Name: Dalevinder Singh

PI: F1067971

EMA

# Background and Context

Aerial imagery that covers approximately 47,710 km2 of publicly managed forests serves as a source of forest surveillance in the federal state of Lower Saxony, Germany. This imagery is used for the monitoring and identification of tree species, informing the careful planning and allotment of fieldwork (Ahlswede *et al*., 2022). However, the identification process is time consuming and requires specialised skills.

To alleviate this issue, as stated by Bayrak *et al*. (2023), convolutional neural networks (CNNs) have gained much attention for their proclivity for tree species classification from multi-channel imagery, as the one above. However, a drawback to using these models is the vast amount of curated data required for training. Labelling large datasets is laborious and costly, limiting the application of these models.

# Aim and Objectives

The aim of this investigation is to evaluate the extent to which the data structure and architecture of a CNN impact its ability to accurately differentiate aerial forest imagery.

The objectives are as follows:

* Develop a baseline CNN model that achieves a reasonable level of discriminatory accuracy of the TreeSatAI Benchmark Archive dataset. This model will act as a control for which all subsequent modifications will be evaluated against.
* Integrate residual links into the baseline CNN architecture to facilitate gradient flow to deeper layers of the system and assess the model’s performance.
* Employ class weighting to the baseline model to increase the model’s sensitivity to minority classes and assess the model’s performance.
* Artificially increase the diversity of the TreeSatAI Benchmark Archive dataset through data augmentation and assess the baseline models performance.
* Apply the baseline model to the TreeSatAI Benchmark Archive dataset at a different level of classification and assess how robust the model is to class complexity.

# Methods

## Data

The investigation will be conducted using the TreeSatAI Benchmark Archive dataset. There are 50,381 images each made up of 304 x 304 pixels covering four spectral bands, RGB and NIR (Ahlswede *et al*., 2022). There are three levels of classification: level 3 holds the species type (20 classes), level 2 groups the species into forest management categories (10 classes) and, level 1 holds the leaf type (3 classes).

## Baseline model

Data Preprocessing:

The dataset will be split into three categories that use the level 2 classifications: train, validation and test. Each will include 35,266, 5,038 and 10,077 examples, respectively. To enable faster processing and quicker convergence the data will be batched, with each batch including up to 128 examples. Like Bayrak *et al*. (2023), the images will be rescaled to 288 x 288 pixels to ease the computational cost of training and testing the model.

Model Structure:

Due to the nature of the classification problem the baseline model needs to be complex enough to differentiate between the 10 classes, hence, the deep CNN architecture below was chosen rather than a simpler shallower model.

| Block | Layer | Units | Filters | Kernel Size | Padding | Activation | Max-pooling Size | Drop-out Rate |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Conv1 | Convolution |  | 8 | 5x5 | same | ReLU |  |  |
| Convolution |  | 8 | 5x5 | same | ReLU |  |  |
| Max-pooling |  |  |  |  |  | 2x2 |  |
| Drop-out |  |  |  |  |  |  | 0.2 |
| Conv2 | Convolution |  | 16 | 3x3 | same | ReLU |  |  |
| Convolution |  | 16 | 3x3 | same | ReLU |  |  |
| Max-pooling |  |  |  |  |  | 2x2 |  |
| Drop-out |  |  |  |  |  |  | 0.2 |
| Conv3 | Convolution |  | 24 | 3x3 | same | ReLU |  |  |
| Convolution |  | 24 | 3x3 | same | ReLU |  |  |
| Max-pooling |  |  |  |  |  | 2x2 |  |
| Drop-out |  |  |  |  |  |  | 0.2 |
| Conv4 | Convolution |  | 32 | 3x3 | same | ReLU |  |  |
| Max-pooling |  |  |  |  |  | 2x2 |  |
| Drop-out |  |  |  |  |  |  | 0.2 |
| Dense | Flatten |  |  |  |  |  |  |  |
| Dense | 96 |  |  |  | ReLU |  |  |
| Drop-out |  |  |  |  |  |  | 0.5 |
| Dense | 10 |  |  |  | Softmax |  |  |

Table 1 Baseline Convolutional Neural Network Structure

This model is sequentially structured with 4 convolutional blocks followed by a dense block, allowing data to stream linearly layer by layer. The number of filters increases for each successive convolutional block and at the same time the kernel size is reduced. This structure efficiently facilitates feature extraction and the learning capacity of complex patterns, which is crucial for accurate classification.

