Investigation into using Machine Learning for Tree Species Identification

Background and Context

Following the development of high-resolution satellite images which have “rich spatial, color, and texture information” (Bayrak *et al*. 2023), the required quality has been achieved where they can be used to train convolution neural networks (CNNs) in various tasks.

This report uses a dataset of roughly 50,000 images of publicly managed forests spread across approximately 47,710km^2, collected from flight patterns between 2011 and 2020 over the Lower Saxony region of Germany ((Ahlswede *et al.* 2022). The dataset is licensed under the Creative Commons Attribution 4.0 License, which permits copying, sharing and adapting of the material for any purpose. Each image in the dataset is a 60-by-60 metre area of a forest, with an assigned class of tree species, including cleared areas, spread across three different levels of detail.

There has been an increased use of these images “in support of forest ecosystem monitoring and sustainable forest management” (Holzwarth et al., 2020), in addition to research into using CNNs for a variety of tree identification uses including “tree species discrimination, forest damage detection, and tree mortality mapping” (Bayrak *et al*. 2023).

Whilst there are benefits to the use of CNNs in this area allowing researchers to “minimize fieldwork, which is time-consuming and expensive” (Bayrak *et al*. 2023), there are also limitations to these models including the risks of imbalanced datasets leading to biased models.

Aims and Objectives

This report will investigate the effectiveness of CNNs in identifying tree species. Specifically, it will aim to identify which methods have the greatest impact on the performance of a CNN by comparing the performance of three slightly modified models to a baseline model. It will also assess whether data augmentation can improve the performance of CNNs by comparing the relevant metrics of previously created models on the regular dataset, and an augmented training dataset.

Methods

The baseline CNN model consisted of two convolution layers with eight filters, followed by a pooling layer. It then had two more convolution layers, these with four filters, before another pooling layer. Finally, the model has a dropout layer to help reduce the likelihood of overfitting, a flatten layer and two dense layers with 128 and 10 filters respectively. All layers used a ReLU activation, except for the final layer which uses a softmax activation function to present a distribution of the probability an image corresponds to each class. The model was compiled using an Adam optimiser with a learning rate of 0.001 and ‘categorical crossentropy’ for its loss function. Each model was trained for 25 epochs before its training and validation loss and accuracy were plotted. It was trained for 25 epochs as previous iterations of the baseline model were still improving after 20 epochs, but were beginning to overfit, so the number of epochs was increased to determine the optimal model whilst minimising the chance of overfitting. Finally, the model was evaluated on the test data.

This was repeated to create three additional models, each with a modification, to compare performance and find which method most improved model performance. One model was created with quadruple the number of filters in each convolution layer, as this would create more filter maps within each layer, which could improve performance by allowing a larger variety of patterns to be captured. A second model uses two additional convolution layers with four filters between pooling layers. This could improve performance as it would allow for capturing more abstract features in the images which could help with classification. The final model had an increased learning rate of 0.01, as this could allow the model to reach the optimal settings within the epochs, before overfitting occurs.

Some changes were made to the training dataset for these models. The image scale was reduced from 304x304 pixels to 152x152 pixels as when the models were trained on larger images, the kernel would become exhausted due to the dataset being too large. The level two classes were chosen for the dataset to allow for sufficient differentiation between tree species, without being too detailed as to make training the models difficult. Finally, as the dataset was imbalanced, featuring over 5000 images of pine trees but less than 1000 of fir trees, the dataset was weighted rateably by number in class to total images multiplied by 10. By weighting the dataset, accuracy should become a viable metric for analysis as the accuracy paradox is lessened.

To investigate whether data augmentation could improve the models, a new dataset was created by translating the training dataset with a height and width factor of 0.2 and rotating the data with a factor of 0.1. The models were then retrained on this augmented dataset to compare the models’ performance.

Results

Figures 1-4 below show training and validation, loss and accuracy for models trained on the weighted data. Figures 5-8 shows the same for models trained on the augmented data.

Figure 1 below shows the baseline model is improving after 25 epochs, although the rate of improvement decreased after around 8 epochs, for both accuracy and loss. The lines for training and validation loss are beginning to separate after around 20 epochs, suggesting overfitting is beginning.

