Assessing Applications of Self-Supervised Learning for Tree Species Classification from LiDAR Point Clouds

Jake Pott[1], Baoxin Hu[2,*], F. Wayne Bell[3]

[1] Dept. of Earth and Space Science and Engineering, York University, Toronto, Canada – jakepott@yorku.ca
[2] Dept. of Earth and Space Science and Engineering, York University, Toronto, Canada – baoxin@yorku.ca
[3] Forest Ecology and Silviculture, Ontario Forest Research Institute,
1235 Queen St. East, Sault Ste. Marie, Ontario, P6A 3W9 – wayne.bell@ontario.ca
[*] Corresponding Author – Baoxin Hu – baoxin@yorku.ca

Keywords: LiDAR, PointMAE, PointNet, Species Classification, Self-Supervised Learning

Abstract

Individual tree species classification from LiDAR (Light Detection And Ranging) point clouds has significant potential to support forest inventory and management, yet remains challenging due to complex three-dimensional canopy structures and the limited availability of labelled ground truth data. This study investigates self-supervised learning for tree species classification from LiDAR point clouds by comparing the PointMAE, a masked autoencoder-based model, with two supervised baselines, PointNet and PointNet++. Using the FOR-species20k dataset, two experiments were conducted: a 33-species classification and a 6-species classification, each evaluated with point cloud sizes of 2048 and 8192 points. Using 2048 points, the PointMAE achieved the highest overall accuracy in both experiments (0.67 and 0.89 respectively), outperforming PointNet++ (0.63 and 0.84) and PointNet (0.39 and 0.75). Across all models, performance decreased when using 8192 points, indicating sensitivity to point cloud density and sampling. Per-species analysis showed that coniferous species with distinctive crown geometries were the easiest to classify, while broadleaf species with similar crown forms, particularly *Carpinus betulus*, were the most challenging. These results show that self-supervised pretraining can improve classification accuracy over fully supervised approaches, highlighting its value for forestry applications where labelled data are limited.

1. Introduction

Accurate individual tree species classification is crucial for sustainable forest inventory management and forest biomass estimation. Advances in remote sensing, particularly Light Detection and Ranging (LiDAR), enable three-dimensional (3D) analysis of forest structure through point clouds. At the individual-tree level, LiDAR point clouds capture fine-scale crown geometry and vertical structure, which are highly informative for tree species differentiation (Li et al., 2025). However, the rich structural detail captured by LiDAR point clouds also introduces substantial intra- and inter-species variability, making reliable tree species classification challenging in structurally complex forest canopies (Ma et al., 2024).

Recent developments in deep learning, such as point-based and transformer-based models, have led to increasingly accurate results for pattern recognition tasks, including tree species identification. Two types of models are particularly relevant in this context. Point-based models, such as PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) operate directly on 3D LiDAR point clouds, enabling the modelling of local geometric structure through neighbourhood aggregation. Transformer-based models have also garnered attention due to the flexibility of their self-attention mechanisms and strong performance across a range of domains, including natural language processing (Vaswani et al., 2017) and computer vision (Dosovitskiy et al., 2020). However, despite their performance gains, deep learning models require large labelled datasets to perform optimally, posing a major challenge given the time- and cost-intensive nature of collecting ground-truth species labels (Wang et al., 2024).

In response to the limited availability of labelled data, alternative learning paradigms that reduce reliance on manual annotation warrant consideration. Unsupervised learning addresses label scarcity by discovering intrinsic patterns in unlabelled data; however, it typically lacks task-oriented supervisory signals required for effective downstream classification. In contrast, self-supervised learning (SSL) introduces structured supervision through automatically generated pretext tasks, enabling the learning of transferable feature representations without human-annotated labels (Rani et al., 2023). As a result, a self-supervised learning framework is adopted in this study to reduce reliance on manual species annotations while learning transferable feature representations from LiDAR point clouds.