ReLU activations are employed to introduce non-linearity into the system, which is necessary for learning complex patterns whilst being computationally efficient. The max-pooling layers reduce the spatial dimensionality of each output, in turn reducing the computational cost of the neural network and helping with issues of translational variance. The drop-out layers help prevent overfitting, and finally, a dense layer with 10 units corresponding to the number of classes employs a softmax activation to generate a probability distribution for classification.

Training:

As this is a multiclass problem, categorical cross entropy will be used as the loss function during training. ADAM will be used as the optimiser due to its quick convergence with a learning rate of 0.0005 and training will run for 25 epochs. Accuracy and loss will be used to evaluate overfitting during training by applying the model to the validation dataset at each epoch.

Evaluation:

The trained model will be evaluated against the test dataset to determine its performance against unseen data. The performance will be evaluated using the metrics accuracy and weighted precision, recall and F1 score due to the number of classes.

## Baseline Modifications

### Residual Links Model

Deep neural networks tend to suffer from the vanishing gradient problem in which gradient flow struggles to effectively backpropagate to the initial layers of the system. Leading to slower convergence and ineffective training (The Open University, 2024). As the baseline model consists of four sequential convolution blocks, residual links will be added from the end of block 1 to the end of block 3 and from the end of block 2 to the end of block 4. As the shape of the output changes throughout the system, the earlier block outputs need to be transformed to match the later block outputs. This will be achieved using two additional convolutional layers:

| Layer | Filters | Kernel Size | Strides | Notes |
| --- | --- | --- | --- | --- |
| Convolution | 24 | 1x1 | 4 | Transforms block 1s output to (36,36,24) |
| Convolution | 32 | 1x1 | 4 | Transforms block 2s output to (18,18,32) |

Table 2 Additional Convolutional Layers Facilitating Residual Links

### Weighted Classes Model

Class weighting will be added algorithmically to the baseline model. This is preferential to oversampling due to the reduced computational overhead thus speeding the training of the system. Each class weight will be calculated using the following formula:

### Data Augmentation Model

The baseline model will be trained using an augmented version of the TreeSatAI Benchmark Archive dataset. As satellite imagery can vary in terms of lighting and exposure, the data will be normalised to ensure that the pixel values are standardised allowing for better training and learning. Furthermore, it is likely that this type of imagery is subject to translational variance. Therefore, to ensure that the model is robust, the image's position will be randomly translated by 40%.

### Class Complexity Model

Due to the relative complex architecture of the baseline model, the more granular level 3 classifications will be used to understand how the performance of the model varies with the classification level.

# Results and Evaluation

## Training

### Baseline Model

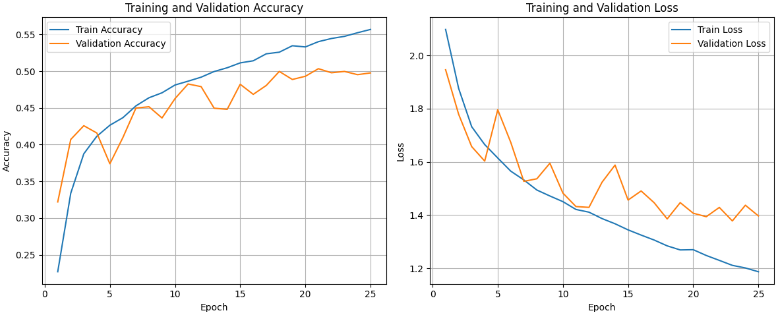


Figure 1 Baseline Model Training

Both datasets show the accuracy and loss respectively increase and decrease throughout the course of training, indicating that the model is learning. The validation metrics follow the training metrics closely suggesting that the model generalises well to unseen data. However, there are clear signs of overfitting as seen by the divergence in the validation’s accuracy and loss by epoch 25. Nonetheless this appears to be a good baseline model offering an accuracy much higher than random (this would be 10% for this classification level), with room for improvement.

### Residual Links Model

### 

Figure 2 Residual Links Model Training

This shows a vast deterioration in training when compared to the baseline model. The inclusion of residual links has exacerbated the overfitting issue already present in the baseline model, as seen through the large discrepancies in the validation’s accuracy and loss. This potentially suggests that the baseline model's architecture does not suffer from the vanishing gradient problem.

