# Leveraging Transformer Models and Ancillary Data for Accurate Plant Trait Prediction

**Juhyun Lee**
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
j866lee@uwaterloo.ca

## Abstract

This project aims to predict plant traits using citizen science photographs and ancillary data, including climate and soil conditions. Building on previous work that employed CNNs, we utilized Transformers and ensembling technique in conjunction with a larger dataset from the iNaturalist platform to integrate visual and environmental features. We focused on predicting six specific plant traits: X4, X11, X18, X26, X50, and X3112. Model performance was assessed using the $R^2$ score to measure prediction accuracy. Our results demonstrate the potential of combining image-based data with environmental information for accurate plant trait prediction, contributing valuable insights for biodiversity research and ecological monitoring. Source Code: https://github.com/ljh0423/cs480project

## 1 Introduction

Introduction In recent years, advancements in artificial intelligence (AI) and citizen science have revolutionized the field of ecological research, particularly in the study of plant traits. Predicting plant traits—such as leaf size, shape, and color—has traditionally relied on expert knowledge and labor-intensive fieldwork. However, with the proliferation of digital technologies and large-scale image databases, there is now an unprecedented opportunity to leverage these resources to automate and enhance the prediction of plant traits.

The pioneering work by Schiller et al. (2021) demonstrated the potential of using Convolutional Neural Networks (CNNs) to predict plant traits from photographs. Building on this foundation, our project aims to extend the capabilities of AI in this domain by utilizing a larger dataset and incorporating ancillary information related to local climate, soil conditions, and satellite data.

### 1.1 Motivation and Objectives

The motivation behind this project is twofold. First, we seek to enhance the accuracy and scope of plant trait predictions by integrating comprehensive data sources, including high-resolution plant images and detailed environmental information. Second, we aim to demonstrate the practical application of advanced deep learning techniques in citizen science and ecological research, potentially transforming how plant traits are studied and monitored.

To achieve these goals, we have paired images from the iNaturalist database with extensive trait data curated by scientists over decades. Each image is accompanied by ancillary information providing context about the plant's environment. Our approach involves using advanced models, specifically Swin Transformer and EfficientViT, to extract features from the images. Additionally, we have developed a combined model that integrates EfficientViT with a simple neural network for the numer-

Submitted to CS480/680, 2024 Spring. Do not distribute.

ical ancillary data. This ensemble approach leverages the strengths of each model type, improving prediction accuracy for the six specific plant traits.

### 1.2 Model Structure

- **Swin Transformer**: This model extracts hierarchical visual features from input images using a transformer architecture with shifted windows, capturing both local and global contexts.
- **EfficientViT**: Known for its efficiency, this model processes images and is particularly suited for large-scale datasets.
- **Neural Network**: Used for processing the ancillary numerical data, consisting of environmental information.

The final ensemble model integrates predictions from Swin Transformer, EfficientViT, and a neural network handling ancillary data. This approach captures a comprehensive representation of plant traits to make effective predictions on the output traits.

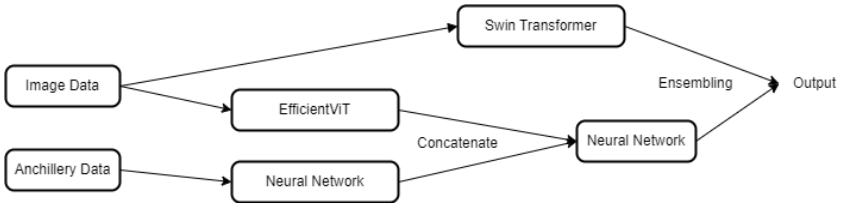


Figure 1: Model structure for plant trait prediction, combining Swin Transformer and EfficientViT for visual analysis with a simple neural network for ancillary data integration.

## 2 Related Works

Recent advancements in plant trait prediction have increasingly leveraged machine learning and computer vision. Schiller et al. (2021) used Convolutional Neural Networks (CNNs) to predict traits such as leaf area and plant height from thousands of annotated images. Their study showcased CNNs' capability to predict traits with high accuracy but was limited by a relatively small dataset and the exclusion of ancillary environmental data, which impacted the model's overall accuracy and generalizability.

Building on these foundations, Kattenborn et al. (2020) introduced a multi-modal approach by integrating aerial imagery with ground-based sensor data, such as soil moisture and temperature. This integration enhanced model accuracy, demonstrating the benefit of including environmental factors. However, their study was constrained geographically and did not address plant traits at the species level, which may limit the applicability of their findings to broader ecological contexts. Ustin and Gamon (2010) reviewed remote sensing technologies for estimating traits like chlorophyll content and canopy structure but focused primarily on aerial and satellite imagery. This approach lacked the fine-grained detail of ground-based photographs and did not incorporate modern deep learning techniques, which could enhance prediction accuracy.