Within this self-supervised learning paradigm, Masked Autoencoders (MAEs) have demonstrated strong performance in natural language processing and computer vision by learning rich feature representations through the reconstruction of masked input data (He et al., 2021; Baevski et al., 2022). MAEs have recently been extended to 3D point clouds through models such as PointMAE (Pang et al., 2022). The PointMAE divides a point cloud into patches, masks a large portion of them, then trains the model to reconstruct the missing geometry from features learnt through the visible patches. This self-supervised pretraining enables the model to learn meaningful structural and geometric features without requiring manually-annotated species labels, directly addressing the challenge of limited ground truth available in forestry applications. Following pretraining, the encoder can be finetuned for downstream classification using a substantially smaller set of labelled samples.

This study evaluates the potential of self-supervised learning for

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-225-2026 | © Author(s) 2026. CC BY 4.0 License.

tree species classification from LiDAR point clouds by comparing the PointMAE against two supervised baselines, PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b). Using the FOR-species20k benchmarking dataset (Puliti et al., 2024), two experiments are conducted: a 33-species classification evaluated through an online benchmark, and a 6-species classification evaluated locally. In addition, the effect of point cloud size is examined by testing samples with 2048 and 8192 points. The main contributions of this study are as follows:

1. A systematic evaluation of self-supervised learning is presented for individual tree species classification from LiDAR point clouds, focusing on the PointMAE model and comparing self-supervised pretraining with fully supervised point-based baselines (PointNet and PointNet++) on the large and heterogeneous FOR-species20k dataset.
2. The effects of point cloud size and species-level structural variability on classification performance are analysed.
3. The practical implications of self-supervised learning for forestry applications where labelled data are limited are discussed.

2. Methods

2.1 Dataset

The dataset used for this study is the FOR-species20k, a large-scale publicly available benchmarking dataset that consists of individual tree point clouds and is designed for tree species classification (Puliti et al., 2024). The dataset consists of over 20,000 tree point clouds across 33 species, with data samples being captured through terrestrial (TLS), mobile (MLS), and UAV laser scanning (ULS). It is important to note that the FOR-species20k does not apply harmonisation or normalisation across samples from different scanning platforms. The FOR-species20k dataset combines 25 different datasets across various European, North American, and Australian forests, covering different dominant genera such as *Pinus*, *Fagus*, *Picea*, and *Carpinus*. Within the dominant species, the dataset covers each developmental stage from young saplings to mature trees (Puliti et al., 2024). The dataset is split into training and testing sets with a 90:10 ratio, with the testing set labels being withheld. For the training dataset, each tree has associated metadata about the species label, genus, scanning platform, and approximate tree height.

This study explores two different experimental setups: one which utilises the entire dataset and is evaluated through an online benchmarking website from Puliti et al. (2024), and a second which utilises the 6 species with the most samples from the training dataset and is evaluated locally. The six species selected for the second experiment are shown in Table 1. Even with the dataset being trimmed to just six species from exclusively the training set, the total number of samples available is 11,258. This study also explores the resulting differences when the data is downsampled to different point cloud sizes, namely 2048, denoted 2k, and 8192, denoted 8k. For experiment 1, the training dataset was split into 85% train and 15% validation datasets using stratified sampling, with the unlabelled test dataset being used for testing. For experiment 2, only the 6 species with the most samples, shown in Table 1, were used, with the samples being split into 70% train, 15% validation, and 15% testing datasets using stratified sampling.

Species	Samples
Pinus sylvestris	3296
Fagus sylvatica	2482
Picea abies	1983
Carpinus betulus	1243
Acer campestre	1240
Quercus faginea	1014

Table 1. The six most represented species in the training dataset, used for testing on a smaller number of species.

2.2 Data Preprocessing

For preprocessing the dataset, all of the point clouds were normalised by centering the points around the origin and scaling them to fit within a unit sphere. By scaling every point cloud to be within the range $[-1, 1]$, variance associated with differences in tree size and height is reduced. After normalisation, FPS (Farthest Point Sampling) is applied to downsample the point clouds to the same size. For data augmentations, random rotation (full 360° rotation around the z-axis, and $\pm 15^\circ$ tilt for the x and y axes), random scaling in the range $[0.8, 1.2]$, and Gaussian noise with a standard deviation of $\sigma = 0.01$ and a maximum magnitude of 0.05 were all applied to the training dataset. Applying each augmentation independently to generate additional samples increased the size of the training dataset by a factor of four.