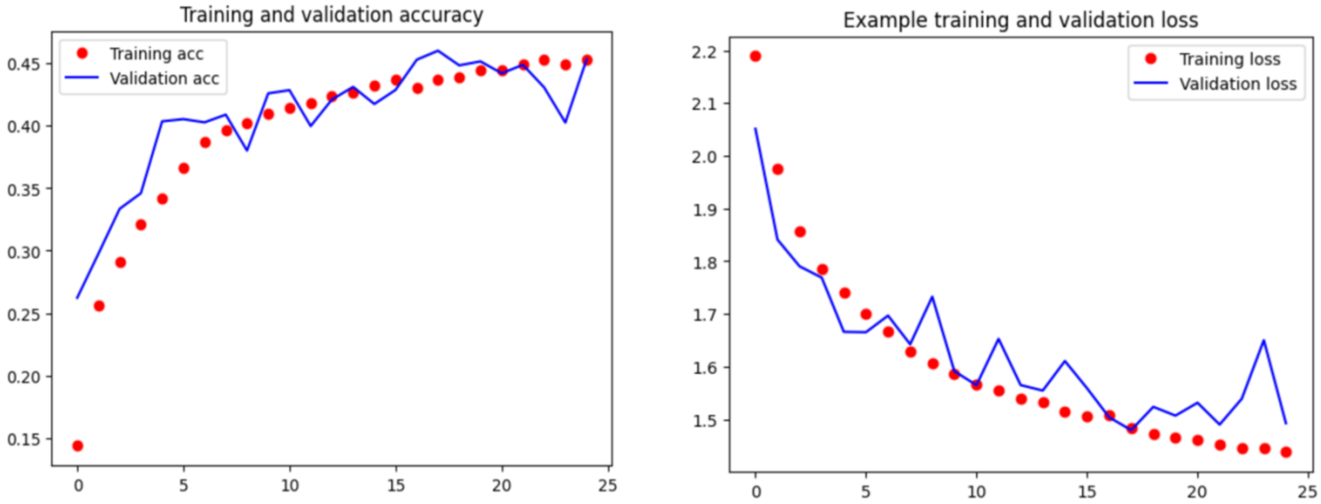

Figure 1: Baseline model with weighted training data

Figure 2 below shows the filters model is still improving in accuracy and loss after 25 epochs, however, improvement rate is greatly reduced after around 8 epochs. Additionally, the lines for training and validation accuracy are beginning to separate after around 20 epochs, and 15 epochs for loss, suggesting the model is starting to suffer from overfitting.

