Investigating the Challenges faced by Convolutional Neural Networks in Tree Species Recognition

Background and Context

Monitoring forests is crucial for understanding the impacts of deforestation, habitat loss, and climate change, as well as informing sustainable forest management practices. Machine learning techniques, if applied to aerial and satellite imagery, could provide valuable insights for forest conservation efforts.

Previous work in this domain by Bayrak et al. (2023) has explored the use of machine learning, particularly deep learning techniques such as convolutional neural networks (CNNs), for analysing aerial imagery. However, the application of these models faces several issues: the large number of classes and potential class imbalances, the difficulty of distinguishing between visually similar tree species, and computational challenges due to the high-resolution nature of modern aerial images. Efficient model architectures and techniques are therefore essential for practical deployment.

This investigation used data derived from the TreeSatAI Benchmark Archive introduced by Ahlswede et al. (2023). It consists of 50,381 aerial images of forests located in Lower Saxony in Germany. Each image covers an area of 60 x 60 metres and is assigned a single label from a total of 20 species, which are further grouped into 10 forest management classes, and then grouped into three broad categories (Figure 1). The images are 304 × 304 pixels and encoded as four-channel PNG images. The PNG's alpha channel has been used to represent the near-IR light band in the original images.

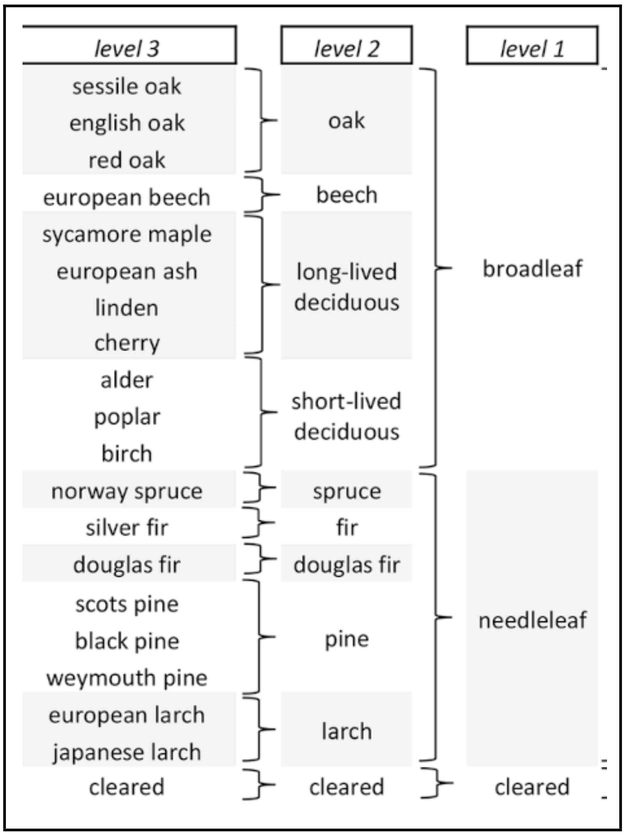


Figure 1:TreeSatAI Label Level Structure.

Within the dataset's Level 2 classes (the 10 forest management classes) a major class imbalance is clearly seen in the low number of samples for the Fir class and to a lesser extent the Douglas Fir and Larch classes (Figure 2).