2.3 PointMAE

PointMAE was first introduced by Pang et al. (2022) as a self-supervised learning framework that applies masked autoencoding to 3D point clouds. PointMAE is formulated as a denoising autoencoder, in which noise is introduced into the input point clouds by masking portions of the data, and the model is trained to reconstruct the original geometry from the remaining visible points. The training pipeline follows a two-stage procedure consisting of self-supervised pretraining on unlabelled point clouds, followed by supervised finetuning for downstream tree species classification. The pretraining and finetuning pipeline follows the implementation of Pang et al. (2022), with modifications to the data augmentation and finetuning configuration to suit the FOR-species20k dataset. The details of each stage are described in the following subsections.

2.3.1 Stage 1: Self-Supervised Pretraining The full FOR-species20k dataset was used for pretraining in both experiments, without any species labels. Since the pretraining objective operates on exclusively point coordinates, no class discriminative information is introduced.

For the 2k dataset, each point cloud is tokenised into 64 patches of 32 neighbouring points each; for the 8k dataset, 128 patches of 64 points are used. Each patch can overlap and is represented by subtracting its centroid from the local coordinates, which are then projected into a 384-dimensional token. Positional embeddings are produced by a learnable MLP that maps each patch centroid to the same embedding dimension (Pang et al., 2022).

Since patches are able to overlap, masking is applied at the patch level rather than per-point to preserve the completeness of each patch’s local geometry. Pang et al. (2022) found the most success with a masking ratio within the range of $60\% - 80\%$. As such, a 70% masking ratio is used for this study, applied by randomly selecting patches to mask. The masked patches are then replaced by shared learnable mask tokens, a trainable embedding projected to the decoder’s input dimension (192-dim).

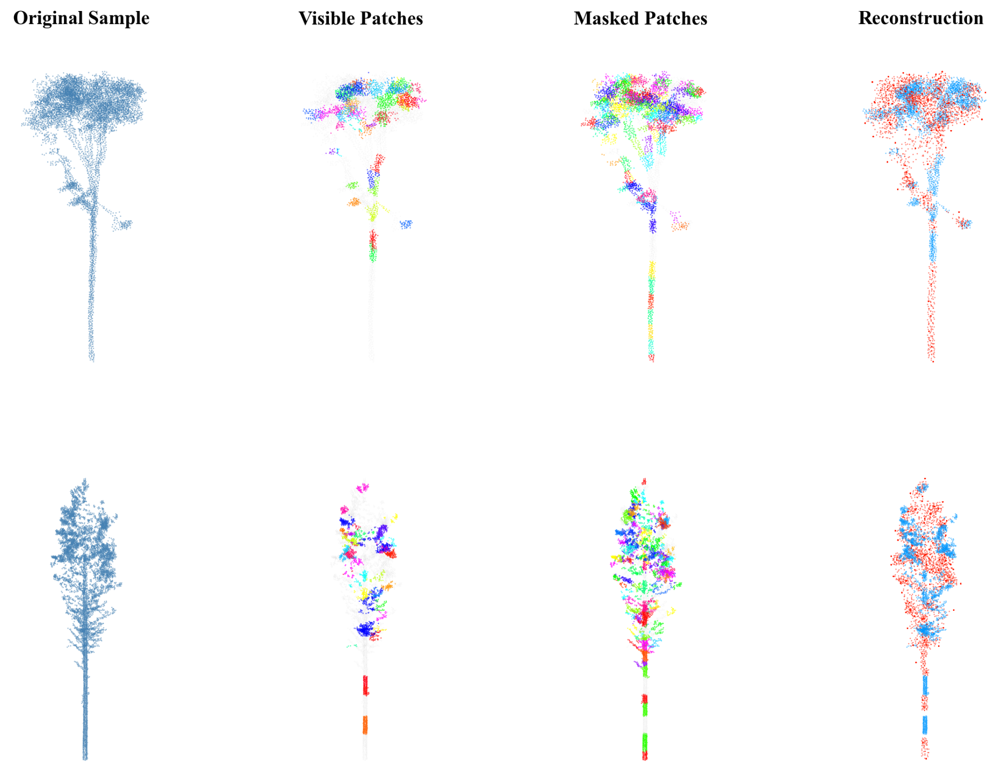

Figure 1. Example of how the visible and masked patches look on *Eucalyptus miniata* (top) and *Pinus sylvestris* (bottom) samples from the 8k dataset, and how the masked patches get reconstructed on epoch 300. Each patch is visualised through a different colour. Reconstructed points are shown in red.