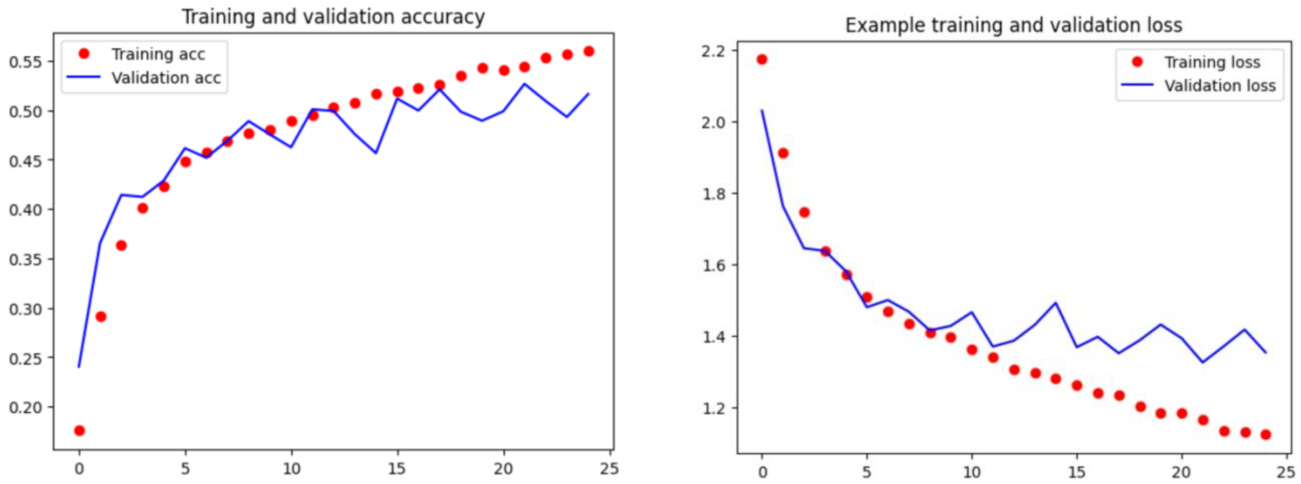

Figure 2: Filters model with weighted training data

Figure 3 shows that the layers model is still improving after 25 epochs, but the rate of improvement has reduced greatly with accuracy appearing to increase by only 0.01 in the final 5 epochs. The lines for validation and accuracy are still close suggesting overfitting has not occurred.

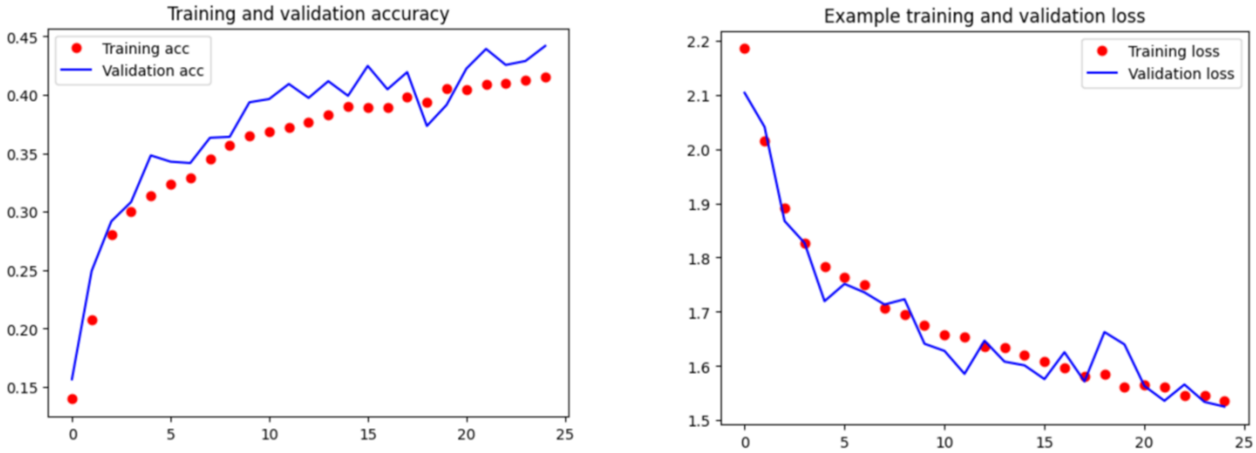

Figure 3: Layers model with weighted training data

Figure 4 shows that the higher learning rate model has not improved over its training. Both the accuracy and loss for training and validation have oscillated around a similar accuracy, suggesting that the model is overshooting the optimal values and is stuck at non-optimal points.