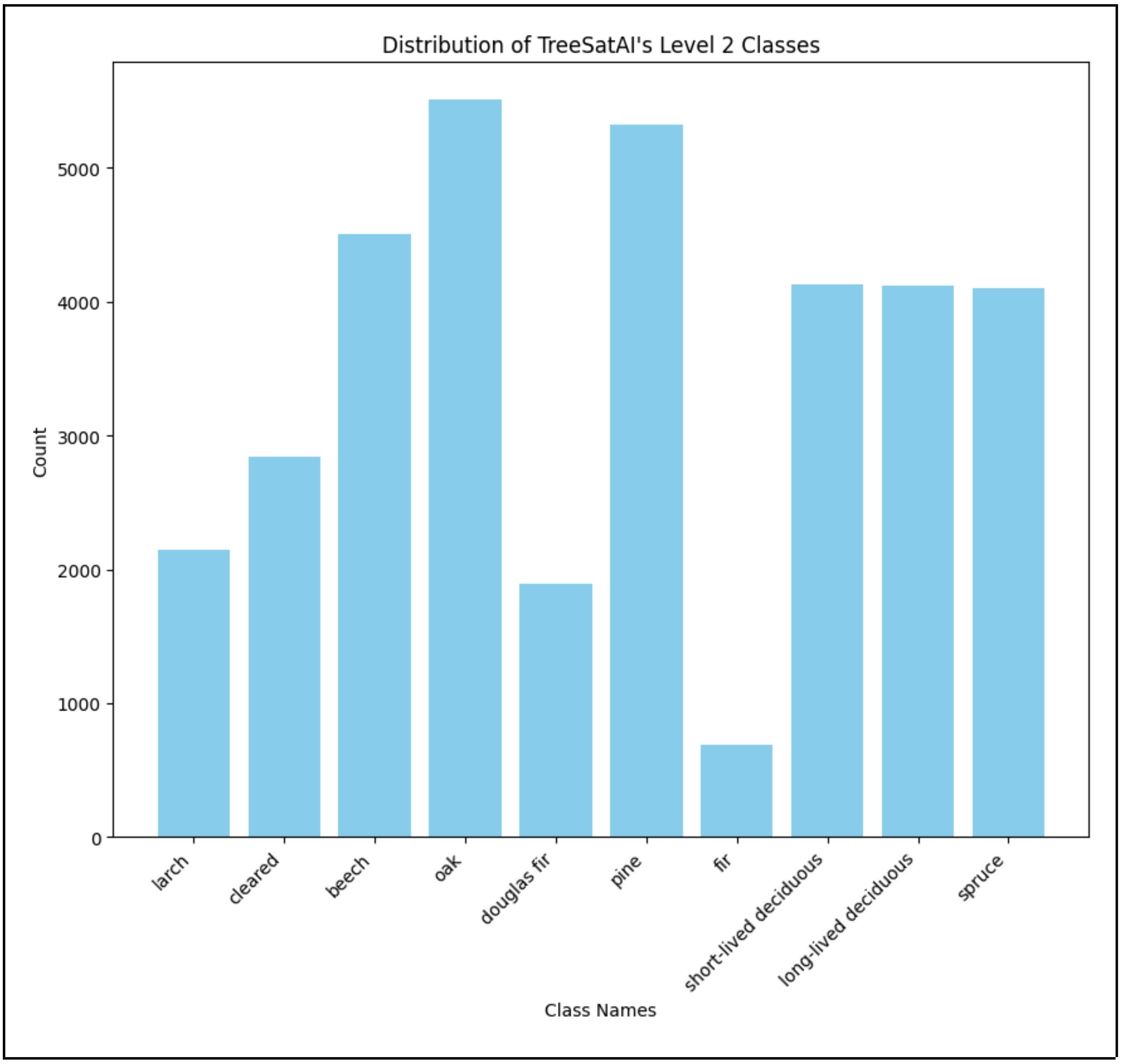


Figure 2: Distribution of TreeSatAI's Level 2 Classes

Aim and objectives

The primary aim of this investigation is to investigate the challenges faced in the development of convolutional neural network (CNNs) models for tree species recognition. Particularly the key issue of class imbalance due to the naturally imbalanced distribution of tree species seen in the real world TreeSatAI dataset.

To achieve this aim, the following objectives were established:

1. Investigate the impact of varying classification granularity on model performance, taking note of how reducing granularity in this case reduces the tree class imbalance.
2. Investigate the impact of data augmentation techniques, such as random translations, to increase the diversity of the training data and mitigate overfitting, particularly in the presence of limited training samples for certain classes.
3. Investigate the impact of proportional weighting during model training to attempt to mitigate the potential biases introduced by the class imbalance.
4. Investigate the impact of increasing model depth to assess the potential benefits of capturing more complex features.

Methods

To achieve these aims, five convolutional neural network (CNN) models using the TensorFlow and Keras libraries for Python were developed:

Model 1 was a base CNN model that the later models built upon and it acted as a point of comparison for the performance of all later models. This base CNN model consisted of three pairs of convolutional and pooling layers, then followed by a dropout layer, a flattening operation, and two final dense layers. An illustrative diagram of this base model architecture is presented in Figure 3.