Visible tokens are passed through a transformer encoder (384-dim, 6 heads); the resulting encoded tokens, together with the learnable mask tokens, are received by a decoder (192-dim, 6 heads) to reconstruct the point coordinates of the masked patches. Feeding only the visible tokens into the encoder reduces the number of input tokens and avoids leakage of positional information that would degrade feature learning (Pang et al., 2022). The reconstruction objective is the Chamfer Distance between the predicted and ground-truth coordinates in the masked regions. Some examples of the patches used and the reconstructed point cloud can be seen in Figure 1.

Augmentations (random rotation, scaling, Gaussian noise) are applied to all samples each epoch. Pretraining runs for 300 epochs across both the 2k and 8k configurations, using the Adam with decoupled weight decay (AdamW) optimiser with a learning rate of 0.001 and a weight decay of 0.05, gradient clipping at a maximum norm of 1.0, and a learning rate schedule of 10 epochs of linear warmup followed by cosine annealing to 1×10^{-5}.

2.3.2 Stage 2: Supervised Finetuning Following pretraining, the encoder is finetuned for species classification using labelled training samples. The pretrained tokeniser, positional embedding, and encoder are retained from pretraining. Unlike during pretraining, all patches are fed into the encoder, and the resulting encoded tokens are aggregated into a global feature vector via concatenated max-pooling and mean-pooling.

A classification head consisting of two fully connected layers (512 and 256 units) with batch normalisation, ReLU activations, and 20% dropout is appended to map the global feature vector to species predictions. To address class imbalance, we use inverse-frequency class weighting in the cross-entropy loss with label smoothing of 0.1 to reduce overfitting.

A differential learning rate is applied, with the encoder parameters being trained at $0.2\times$ the base learning rate to preserve pretraining representations, while the classification head trains at the full rate of 1×10^{-4}. Finetuning runs for 150 epochs using AdamW (weight decay of 0.05) with 10 epochs of linear warmup followed by cosine annealing to 1×10^{-6}, and gradient clipping at a maximum norm of 1.0.

2.4 PointNet and PointNet++

PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) serve as fully supervised baselines, trained end-to-end on labelled species data without a pretraining stage. PointNet aggregates per-point features into a global representation via max-pooling, while PointNet++ extends this with a hierarchical set-abstraction architecture that captures local geometric structure at multiple scales.

For the PointNet implementation, an orthogonality regularisation term is applied to the feature transformation matrix to preserve distances and angles between points. For PointNet++, two set abstraction layers with multiple radii (0.1/0.2/0.4 and 0.2/0.4/0.8) extract hierarchical features, followed by a global set abstraction and classification head with dropout rates of 0.4 and 0.5.

Both models use negative log-likelihood loss with inverse-frequency class weighting, AdamW with a learning rate of

Model	Exp.	Points	OA	Precision	Recall	F1
PointNet	1	2048	0.39	0.38	0.37	0.35
	1	8192	0.37	0.35	0.35	0.32
	2	2048	0.75	0.71	0.73	0.71
	2	8192	0.65	0.58	0.60	0.58
PointNet++	1	2048	0.63	0.67	0.61	0.63
	1	8192	0.48	0.52	0.48	0.47
	2	2048	0.84	0.81	0.82	0.82
	2	8192	0.83	0.80	0.81	0.80
PointMAE	1	2048	**0.67**	**0.67**	**0.66**	**0.66**
	1	8192	0.65	0.69	0.64	0.64
	2	2048	**0.89**	**0.87**	**0.87**	**0.87**
	2	8192	0.76	0.78	0.79	0.76

Table 2. Test results across all models, experiments, and point cloud sizes. Precision, recall, and F1 are macro-averaged. Bold indicates best performance per experiment.

0.001 and weight decay of 0.05, 10 epochs of linear warmup followed by cosine annealing, and gradient clipping at a maximum norm of 1.0. Both are trained for 200 epochs with early stopping applied at a patience of 25 epochs based on validation accuracy.