Ma et al. (2020) utilized deep learning for high-throughput phenotyping, predicting traits such as plant height and leaf count while emphasizing the importance of temporal data due to trait variations over time. Nonetheless, their study was conducted in controlled environments, limiting the model's robustness and applicability to natural settings where variations are more pronounced. LeCun et al. (2015) provided a broad overview of deep learning techniques, which influenced subsequent studies in plant trait prediction. While their review highlighted the significance of CNNs in visual data analysis, it did not delve into the specifics of integrating multi-modal data or addressing the challenges posed by diverse datasets.

This report builds upon these prior works by addressing the specific problem of predicting plant traits from ground-based photographs and ancillary environmental data. Unlike previous studies, our approach aims to leverage the strengths of transformer-based models using Swin Transformer and EfficientViT models in capturing complex visual patterns.

# 3 Main Results

## 3.1 Problem Formulation

The task of predicting plant traits from images and ancillary data is framed as a multi-output regression problem. Given a dataset $X = \{x_i\}_{i=1}^{N}$ consisting of images and associated environmental information, the goal is to predict a vector $Y = \{y_i\}_{i=1}^{N}$ representing six plant traits.The model's objective is to maximize the coefficient of determination ($R^2$ score), which was created and used as a loss function criteria in training the models. The $R^2$ score is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \text{where} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Smooth L1 Loss is also used, which combines the quadratic loss for small errors and linear loss for large errors, providing a balance between sensitivity to small differences and robustness to outliers.

## 3.2 Methodology

### 3.2.1 Data Preprocessing

Similar data preprocessing methods from Schiller et al. 2021 were employed, including image normalization and random transformations to avoid overfitting, outlier filtering and standardization of ancillary data and output data, to reduce the impact of outliers and improve the accuracy and stability of the trained models. Different batches sizes were tested to ensure efficiency of the training process.

## 3.3 Design Choices and Justifications

Swin Transformer uses hierarchical approach with shifted windows, allowing it to capture intricate local and global features from images. Its ability to process images at multiple scales helps in understanding detailed patterns and context within the plant photographs.

EfficientViT is an optimized version of ViT, which is computationally efficient and scalable, making it suitable for handling large-scale datasets and its design reduces computational overhead while maintaining robust performance, which is crucial for processing image data.

We used the above pretrained models to encode features from the image data, and add it to the features from anchillery data to balance complexity and efficiency while capturing diverse data modalities.

## 3.4 Ablation Studies

The following configurations were tested to understand the contributions of each component:

- Model 1: **EfficientViT + Ancillary Data**: Focused on the integration of visual and numerical data. (Liu et al. 2021)
- Model 2: **Swin Transformer**: Served as a baseline for hierarchical visual feature extraction. (Zhou et al. 2023)
- Model 3: **Combined Model**: Use of ensembling to combine the results of the two models.

Initial testing with the **EfficientViT + Ancillary Data** model achieved an average R2 score of 0.429. The trait-specific R2 scores at different epochs are shown in Figure 2. The drop in validation score at epoch 6 exhibited signs of overfitting, so 5 epochs were used in training.
For the **Swin Transformer only** model, R2 scores were slightly fluctuant from those for model 1, shown in Table 1.

With inspiration from Yang and Browne 2004, weighted averaging was tested to identify the optimal weights of the results, which was found to be (**0.4,0.6**) for Model 1 and Model 2 respectively.