Model 2 was a duplicate of the first and was the only one to be trained on the level 1 (needle, broadleaf, or cleared) labels, to specifically illustrate effects the label granularity has on model performance. All other models instead used the level 2 (the 10 forest management classes) labels.

For Model 3 data augmentation techniques were added to artificially increase the diversity of the training data and mitigate the risk of overfitting. Specifically, the input data was normalised and a random translation was added, shifting each input image by a random amount within ± 20% of the image dimensions. This model hoped to investigate how this alteration of the training dataset impacts the performance in the presence of the limited samples for certain classes.

Model 4 employed proportional weighting to address the class imbalance present in the Dataset with the Level 2 label's applied, where certain tree species were more prevalent than others, as demonstrated previously. By giving higher weights to the minority classes, the hope is the model will learn discriminative features for all classes, including the under-represented minority classes, potentially improving overall performance and reducing bias towards majority classes.

And finally, Model 5 had an augmented architecture. It incorporates additional convolutional layers, having 2 identical convolutional layers for each pooling layer. This model aimed to investigate the potential effects that increasing the model's convolutional filters and depth has on capturing more complex features. And if these changes are advantageous for distinguishing between multiple forest management classes.

Residual connections were considered as an alternative model approach in this investigation. However they introduce additional complexity to the network architecture, as they require careful design to integrate effectively, and as the base CNN model was not excessively deep, the training process should remain stable enough without being impacted by the vanishing gradient problem.

The TreeSatAI dataset was split into separate training (70% of the total dataset) and validation (10% of the total dataset). The validation subset being used to gauge overfitting during the training process. All these models: were trained for 10 epochs, had a batch size of 128, used Keras' default learning rate of 0.001, used categorical crossentropy as the loss function, and used the Adam optimization algorithm. These hyperparameters were kept consistent as control variables across models to allow for fair comparisons.

Once trained the models were evaluated on a test set, consisting of the remaining 20% of the TreeSatAI dataset. The evaluation process focused on three key metrics: accuracy, precision, and recall.

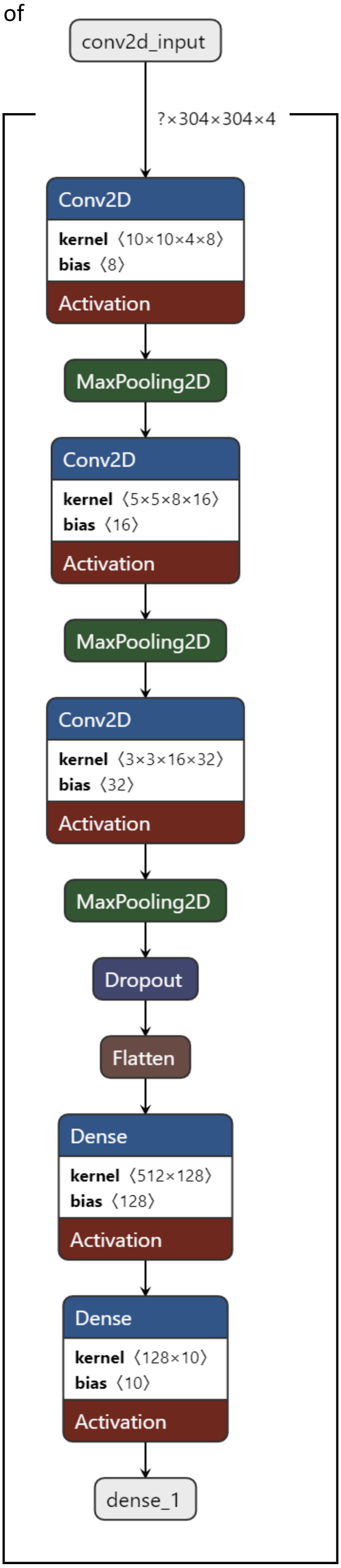


Figure 3:Base CNN model Structure.

As while accuracy provides an overall indication of the model's performance, it can be misleading in cases of class imbalance as the model's performance on minority classes may be underrepresented.

Precision instead quantifies the proportion of true-positive instances among all instances classified as positive by the model. In the context of tree species recognition in this investigation, high precision indicates that a model is making fewer false-positive predictions, which is crucial for reliable identification of tree species.