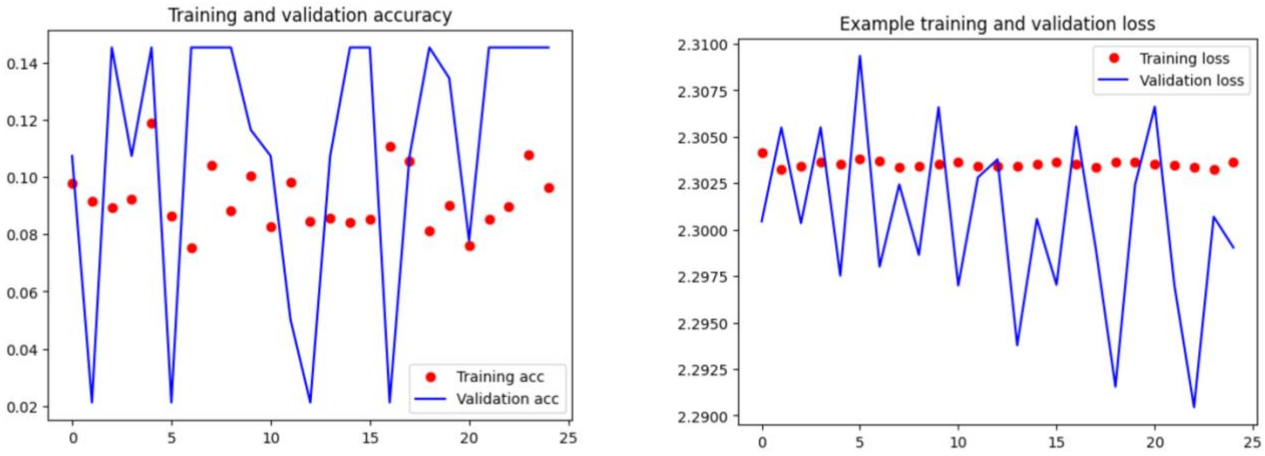

Figure 4: Higher learning rate model with weighted training data.

Figure 5 shows that when trained on the augmented data, the baseline model is still improving in accuracy and decreasing in loss over the 25 epochs. However, the rate of improvement is degrading rapidly. The lines for validation and training accuracy and loss are still close together, suggesting overfitting has not occurred.

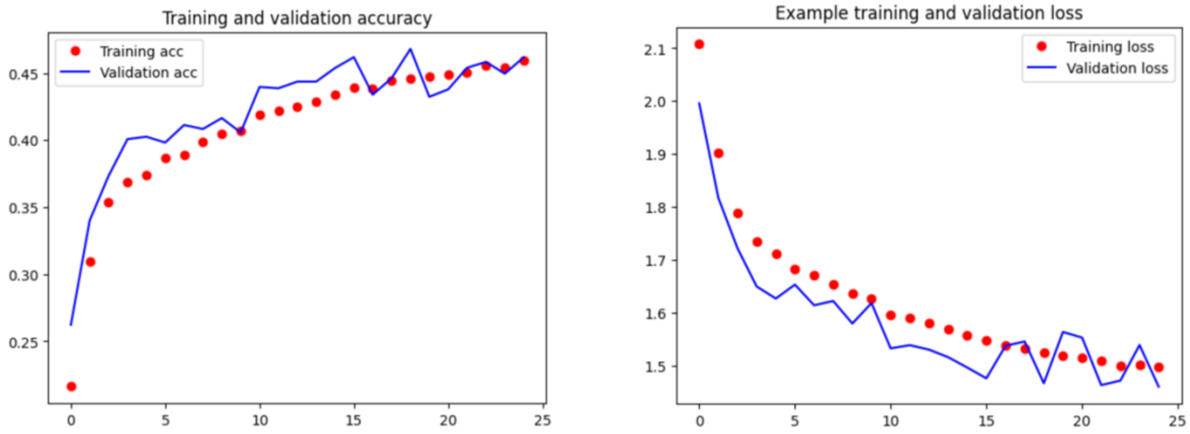

Figure 5: Baseline model with augmented training data.

Figure 6 shows that when trained on the augmented data, the filters model's accuracy is increasing, and the loss is decreasing for all 25 epochs, but the rate has reduced greatly. The lines for validation and training accuracy and loss are still relatively close suggesting overfitting has not occurred for this model.

