5270/ESA-c5d3d65>.
- ECMWF (1950-Present), 'ERA5-Land Monthly Aggregated ECMWF Climate Reanalysis', https://developers.google.com/earth-engine/datasets/catalog/ECMWF ERA5_LAND_MONTHLY_AGGR.
- Food and Agriculture Organization of the United Nations (2020), *Global Forest Resources Assessment 2020: Main Report*, Food and Agriculture Organization of the United Nations, Rome.
- Gavrikov, P. (2020), 'visualkeras', <https://github.com/paulgavrikov/visualkeras>.
- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O. and Townshend, J. R. G. (2013), 'High-resolution global maps of 21st-century forest cover change', *Science* **342**(6160), 850–853.
- He, K., Zhang, X., Ren, S. and Sun, J. (2015), 'Deep residual learning for image recognition'.
- He, T., Zhou, H., Xu, C., Hu, J., Xue, X., Xu, L., Lou, X., Zeng, K. and Wang, Q. (2023), 'Deep learning in forest tree species classification using sentinel-2 on google earth engine: A case study of qingyuan county', *Sustainability* **15**(3).

- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellán, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020), 'The ERA5 global reanalysis', *Quarterly Journal of the Royal Meteorological Society* **146**(730), 1999–2049.
- Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A., Djalante, R., Ebi, K., Engelbrecht, F., Guiot, J., Hijiodka, Y., Mehrotra, S., Payne, A., Seneviratne, S., Thomas, A., Warren, R., Zhou, G. and Tschakert, P. (2018), *Impacts of 1.5°C global warming on natural and human systems*, IPCC.
- Krist, F. J., Ellenwood, J. R., Woods, M. E., McMahan, A. J., Cowardin, J. P., Ryerson, D. E., Sapiro, F. J., Zweifler, M. O. and Romero, S. A. (2014), '2013-2027 national insect and disease forest risk assessment.', *Fort Collins, CO: US Forest Service, Forest Health Technology and Enterprise Team*.
- Lefsky, M. A., Harding, D., Cohen, W. B., Parker, G. G. and Walton, J. (2002), 'Lidar remote sensing for ecosystem studies', *BioScience* **52**(1), 19–30.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. and Talwalkar, A. (2018), 'Hyperband: A novel bandit-based approach to hyperparameter optimization', *Journal of Machine Learning Research* **18**(185), 1–52.
- Magalhães, J. G. d. S., Amoroso, M. M. and Larson, B. C. (2021), 'What evidence exists on the effects of competition on trees' responses to climate change? a systematic map protocol', *Environmental Evidence* **10**(1), 34.
- Mauri, A., Strona, G. and San-Miguel-Ayanz, J. (2017), 'EU-Forest, a high-resolution tree occurrence dataset for Europe', *Scientific Data* **4**(1), 160123.
- Mehmood, K., Anees, S. A., Muhammad, S., Hussain, K., Shahzad, F., Liu, Q., Ansari, M. J., Alharbi, S. A. and Khan, W. R. (2024), 'Analyzing vegetation health dynamics across seasons and regions through ndvi and climatic variables', *Scientific Reports* **14**(1), 11775.
- Pazos-Outón, L. M., Vasconcelos, C. N., Raichuk, A., Arnab, A., Morris, D. and Neumann, M. (2024), 'Planted: a dataset for planted forest identification from multi-satellite time series'.
- URL:** <https://arxiv.org/abs/2406.18554>
- Pettorelli, N., Duncan, A. and Williamson, D. (2016), 'Satellite remote sensing for applied ecologists: opportunities and challenges', *Journal of Applied Ecology* **53**(4), 927–936.
- Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E. and Rossiter, D. (2021), 'SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty', *SOIL* **7**(1), 217–240.
- Sexton, J. O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., Collins, K. M., Channan, S., DiMiceli, C. and Townshend, J. R. (2013), 'Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error', *International Journal of Digital Earth* **6**(5), 427–448.

- Toledo, M., Poorter, L., Peña-Claros, M., Alarcón, A., Balcázar, J., Leaño, C., Licona, J. C., Llanque, O., Vroomans, V., Zuidema, P. and Bongers, F. (2011), 'Climate is a stronger driver of tree and forest growth rates than soil and disturbance', *Journal of Ecology* **99**(1), 254–264.
- Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E. and Steininger, M. (2003), 'Remote sensing for biodiversity science and conservation', *Trends in Ecology & Evolution* **18**(6), 306–314.
- Vose, J. M. et al. (2018), Ch. 6: Forests, in 'Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II', U.S. Global Change Research Program, Washington, DC, chapter 6, p. 243.
- Watson, J. E. M., Evans, T., Venter, O., Williams, B., Tulloch, A., Stewart, C., Thompson, I., Ray, J. C., Murray, K., Salazar, A., McAlpine, C., Potapov, P., Walston, J., Robinson, J. G., Painter, M., Wilkie, D., Filardi, C., Laurance, W. F., Houghton, R. A., Maxwell, S., Grantham, H., Samper, C., Wang, S., Laestadius, L., Runting, R. K., Silva-Chávez, G. A., Ervin, J. and Lindenmayer, D. (2018), 'The exceptional value of intact forest ecosystems', *Nat. Ecol. Evol.* **2**(4), 599–610.
- Wessel, M., Brandmeier, M. and Tiede, D. (2018), 'Evaluation of different machine learning algorithms for scalable classification of tree types and tree species based on sentinel-2 data', *Remote Sensing* **10**(9).
- URL:** <https://www.mdpi.com/2072-4292/10/9/1419>
- Xi, Y., Ren, C., Tian, Q., Ren, Y., Dong, X. and Zhang, Z. (2021), 'Exploitation of time series sentinel-2 data and different machine learning algorithms for detailed tree species classification', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 7589–7603.
- Zheng, Y. and Wu, Z. (2019), 'Deep learning in remote sensing: A comprehensive review and list of resources', *ISPRS Journal of Photogrammetry and Remote Sensing* **152**, 287–309.