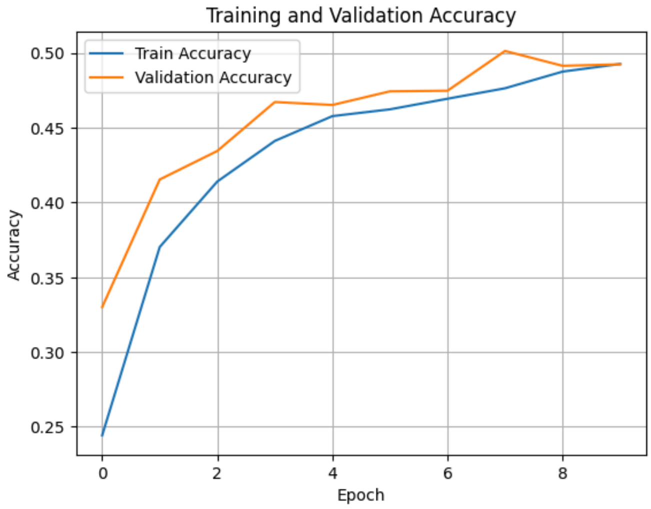
Recall meanwhile, measures the proportion of true positive instances that the model correctly identified out of all actual positive instances. For this investigation high recall ensures that the model is effectively detecting and identifying instances of each tree species, including those from minority classes.

Results

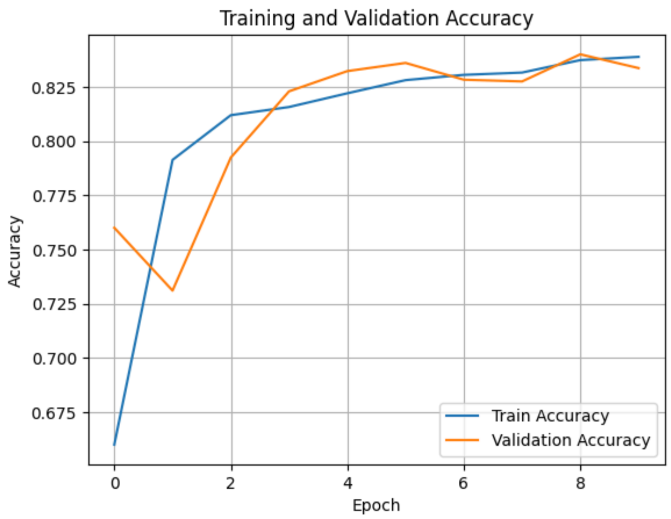
The training and validation performance graphs are presented below in Table 1.

None of the models showed strong signs of overfitting although Model 5 could potentially be beginning to overfit, though more training epochs would be needed to be sure. Model 2 shows a very quick performance increase but then levels off showing that the simple model is likely already reaching its permanence limit. The rest of the models still have relatively steep accuracy curves showing that they haven't reached their peak performance, Model 5 especially shows signs that it could do with training for more epochs, however if it does start to show signs of overfitting then more epochs won't be beneficial.

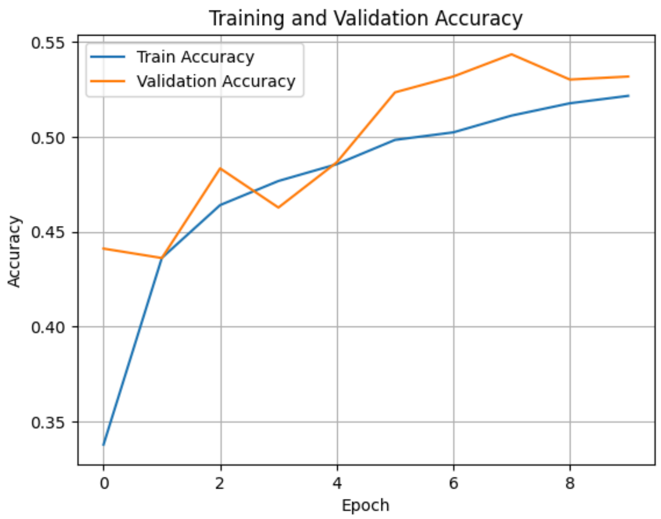
Table 1: CNN Models' Training Performance



