

Enhanced Multimodal Swin Transformer for Fine-Grained Plant Species Classification Using Contextual Metadata

Armayau Saadu¹; Dr. Karatu Musa Tanimu²; Dr. Sirajo Abdullahi Bakura³

^{1,2,3}Department of Computer Science, Federal University Birnin Kebbi, Nigeria

Publication Date: 2026/10/28

Abstract

Fine-grained plant species classification is difficult because closely related species can exhibit highly similar morphological characteristics, limiting the effectiveness of image-only recognition. This study developed an Enhanced Multimodal Swin Transformer that combines visual features from PlantCLEF images with contextual metadata describing taxonomic hierarchy, geographical distribution, seasonal occurrence, and ecological characteristics. A total of 91,673 images representing 1,000 plant species were partitioned into training (63,366), validation (14,081), and test (14,226) sets. Global Biodiversity Information Facility (GBIF) data were retrieved and linked to species using scientific names. ResNet50 and EfficientNet-B0 were trained as image-only baselines under the same dataset partitions, while the proposed framework used a pertained Swin Transformer visual branch, metadata encoder, metadata reliability module, and attention-based multimodal fusion. The proposed model achieved 97.12% best validation accuracy and 97.15% test accuracy, with weighted precision, recall, and F1-score of 97.48%, 97.15%, and 97.07%, respectively. These results exceeded ResNet50 and EfficientNet-B0 test accuracy by 24.23 and 27.22 percentage points, respectively. Confusion matrices showed fewer off-diagonal errors for the proposed model, while Grad-CAM visualizations indicated attention to biologically meaningful structures such as flowers, leaves, and reproductive organs. The findings support multimodal transformer-based learning as an effective approach for fine-grained plant identification and biodiversity monitoring.

Keywords: Biodiversity Monitoring; Contextual Metadata; Fine-Grained Classification; Plant Species Identification; Swin Transformer.

I. INTRODUCTION

➤ Background

Fine-grained visual classification (FGVC) concerns the discrimination of closely related categories within a broader visual class. Its central difficulty is the combination of substantial intra-class variation with small inter-class differences, which requires models to learn highly discriminative visual representations [1]. Earlier FGVC methods frequently relied on manually annotated object parts or bounding regions to isolate subtle features, but such annotations are costly and limit practical scalability [2,3,4].

These challenges are particularly important in biodiversity informatics, where species may differ only in subtle characteristics of leaves, flowers, stems, fruits, or reproductive structures. Automated recognition can support biodiversity monitoring, ecological assessment,

conservation, and plant identification, but image-only models may lack contextual evidence that human experts naturally use during identification. Biodiversity datasets also exhibit variation in image quality, ecological conditions, and species representation [8,9,10].

Transformer-based vision models provide an alternative to conventional convolutional architectures by modelling relationships across image regions and learning hierarchical representations. The Vision Transformer demonstrated the effectiveness of transformer architectures for image recognition [5], while the broader self-attention framework underlying transformers provides a mechanism for modelling relationships among input representations [6]. More recently, metadata fusion with transformer-based visual features has shown promise for plant identification, indicating that contextual information can complement visual evidence [7].

➤ Problem Statement

Fine-grained plant species recognition remains challenging because visually similar species can produce ambiguous image representations. Conventional CNN-based approaches may perform adequately on general image classification while remaining limited when subtle morphological distinctions are required. In addition, biodiversity observations may contain contextual information—such as taxonomy, geography, seasonality, and ecological occurrence—that is not directly available from the image. The research problem addressed in this study is therefore how to improve fine-grained plant species classification by combining visual representations with complementary contextual metadata while retaining interpretable evidence for model decisions.

➤ Aim and Objectives

The aim of this study was to enhance existing transformer-based fine-grained plant species classification through multimodal data integration to improve species identification accuracy and robustness.

- To enhance fine-grained plant species classification using a transformer-based visual model.

- To integrate visual information with taxonomic, geographical, seasonal, and ecological metadata.
- To benchmark the proposed framework against image-only deep learning models using standard performance metrics.

II. MATERIALS AND METHODS

➤ Research Design

The study adopted an experimental and applied artificial intelligence design. The proposed multimodal model was trained and evaluated against two image-only CNN baselines under identical dataset partitions. This design enabled direct assessment of the contribution of contextual metadata to fine-grained plant species classification.

➤ Dataset and Data Partitioning

The experiments used a PlantCLEF subset containing 91,673 images representing 1,000 plant species. The data were divided into 63,366 training images, 14,081 validation images, and 14,226 test images. The same partitions were used for all models to maintain a fair comparison.

Table 1 Dataset Composition Used for Model Development and Evaluation.

Subset	Images
Training	63,366
Validation	14,081
Test	14,226
Total	91,673
Plant species	1,000

➤ Contextual Metadata Acquisition and Processing

PlantCLEF images were linked to contextual information retrieved from the Global Biodiversity Information Facility (GBIF) using the scientific name of each species. The GBIF Species Match API was used for taxonomic information, while the GBIF Occurrence API provided geographical and ecological occurrence information. Metadata included Kingdom, Phylum, Class, Order, Family, Genus, recorded countries, mean latitude, mean longitude, most common observation month, and number of recorded occurrences. Categorical attributes were represented using embedding layers, whereas numerical attributes were normalized before fusion with visual features.