3. Results

Table 2 presents the results across both experiments and point cloud sizes. In both experiments, PointMAE achieved the highest overall accuracy at the 2048 point cloud size, outperforming PointNet++ by 0.04 OA in experiment 1 (0.67 vs. 0.63) and by 0.05 in experiment 2 (0.89 vs. 0.84). PointNet consistently ranked lowest across all experiments and point cloud sizes.

Across all models, performance decreased when the point cloud size was increased from 2048 to 8192, though the magnitude of the drop varied considerably by model. The PointNet++ showed the largest drop among all models in experiment 1 (0.15 OA) but the smallest in experiment 2 (0.01 OA). A notable exception to PointMAE's overall lead emerged on the 8k dataset in experiment 2, where PointNet++ outperformed PointMAE by 0.07 OA (0.83 vs. 0.76).

Figure 2 shows the normalised confusion matrix for the PointMAE on the 2k dataset in experiment 1. Among the 33 species, 8 achieved above 80% recall, namely *Picea glauca*, *Pinus contorta*, *Pinus radiata*, *Pinus resinosa*, *Corylus avellana*, *Eucalyptus miniata*, *Populus deltoides*, and *Quercus rubra*. Within-genus confusion was most pronounced among *Quercus*, *Pinus*, and *Fraxinus* species, where geometrically similar crowns within the same genus produced the largest off-diagonal concentrations.

Species	Precision	Recall	F1	Support
Acer campestre	0.82	0.78	0.80	186
Carpinus betulus	0.61	0.73	0.66	186
Fagus sylvatica	0.91	0.85	0.88	372
Picea abies	0.97	0.93	0.95	298
Pinus sylvestris	0.95	0.96	0.96	495
Quercus faginea	0.93	0.97	0.95	152
Macro avg	0.86	0.87	0.87	1689

Table 3. Per-species test results for the PointMAE with the 2048 point cloud size in experiment 2.

Per-class results for the PointMAE in experiment 2 (Table 3) show that *Pinus sylvestris*, *Picea abies*, and *Quercus faginea* achieved the highest F1-scores (0.96, 0.95 and 0.95 respectively), while *Carpinus betulus* was the most challenging (0.66). This pattern was consistent across all models in experiment 2.

4. Discussion

4.1 Self-Supervised Pretraining vs. Supervised Baselines

The PointMAE's advantage over both supervised baselines at the 2048 point cloud size suggests that self-supervised pretraining captures more transferable geometric features than end-to-end supervised training achieves with equivalent labelled data. However, this advantage is not unconditional: the reversal on the 8k dataset in experiment 2, where PointNet++ outperformed PointMAE, indicates that SSL benefits are configuration-dependent and that pretraining must be tailored to point density to remain effective.

It should be noted however, that the PointMAE encoder was pretrained on the full FOR-species20k dataset, including the geometric structure of test samples that the supervised baselines never observed during training. Although no species labels were used during pretraining, this asymmetry in data exposure means that the performance gap between PointMAE and the supervised baselines cannot be attributed solely to the benefits of self-supervised pretraining. Future work should pretrain using only the training split to isolate the contribution of SSL from incidental geometric familiarity with the test set.

One limitation of this evaluation is that neither of the baseline models used, the PointNet and PointNet++, shared architectural similarities with the PointMAE. The lack of another SSL model, such as the PointM2AE (Zhang et al., 2022), and a transformer-based model, such as the Point Transformer (Zhao et al., 2021), makes it difficult to discern whether the increased performance is due to self-supervised pretraining or the self-attention mechanisms shown in transformer models. Alongside introducing similar architectural models, another way to verify the benefits is to forgo pretraining, finetune the PointMAE's encoder, and compare the results against the pretrained model. Performing an ablation study on the encoder in this manner would also help determine whether the data exposure asymmetry noted above inflates the apparent benefit of SSL.