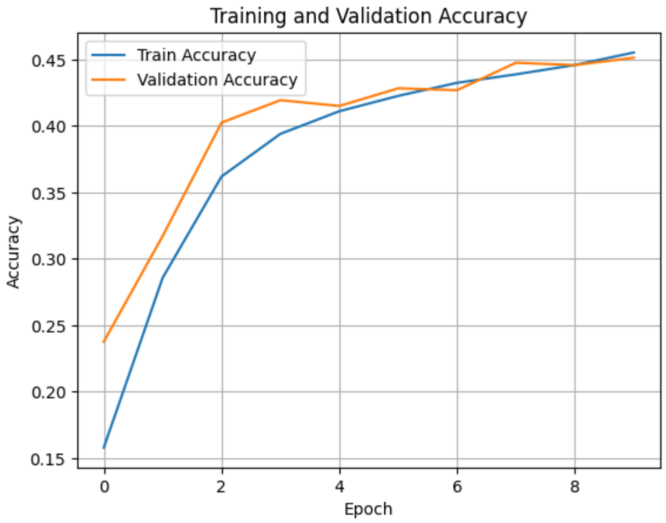
Model 1: Base Model with level 2 labels



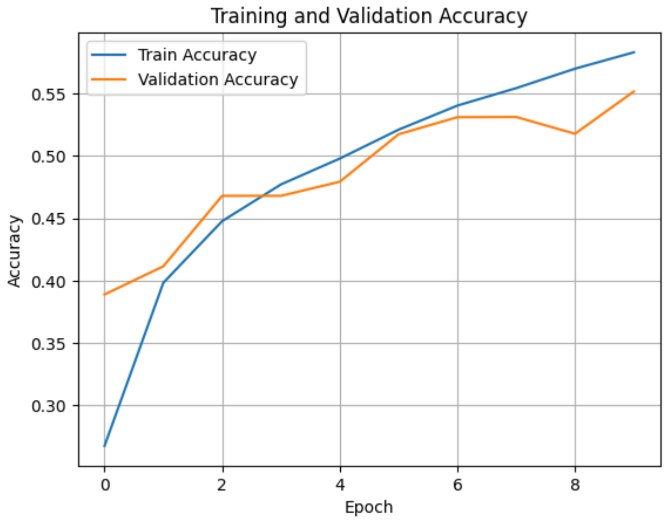
Model 2: Base Model with Level 1 labels



Model 3: Data Augmentation with level 2 labels



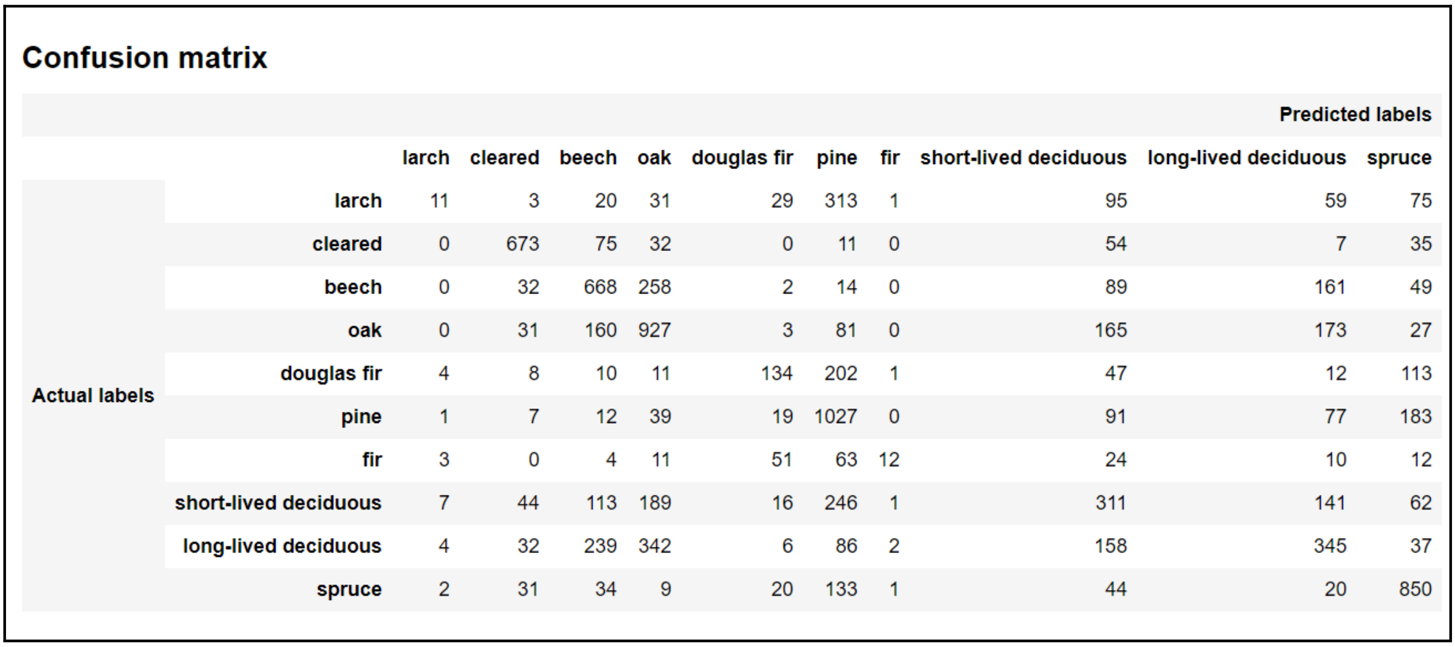
Model 4: Proportional Weighting with level 2 labels