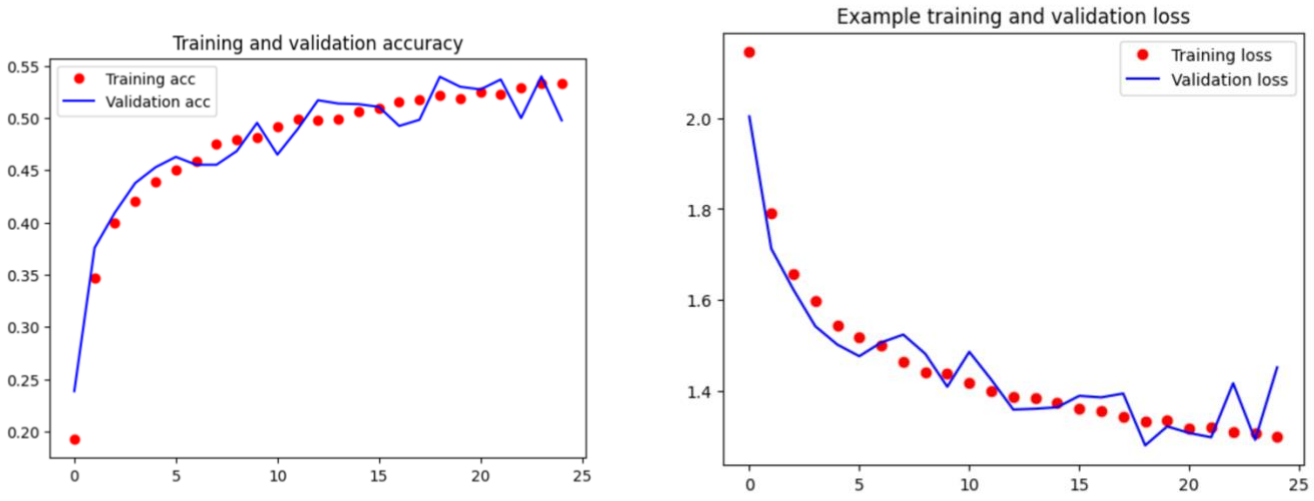

Figure 6: Filters model with augmented training data.

Figure 7 shows that when trained on the augmented data, the layers model's the accuracy for training and validation increased over the training period but appears to have stopped improving in the last 4 epochs. A similar trend is visible in the loss as the final epochs appear to oscillate around the same value. Furthermore, the loss and accuracy for the validation data fluctuates wildly, and is separated from the training lines, suggesting the model is overfitting.

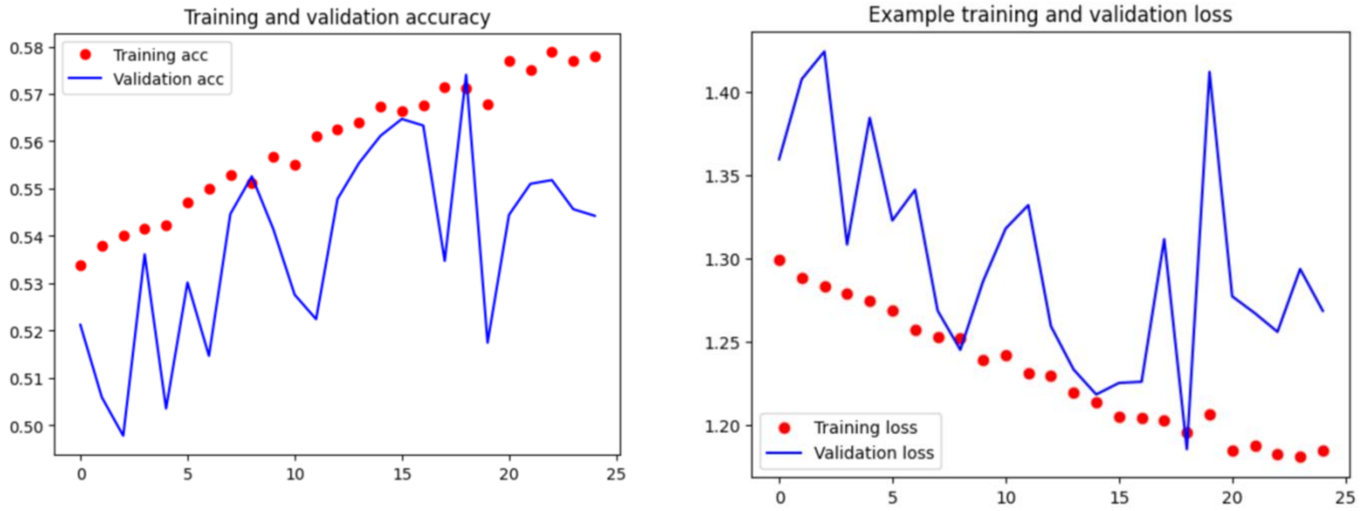

Figure 7: Layers model with augmented training data.

Finally, figure 8 shows that when trained on the augmented data, the model with a higher learning rate has similar training and validation accuracy and loss. Both appear to oscillate around the same value for the course of training suggesting the model is overshooting the optimal values and is stuck at non-optimal points.