4.2 Effects of Point Cloud Size and Species-Level Structure

The diversity of the samples in the FOR-species20k dataset, while a boon for training a generalisable model, introduces significant challenges for classification. The varying scanner types utilised each produced point clouds with different densities and point distributions, which combined with the absence of harmonisation or normalisation across each sample, increases the difficulty of learning consistent features across species and platforms. This variation can also directly impact preprocessing

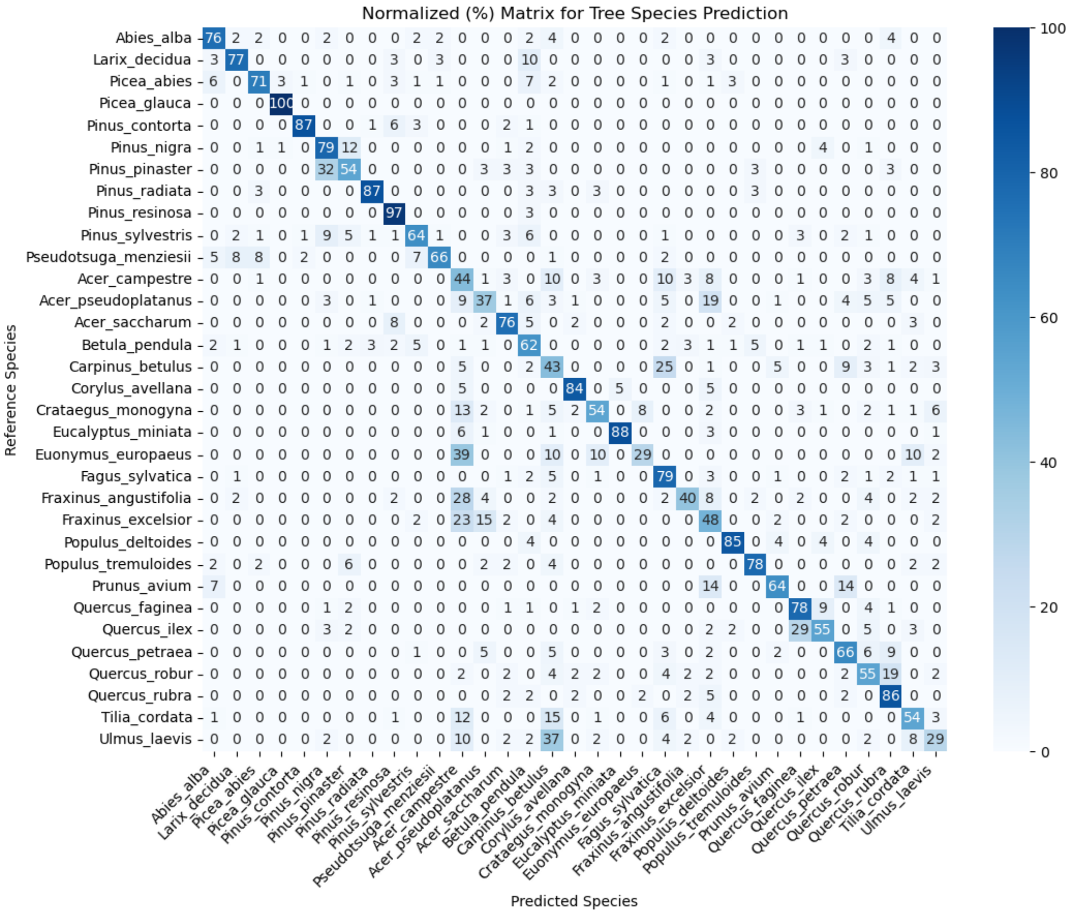

Figure 2. Normalised confusion matrix for the PointMAE with the 2048 point cloud size in experiment 1 (33 species). Values represent the percentage of true samples predicted as each class. Figure as returned by the FOR-species20k benchmarking platform (Puliti et al., 2024).

steps such as FPS downsampling, as its greedy approach to downsampling could potentially discard structurally important regions in unevenly dense point clouds.

The performance drop with increased point cloud size likely reflects the fixed hyperparameter configuration used across both settings. With the exception of the PointMAE’s patch count and patch size, no other parameters were adjusted for the 8k setting, including the learning rate, neighbourhood radii in the PointNet++, or the number of pretraining epochs for the PointMAE. Under these conditions, the additional points may have introduced extra noise rather than meaningful structural detail. The differential impact across models and experiments is also notable: in experiment 2, PointNet++’s hierarchical multi-scale architecture showed considerably greater resilience than PointNet or PointMAE, yet this pattern did not hold in experiment 1, where PointNet++ experienced the largest drop among all models. This suggests that the benefits of multi-scale feature extraction at higher point densities may be modulated by classification task complexity. The larger drop for the PointMAE in experiment 2 could reflect overfitting during pretraining, as 300 epochs on larger point clouds with more visible tokens may produce less generalisable feature representations. Future work should investigate the effects of varying pretraining duration and configuration on PointMAE performance at higher point densities.