Model 5: Extra Convolutional Layers with level 2 labels

Figure 4 shows the confusion matrix generated by Model 1 when evaluated on the test dataset. From the matrix it's clear how the model struggles on the minority classes: especially Fir, where it only correctly identifies 12 instances and instead misidentifies them as Pine or Douglas Fir, and Larch where again the model tends to mislabel them as pine. Pine being one of the most over represented classes in the dataset as seen in figure 2 previously.

The Cleared class, that being land cleared of trees, shows pretty good results despite being a slight minority class, this is presumably due to the greater visual distinction between instances of this class and all the others. Possibly this indicates that a model can still achieve good results for a minority class if there are more distinct differences between them and the majority class.



Confusion matrix

											Predicted labels
		larch	cleared	beech	oak	douglas fir	pine	fir	short-lived deciduous	long-lived deciduous	spruce
Actual labels	larch	11	3	20	31	29	313	1	95	59	75
	cleared	0	673	75	32	0	11	0	54	7	35
	beech	0	32	668	258	2	14	0	89	161	49
	oak	0	31	160	927	3	81	0	165	173	27
	douglas fir	4	8	10	11	134	202	1	47	12	113
	pine	1	7	12	39	19	1027	0	91	77	183
	fir	3	0	4	11	51	63	12	24	10	12
	short-lived deciduous	7	44	113	189	16	246	1	311	141	62
	long-lived deciduous	4	32	239	342	6	86	2	158	345	37
	spruce	2	31	34	9	20	133	1	44	20	850

Figure 4: Model 1 Confusion Matrix.

Table 2 has the performance results, listing the Accuracy, Precision, and Recall of each of the five models when evaluated on the testing set.

Table 2: Model Evaluation Results

Model	**Accuracy**	**Precision**	**Recall**
Model 1: Base Model with level 2 labels	0.492	0.687	0.297
Model 2: Base Model with Level 1 labels	**0.824**	0.833	0.817
Model 3: Data Augmentation with level 2 labels	**0.529**	0.710	**0.361**
Model 4: Proportional Weighting with level 2 labels	0.469	**0.707**	**0.219**
Model 5: Extra Convolutional Layers with level 2 labels	**0.554**	**0.724**	**0.375**

Evaluation of results

The base Model 1 achieved a relatively low accuracy of 0.492 and a very low recall of 0.297 when trained on the more granular level 2 labels. This suggests that distinguishing between the 10 management classes is a challenging task for the base model architecture. This same model architecture, when trained on the level 1 labels for Model 2, instead achieved a much higher accuracy of 0.824 and a balanced precision and recall, indicating better performance on this simpler classification task. This is likely due to the more balanced dataset and also how the 3 classes of Level 1 are possibly more visually distinct from each other. Based on this comparison, it is evident that the classification task becomes more challenging as the granularity increases from the level 1 (needle, broadleaf, or cleared) to the level 2 (10 forest management classes) labels.