### Weighted Classes Model

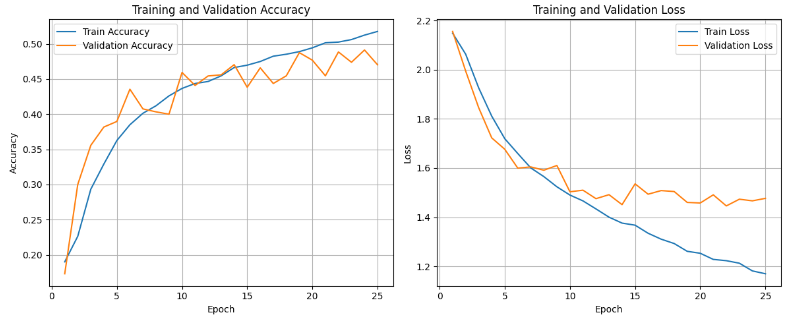


Figure 3 Weighted Classes Model

The training of this model is quite similar to the baseline model. Both models exhibit overfitting primarily signalled through the divergence of the validation loss at roughly epoch 15. However, a key distinction is seen in the size of the fluctuations, they are much less dramatic in this model. This suggests that the model is showing signs of sensitivity to minority classes and making less aggressive updates at each epoch.

### Data Augmentation Model

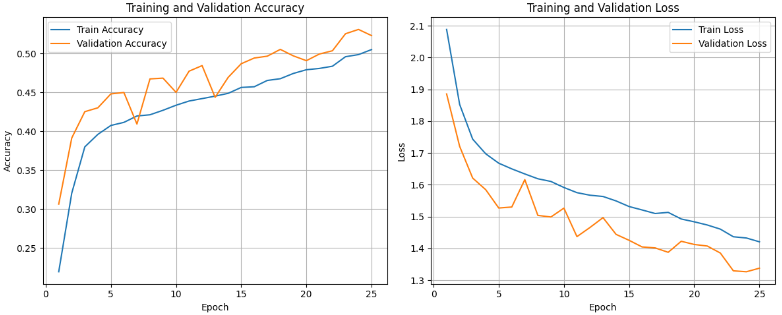


Figure 4 Data Augmentation Model Training

Compared to the baseline, all signs of overfitting appear to have been eliminated. Highlighting the significance of image standardisation and a robust varied dataset. Moreover, the trajectory of both the loss and accuracy strongly signal that this model would benefit from further training. Rather surprisingly, the validation metrics are greater than the training metrics, this is potentially linked to the augmentation of the training data, making it a much more challenging dataset to predict. Overall, in terms of training there appears to be a significant increase in performance.

### Class Complexity Model

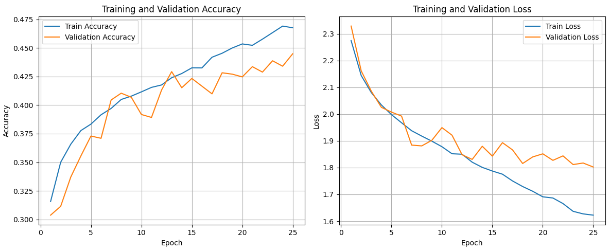


Figure 5 Class Complexity Model Training

Much like the baseline, this model also appears to be overfitting. However, the effect of this is somewhat reduced as indicated by the steeper gradient of the validation loss by epoch 25 and the reduced divergence in accuracies of both datatsets. This suggests that the baseline model is potentially too complex for the level 2 version of the TreeSatAI Benchmark Archive dataset.

## Performance

| Metric | Baseline | Residual Links | Weighted Classes | Data Augmentation | Class Complexity |
| --- | --- | --- | --- | --- | --- |
| Accuracy | 0.516 | 0.494 | 0.475 | 0.522 | 0.445 |
| Weighted Precision | 0.510 | 0.486 | 0.484 | 0.507 | 0.411 |
| Weighted Recall | 0.516 | 0.494 | 0.475 | 0.522 | 0.445 |
| Weighted F1 Score | 0.493 | 0.485 | 0.467 | 0.480 | 0.387 |

Table 3 Performance Metrics for Each Prescribed Model

The above details the performance metrics for the baseline and each of the prescribed models when applied to unseen test data. The baseline model's classification rate is roughly 50%. Its precision suggests that roughly 50% of its positive predictions are correct. The recall indicates that roughly 50% of all positives were identified, and finally, the F1 score suggests a reasonable balance between type 1 and 2 errors. This is not the most performant model but performs much better than average and provides a reasonable basis for comparison with room for improvement.