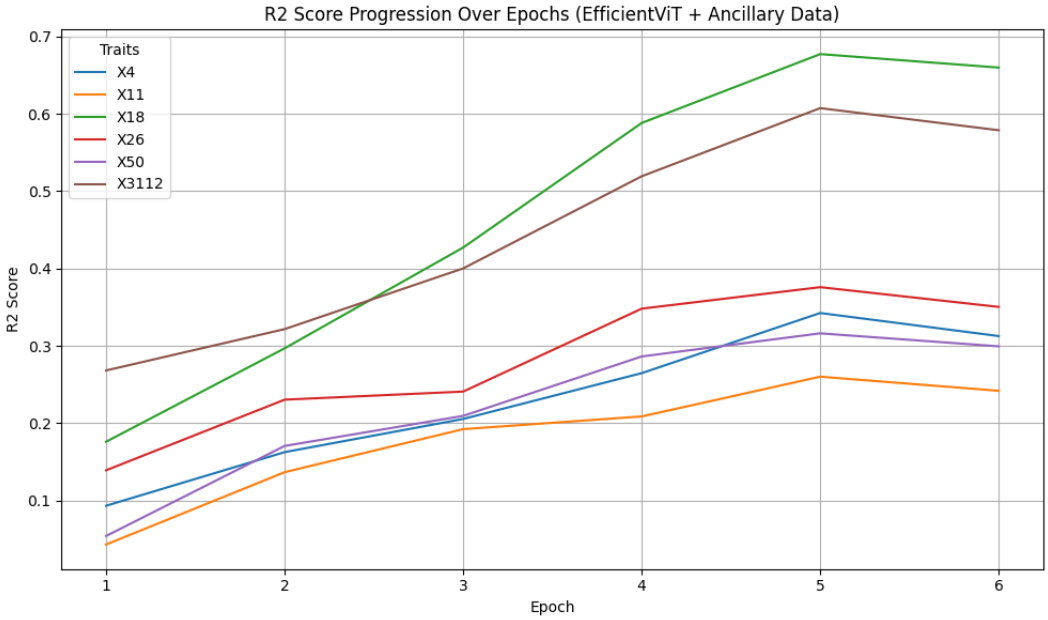


Figure 2: $R^2$ scores for each plant trait predicted using the EfficientViT model combined with ancillary data. The graph illustrates the model's accuracy in predicting the six traits, demonstrating its effectiveness in capturing and integrating both visual and contextual information for each specific trait.

### 3.5 Results and Evaluation

| Trait | R2 Score (Ensemble) | R2 Score (EfficientViT + Ancillary) | R2 Score (Swin Transformer) |
|---|---|---|---|
| X4 | 0.490 | 0.316 | 0.423 |
| X11 | 0.469 | 0.286 | 0.328 |
| X18 | 0.621 | 0.703 | 0.615 |
| X26 | 0.400 | 0.389 | 0.301 |
| X50 | 0.389 | 0.288 | 0.374 |
| X3112 | 0.537 | 0.589 | 0.523 |

Table 1: R2 Scores for Each Trait (Averaged over 2 runs)

The results highlight the benefits of integrating diverse data sources and model types for predicting plant traits. Combining the Swin Transformer and EfficientViT for image data with a neural network for environmental information led to notable improvements in prediction accuracy. Ensembling these models enhanced robustness and generalizability, effectively leveraging their strengths to handle the complexity and variability of the data, resulting in more reliable predictions.

## 4 Conclusion

In this study, we demonstrated the potential of combining advanced computer vision models and neural networks, specifically Swin Transformer and EfficientViT, with a simple neural network for ancillary data to predict plant traits from photographs and environmental information. Our approach successfully leveraged both visual and contextual data, resulting in robust predictions across various traits. However, we still face limitations such as the efficiency issue with training multiple large models, limited information from ancillary data due to correlation and the challenge of generalizing to diverse plant species beyond the dataset.

Future work could explore several directions to address these limitations and further improve the model's performance. This includes experimenting with smaller model architectures (especially with small input image and pruning correlated ancillary features) to enhance efficiency, integrating more comprehensive ancillary data, and utilizing transfer learning techniques to better capture the diversity of plant traits. Additionally, exploring boosting methods could provide further insights and enhance the model's predictive capabilities.

## Acknowledgement

I thank the professor and TAs who helped me by providing the knowledge and understanding that was necessary and vital, without which I would not have been able to work effectively on this assignment. Figure 1 was created using draw.io. Figure 2 created using Python.

## References

Kattenborn, T., F. E. Fassnacht, S. Schmidtlein, and F. Schiefer (2020). "Mapping plant traits and biomass with a multi-sensor approach". *Remote Sensing of Environment*, vol. 242, p. 111747.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". *Nature*, vol. 521, no. 7553, pp. 436–444.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo (2021). "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022.

Ma, X., W. Chen, Z. Zheng, and M. Liu (2020). "Deep learning for high-throughput phenotyping and bioinformatics". *Plant Physiology*, vol. 182, no. 2, pp. 1101–1120.

Schiller, C., S. Schmidtlein, C. Boonman, A. Moreno-Martínez, and T. Kattenborn (2021). "Deep learning and citizen science enable automated plant trait predictions from photographs". *Scientific Reports*, vol. 11, no. 1, p. 16395.

Ustin, S. L. and J. A. Gamon (2010). "Remote sensing of plant functional types". *New Phytologist*, vol. 186, no. 4, pp. 795–816.

Wightman, R. (2019). "PyTorch Image Models". `https://github.com/huggingface/pytorch-image-models`. Accessed: 2024-08-12.

Yang, S. and A. Browne (Nov. 2004). "Neural network ensembles: Combining multiple models for enhanced performance using a multistage approach". *Expert Systems*, vol. 21, no. 5, pp. 279–288.

Zhou, D., B. Kang, W. Zhao, C. Xie, X. Li, Q. Yu, and J. Feng (2023). "EfficientViT: Lightweight Vision Transformer with Cascaded Token and Channel Aggregation". *arXiv preprint arXiv:2302.08059*.