Applying data augmentation techniques in Model 3 improved the model's performance compared to Model 1, increasing the accuracy to 0.529 and recall to 0.361. This improvement suggests that data augmentation helped in diversifying the training data and improving the model's generalisation capabilities, but leading to only slightly better recognition and prediction ability overall.

Implementing the proportional weighting technique in Model 4 did not yield the desired improvement. It instead resulted in a slight decrease in accuracy to 0.469 compared to Model 1. The precision improved to 0.707, indicating that the model became more conservative in its predictions and made fewer false-positive errors. However, the recall decreased significantly to 0.219, suggesting the model could now be biassed towards the minority classes, effectively underpredicting the majority classes.

The extra convolutional layers added to Model 5’s architecture led to the highest accuracy 0.554 and recall 0.375 among the models trained on level 2 labels. This improvement demonstrates the potential benefit of increasing the model's capacity and depth for capturing more complex features, which can be advantageous for the challenging task of distinguishing between the multiple forest management classes.

With its architectural modifications of additional convolutional layers over the base model, Model 5 showed the most promising results of the Level 2 models. However its practical use outside of this investigation is limited, as despite it achieving the highest accuracy, precision, and recall in this investigation, its results are still low especially with its accuracy of only 0.554.

Discussion of wider implications

It is crucial to acknowledge the potential biases that may arise from the data and CNN models used in this investigation. The TreeSatAI dataset is limited to forests in Lower Saxony, Germany. This geographic constraint could limit the usefulness of any models trained on it when presented with data from other regions with different environmental conditions. Furthermore, the extreme class imbalance present in the dataset can lead to biassed models that perform well on majority classes but poorly on minority classes especially if applied to environments with different tree species compositions.

The use of machine learning models in tree species recognition also raises ethical and social concerns. It may result in the neglect or underrepresentation of less common species, especially rare and endangered species, potentially impacting conservation and biodiversity efforts. And as these models could also potentially be used for decision making processes that impact land management, resource allocation, and environmental policies, there could be unintended consequences from the deployment of these models. Misclassification of tree species from biassed model outputs could lead to incorrect management actions or the prioritisation of certain species over others. Leading to impacts on ecosystems, local communities, and economic activities that are dependent on forest resources.

To mitigate the potential harm, relevant stakeholders, such as forest managers, conservationists, and local communities, should be involved in the development and deployment of these models. They can help identify potential ethical concerns and ensure that the models align with societal values and sustainability goals. Additionally, implementing robust monitoring and evaluation processes can help detect unintended consequences or biases in the model's outputs early in the development process.

Conclusions

The results from the models created for this study provide valuable insights into the challenges posed by class imbalance and the impact of various techniques, such as data augmentation, proportional weighting, and architectural modifications, on model performance. While some techniques showed promising improvements, the class imbalance issue remained a significant challenge, particularly in terms of recall for minority classes. The investigation also highlighted the trade-offs involved in addressing class imbalance by reducing the class granularity, as while it has the potential to drastically increase model performance, it comes at the cost of limiting the models' use case to only broad tree types.

There is a need for further exploration of more advanced techniques or a combination of strategies to effectively mitigate this issue: To further enhance CNN performance and address the class imbalance issue more effectively, additional techniques such as oversampling minority classes/ undersampling majority classes, dimensionality reduction, or the introduction of synthetic data to balance the classes could be explored in future work. Alternatively, expanding the geographic coverage of the training data to include more diverse regions and ecosystems could enhance the models' generalisation capabilities and facilitate their broader application.

Ultimately, the application of machine learning techniques in tree species recognition has the potential to significantly impact forest monitoring and conservation efforts. Accurate and efficient identification of tree species can inform sustainable forestry practices, support targeted conservation strategies, and contribute to climate change mitigation efforts. However, it is crucial to address the challenges posed by class imbalance and ensure that any developed models do not neglect or underrepresented minority tree species, as otherwise the models could end up harming the very forests they are made to help manage and protect.

Word count: 2447

References

Bayrak, O.C., Erdem, F. and Uzar, M. (2023) 'Deep Learning Based Aerial Imagery Classification for Tree Species Identification', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, pp. 471-476. Available at: https://doi.org/10.5194/isprs-archives-XLVIII-M-1-2023-471-2023

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, pp. 681-695. Available at: https://doi.org/10.5194/essd-15-681-2023