Relative to the baseline, the residual links model shows a reduced classification rate, less reliability in its positive predictions and identification of positive results, and is more prone to type 1 and 2 errors. This is unsurprising as evidenced by the large amount of overfitting that was exhibited during training. This suggests that the increased gradient flow exacerbated the already overfitted baseline and led to a poorer performing model.

The weighted classes model also shows poor performance relative to the baseline across all metrics. However, based on the reduced fluctuations seen in training the weighting appears to have limited how quickly the model converges as less attention is paid to the majority classes. It is likely that the lower performance metrics seen here are a function of this slower more stable convergence rather than the weighting negatively impacting the model.

Relative to the baseline, the data augmentation model can be concluded to be more robust. The higher accuracy indicates an increased classification rate across all classes. There is a slight reduction in precision meaning there is reduced reliability in the positive predictions made. However, the higher recall signals that more positives were identified. Whilst the F1 score suggests a poorer balance between type 1 and 2 errors, the training data clearly shows that this model has the capacity to be trained further without overfitting. This model is preferential to the others.

Across all metrics the class complexity model performs poorly relative to the baseline. It can therefore be concluded that the complexity of the classification problem impacts the model’s performance.

# Wider Implications

Deep neural networks in the context of forest management are a great conservational tool for monitoring and identification, however, they are not without their limitations. These models are inherently dependent on the quality of the data they are trained upon. For example, imbalanced data can lead to biases in predictions skewed towards the majority classes, potentially resulting in the misallocation of conservational efforts leading to the damage of local environments.

It is also important to understand how the data constrains the application of the technology. For example, the TreeSatAI Benchmark Archive dataset is based on imagery located in a particular region of Germany, a model trained on this data may not generalise well to other locations. Hence, using this model outside of this geographical region would be reckless and raises ethical concerns for the availability of this technology. Moreover, there is also the temporal constraint of this data, trees are subject to seasonal changes, the data must be robust enough to justify the use of this model across seasons.

To mitigate these risks, diverse datasets are required to broaden the geographical and seasonal scope that these models can be applied to. Data collection aimed to boost minority class representation would also be ideal. As an alternative, algorithmic strategies such as class weighting and oversampling could help to deal with inherent class imbalances (Bennasar, 2024). Finally, transparency and clear guidelines relating to how the model can and cannot be used should be provided to users and organisations of this technology.

# Conclusions

Although the baseline model exhibited signs of overfitting, the training and performance metrics provided a suitable control for which the prescribed modifications could be assessed against. The biggest increase in performance to the baseline was through the data augmentations model. Moreover, by simply changing the classification task, the performance of the baseline reduced significantly. Together, this illustrates the profound link between data robustness and structure and a model's performance.

The inclusion of the residual links to the baseline led to increased overfitting and generally a poorer performing model across all metrics. This sends a clear signal highlighting how the inherent architecture of a model influences its performance. Finally, the weighted classes model was seen to stabilise the model during training, leading to a slower convergence. This demonstrates how hyperparameters impact the training and performance of a model.

There is some doubt relating to the number of residual links used and their placement within the baseline. A further investigation could be to address this uncertainty. Further, it would be beneficial to understand how well these techniques complement one and other in increasing performance. For example, a model utilising data augmentations and weighting may lead to higher accuracies.

To conclude, machine learning is a supplementary tool to be used in the context of forest management and conservation. However, the implications of using this technology need to be carefully understood, monitored, and communicated to the relevant entities to prevent environmental and social harm.

# References

Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B. and Kleinschmit, B. (2022) ‘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://essd.copernicus.org/articles/15/681/2023/#section2 (Accessed: 10 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://isprs-archives.copernicus.org/articles/XLVIII-M-1-2023/471/2023/ (Accessed: 10 May 2024)

Bennasar, M. (2024) ‘Block 6: Data handling’. TM358: Machine learning & artificial intelligence. Available at: https://learn2.open.ac.uk/mod/oucontent/view.php?id=2217059 (Accessed: 7 May 2024)

The Open University (2024) ‘Block 2: Image recognition with CNNs’. TM358: Machine learning & artificial intelligence. Available at: https://learn2.open.ac.uk/mod/oucontent/view.php?id=2152484 (Accessed: 8 May 2024)

Word Count: 2419