Across each model in experiment 2, coniferous species such as *Pinus sylvestris* and *Picea abies* were consistently the easiest to classify, which is likely due to the distinct crown shape in point cloud form. *Quercus faginea* was also consistently easy to classify despite having the least number of samples in experiment 2, suggesting that the structural features of the species are distinctive enough for reliable classification. In contrast, *Carpinus betulus* was consistently the most challenging species to classify across all configurations. As seen in Table 3, both the F1-score (0.66) and precision (0.61) are significantly lower than the next lowest F1 and precision (*Acer campestre*, 0.80 and 0.82 respectively), indicating that other species were frequently misclassified as *Carpinus betulus*, while the lower recall (0.73) suggests that true *Carpinus betulus* samples were being classified as other classes. This pattern of confusion with *Carpinus betulus* samples was also observed by Puliti et al. (2024), with roughly a quarter of all true *Carpinus betulus* samples being predicted to be *Fagus sylvatica* samples with their DetailView method. This cross-family confusion is likely due to the similar convergent crown forms that each of these species exhibits.

The confusion matrix from experiment 1 (Figure 2) extends this pattern to the full 33-species classification, where within-genus confusion was most pronounced among *Pinus*, *Quercus*, and *Fraxinus* species. This is consistent with the crown geometry argument: species within the same genus tend to share similar architectural forms, making them harder to distinguish from point cloud structure alone. Conversely, the eight species exceeding 80% recall (*Picea glauca*, *Pinus contorta*, *Pinus radiata*, *Pinus resinosa*, *Corylus avellana*, *Eucalyptus miniata*, *Populus deltoides*, and *Quercus rubra*) are also consistent with this pattern, as each tends to exhibit relatively distinctive crown architectures. However, several of these species are North American or Australian, and their high recall may also reflect distributional uniqueness within the predominantly European dataset rather than purely morphological distinctiveness. These findings, together with the experiment 2 results, where *Quercus faginea* achieved near-perfect recall despite having the fewest samples, suggests that crown geometry is an important driver of per-species performance, though scanner source and geographic representation in the training data may also contribute.

4.3 Practical Implications for Forestry Applications

A central motivation for SSL in forestry applications is the high cost and time associated with acquiring labelled ground truth data. Collecting ground truth data and individual tree delineation are time consuming and costly (Wang et al., 2024), making large-scale labelled datasets difficult to produce. This study directly demonstrates the SSL pipeline under conditions that reflect the learning paradigm’s intended use: the PointMAE was pretrained on the full FOR-species20k dataset without any labels, then finetuned using only the labelled training split, matching the labelled data available to the supervised baselines. The pretrained model outperformed end-to-end supervised training under equal label budgets in the majority of configurations, with the exception of the 8k dataset in experiment 2. This suggests that within a dataset where individual trees have already been delineated, SSL pretraining on the unlabelled portions of data can improve classification over supervised-only approaches, reducing the number of species annotations required.

However, the benefits of SSL were not evaluated under explicitly label-limited conditions, as both the PointMAE and the supervised baselines were finetuned on the same labelled training set. Future work should aim to evaluate the ability for self-supervised learning to train a model on unlabelled data by constraining the amount of samples used in finetuning, such as through progressively reducing label fractions (e.g. 50%, 25% of available training labels) or using underrepresented species with few labelled samples (e.g. the six species with the least number of samples in the FOR-species20k dataset) in order to quantify the practical label efficiency of SSL pretraining.

5. Conclusion

This study evaluated self-supervised pretraining with the PointMAE for individual tree species classification from LiDAR point clouds, benchmarked against fully supervised PointNet and PointNet++ baselines on the FOR-species20k dataset.