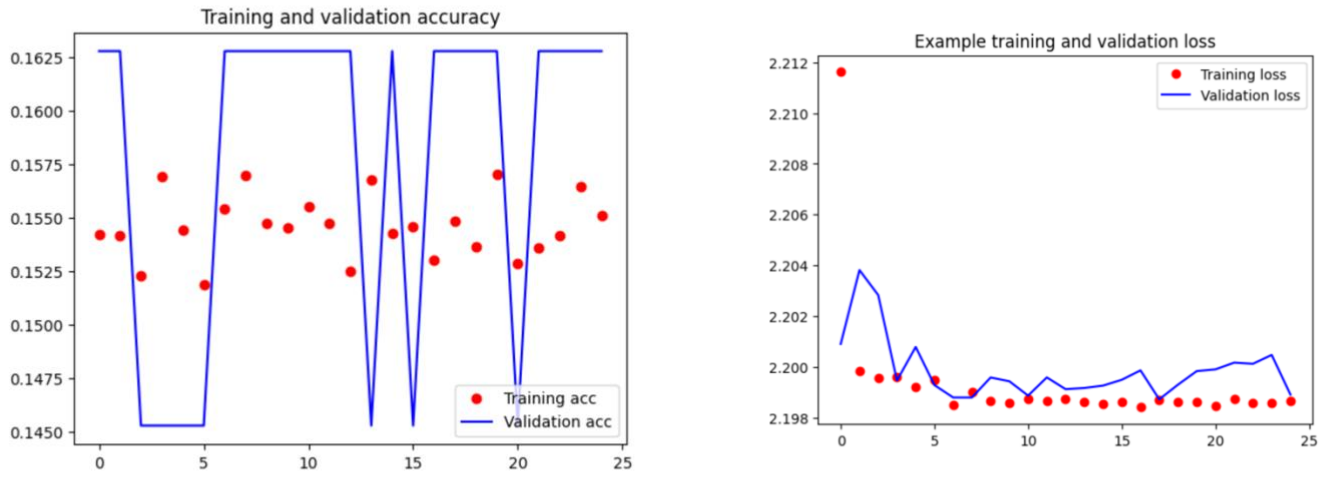

Figure 8: Higher learning rate model with augmented training data.

The accuracies, average precision and average recall for all trained models are shown in the table below. In these circumstances, accuracy is the percentage of species correctly identified, precision is the average percentage of the total tree species labelled as a specific tree that were that tree and recall is the average percentage of the total images of a given tree species that were correctly identified.

The model with the highest accuracy in both datasets is highlighted in bold and whether a model performed better on the weighted or augmented data is shown by underlining. The baseline model had an accuracy of 46%, with a precision of 45% and a recall of 47% when evaluated on the test data. The best performing model was the model with extra filters in each layer on both training datasets. It had an accuracy of 52% on the weighted training dataset and 49% on the augmented training dataset.

Model	Average	Training Data	
		Weighted	Augmented
Baseline	Accuracy	0.4566	0.4704
	Precision	0.4527	0.4796
	Recall	0.4680	0.3915
More Filters	Accuracy	**0.5174**	**0.4871**
	Precision	**0.4836**	**0.5336**
	Recall	**0.5160**	**0.4238**
More Layers	Accuracy	0.4447	0.098
	Precision	0.4134	0.0533
	Recall	0.432	0.0869
Higher Learning Rate	Accuracy	0.1445	0.1555
	Precision	0.0144	0.0156
	Recall	0.1	0.1

Table 1: Different CNNs and their accuracies, average precision and average recall

Evaluation of Results

The baseline model had an accuracy of 46% on weighted data, a precision of 45% and a recall of 47%. Its accuracy improved when trained on augmented data, increasing to 47% with a precision of 48%, however, the recall fell to 39%. This suggests that the model is performing better with most common species in the augmented data but is doing worse analysing rarer species.

Increasing the number of filters in each layer seems to have a positive effect, increasing accuracy by 6% on weighted data, and 2% the augmented training data. Increased precision and recall across both datasets is also observed. This suggests that increasing filters in each convolution layer will improve CNN performance.

Adding convolution layers to the model had little impact on the accuracy for the weighted data, only decreasing by 1%, but saw a larger fall in precision and recall by roughly 4%. There was a sharp decrease in performance of the model when trained on the augmented dataset with accuracy falling to 8%. This suggests that the model was overfitting to the augmented data and was unable to apply underlying trends from the training to test datasets. This suggests that additional convolution layers are not beneficial to a CNN. At best it had limited impact and, at worst caused performance to suffer greatly.

Increasing learning rate was also unsuccessful at improving performance. On both weighted and augmented datasets, the model had a poor accuracy of roughly 15%.

Based on recall being exactly 10%, it appears the model was only predicting one species for all images, possibly because the model was overshooting the optimal settings and was stuck at suboptimal settings.

Using augmented data over weighted data had mixed results on the performance of the models. Three of the models had an increase in precision when trained on the augmented data and two models saw an increase in accuracy. However, all the models had significantly worse recall. This suggests that the models were performing better on the more common tree species, but significantly worse on rarer tree species. Overall, augmented data did not improve the performance of the models, the large fall in recall suggesting that the false positives would be greatly increased. This is an unnecessary risk, and not worth the minor increases in accuracy and precision.

Overall, the best performing model was the filters model when trained on the weighted data as it had the highest accuracy and recall, indicating it will classify with greater consistency and fewer false positives. It also had the second-best precision, still high enough to be reasonable, given the strong accuracy and recall.

There were limitations in the models trained based on the lack of computational power when training models. Smaller images were required to avoid exhaustion errors which could negatively impact performance by making details less noticeable. Additionally, the models had to be simplified to avoid exhaustion errors thereby reducing performance.

Discussion of Wider Implications

Potential biases within the dataset include lack of geographic and tree species diversity. The dataset's source is a specific German region which may not be broadly representative, resulting in less accuracy when applied to other areas. Furthermore, the dataset contains roughly 30 species of trees whereas there are an estimated 73,300 species worldwide (Briggs, 2022). Accordingly, there will be areas with different tree species where the models would be ineffective.

There are also ethical concerns surrounding data collection. The data is from public areas but may now, or in future, include images from private areas, where landowner consent is required alongside maintaining privacy rights.
Additionally, care is required with misclassification risks. If models are solely relied on and a misclassification occurs, it could lead to decisions thought to be helpful which in practice prove harmful.

From a social perspective, the models automate previously manual processes potentially leading to redundancy for fieldworkers who typically perform these review roles. If successful, the models could lead to a rise in unemployment.

There may also be unintended consequences, for example, use by farming companies to locate land suitable farmland, or logging companies to source appropriate production areas. Both could lead to increased deforestation.

To minimise these potential downsides there should be strict regulations to ensure that the use of these models is managed and controlled to avoid misuse, and that there are experienced reviewers to test models for potentially harmful misclassifications.

Conclusions

In this report, a baseline CNN was created to investigate the effectiveness of CNNs for tree species identification. Three further models were created each looking at a potential method for improving performance. These models were then trained on weighted and augmented data to see if augmented data could improve performance.

The preferred model was the one with more filters in each convolution layer, trained on weighted data, because it had an accuracy of 51%. This was the only method that improved performance versus the baseline model. Using augmented data did not appear to be beneficial as it greatly reduced the recall of all models. Ultimately, the models created did not reach sufficient accuracy to be a reliable method for tree species identification.

Using machine learning to identify tree species could help to reduce expensive groundwork, as well as allowing for quicker identification of at-risk areas to help maintain forest health. It should be noted that the models trained in this study did not reach a high enough performance for this, as misclassification was too frequent for reliance.

The performance of these models could be improved in the future work with increased processing power which would allow for more detailed models and larger images. This could lead to the models achieving greater accuracy and producing more reliable results.

(2477 words)

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 Syst. Sci. Data,* 15, pp. 681–695, Available at: https://doi.org/10.5194/essd-15-681-2023 (Accessed 25th May 2024)

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 (Accessed 24th May 2024)

Briggs, H. (2022) *Earth has more tree species than we thought*. Available at: https://www.bbc.co.uk/news/science-environment-60198433 (Accessed 24th May 2024)

Holzwarth, S., Thonfeld, F., ; Abdullahi, S., Asam, S., Da Ponte Canova, E., Gessner, U., Huth, J., Kraus, T., Leutner, B., Kuenzer, C. (2020) 'Earth Observation Based Monitoring of Forests in Germany: A Review' *Remote sensing (Basel, Switzerland)*, 2020-11, Vol.12 (21), p.3570