Self-supervised pretraining improved general classification performance over supervised baselines when using the same labelled finetuning data, with PointMAE achieving 0.67 and 0.89 OA on the 2k dataset in the 33 and 6-species experiments respectively. This advantage, however, diminished at larger point cloud sizes without corresponding hyperparameter adjustments, implying that the benefits of SSL are configuration-dependent.

Increasing the point cloud size from 2048 to 8192 reduced performance across all models, though the impact varied by model and experiment. PointNet++ showed the greatest resilience in experiment 2 but the largest drop in experiment 1, suggesting that the relationship between architecture and point density is modulated by task complexity. Species-level analysis revealed that coniferous species with distinctive crown geometries were consistently the easiest to classify, while broadleaf species with similar crown forms, particularly *Carpinus betulus*, posed the largest challenge.

These results suggest that SSL pretraining on unlabelled point clouds is a promising strategy for forestry applications where annotated data is costly to collect. However, since the PointMAE encoder was pretrained on the full dataset including test set geometry, the observed performance advantage over supervised baselines may be partially attributable to incidental geometric familiarity with the test samples rather than SSL benefits alone. Future work should pretrain using only the training split to isolate SSL contributions from data exposure asymmetry, conduct an ablation study comparing pretrained and non-pretrained encoder finetuning, test finetuning under reduced label budgets and sample sizes, include both a supervised transformer and another SSL baseline to disentangle architectural from SSL contributions, and explore pretraining schedules optimised for high-density point clouds.

Acknowledgements

The authors wish to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Forestry Futures Trust Ontario for their generous financial support of this research. The FOR-species20k dataset was provided by Puliti et al. (2024) and is publicly available at `https://zenodo.org/records/13255198`. The authors thank the data contributors for making this benchmark available to the research community. The PointMAE implementation is based on the official code released by Pang et al. (2022), available at `https://github.com/Pang-Yatian/Point-MAE`, licensed under the MIT Licence. The PointNet and PointNet++ implementations are based on the open-source PyTorch reimplementations by Fei Xia[1] and Xu Yan[2] respectively.

References

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., Auli, M., 2022. data2vec: A general framework for self-supervised learning in speech, vision and language.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked autoencoders are scalable vision learners.

[1] `https://github.com/fxia22/pointnet.pytorch`
[2] `https://github.com/yanx27/Pointnet_Pointnet2_pytorch`

Li, Q., Hu, B., Shang, J., Remmel, T. K., 2025. Two-Stage Deep Learning Framework for Individual Tree Crown Detection and Delineation in Mixed-Wood Forests Using High-Resolution Light Detection and Ranging Data. *Remote Sensing*, 17(9), 1578.

Ma, Y., Zhao, Y., Im, J., Zhao, Y., Zhen, Z., 2024. A deep-learning-based tree species classification for natural secondary forests using unmanned aerial vehicle hyperspectral images and LiDAR. *Ecological Indicators*, 159, 111608.

Pang, Y., Wang, W., Tay, F. E. H., Liu, W., Tian, Y., Yuan, L., 2022. Masked autoencoders for point cloud self-supervised learning.

Puliti, S., Lines, E. R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M. J., Winiwarter, L., Rehush, N., Hristova, H., Murray, B., Calders, K., Terryn, L., Coops, N., Höfle, B., Junttila, S., Krůček, M., Krok, G., Král, K., Levick, S. R., Luck, L., Missarov, A., Mokroš, M., Owen, H. J. F., Stereńczak, K., Pitkänen, T. P., Puletti, N., Saarinen, N., Hopkinson, C., Torresan, C., Tomelleri, E., Weiser, H., Astrup, R., 2024. Benchmarking tree species classification from proximally-sensed laser scanning data: introducing the for-species20k dataset.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., Kumar, K., 2023. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 30(4), 2761–2775.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L., Lu, D., Xu, L., Robinson, D. T., Tan, W., Xie, Q., Guan, H., Chapman, M. A., Li, J., 2024. Individual tree species classification using low-density airborne multispectral LiDAR data via attribute-aware cross-branch transformer. *Remote Sensing of Environment*, 315, 114456.

Zhang, R., Guo, Z., Fang, R., Zhao, B., Wang, D., Qiao, Y., Li, H., Gao, P., 2022. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.