➤ Baseline Models

ResNet50 and EfficientNet-B0 were used as image-only baselines. Both models used pretrained ImageNet weights, were adapted to classify the 1,000 target species, and were trained without contextual metadata. Their results established reference performance for evaluating the proposed multimodal approach.

➤ Proposed Enhanced Multimodal Swin Transformer

The proposed architecture consisted of five principal components: a visual feature extractor, metadata encoder, metadata reliability module, attention-based multimodal fusion module, and classification head. A pretrained Swin Transformer extracted hierarchical image representations. The metadata encoder transformed categorical and numerical contextual attributes into a compact representation. The metadata reliability module estimated the confidence of contextual information so that incomplete or less informative metadata could exert an appropriately reduced influence. Visual and metadata representations were then combined using Multi-Head Attention to form a unified representation for classification.

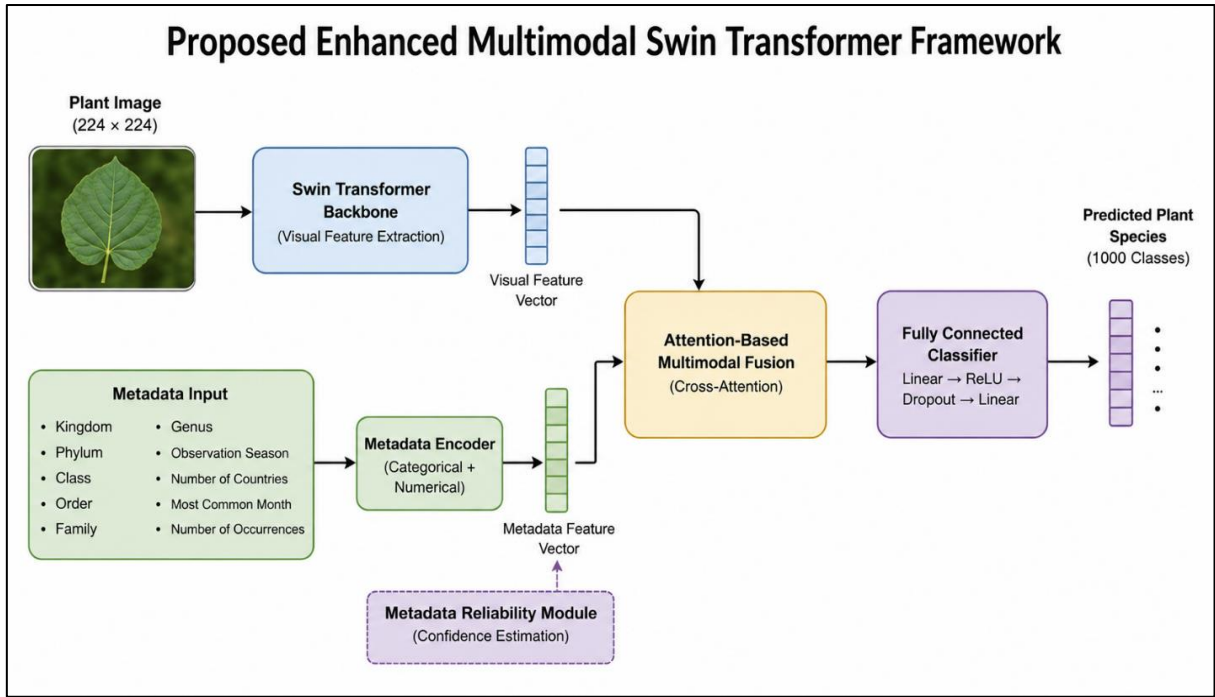


Fig 1 Overall Architecture of the Enhanced Multimodal Swin Transformer. The Framework Combines Hierarchical Visual Features with Encoded Contextual Metadata Through Reliability-Aware Attention-Based Fusion.

➤ Training Configuration

All models were trained using the same dataset partitions. The optimisation procedure used AdamW, cross-entropy loss, cosine annealing, gradient clipping,

and Automatic Mixed Precision (AMP). Training was limited to 50 epochs, with early stopping after seven consecutive epochs without validation improvement.

Table 2 Training Configuration Used in the Experiments.

Parameter	Value
Optimizer	AdamW
Loss function	Cross-entropy loss
Learning-rate scheduler	Cosine annealing
Maximum epochs	50
Early-stopping patience	7 epochs
Mixed precision	AMP enabled
Gradient clipping	Enabled

➤ Experimental Environment

Table 3 Experimental Computing Environment.

Component	Specification
Development platform	Kaggle Notebook
Programming language	Python
Deep learning framework	PyTorch
Vision library	Torchvision
Transformer library	TIMM
Data processing	NumPy, Pandas
Evaluation	Scikit-learn
Visualization	Matplotlib
Explainability	PyTorch Grad-CAM
Hardware accelerator	NVIDIA GPU (Kaggle)

➤ Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, and F1-score. Confusion matrices were generated to examine the distribution of correct and incorrect predictions across the 1,000 species. Weighted

precision, weighted recall, and weighted F1-score were used for the reported overall classification performance. Grad-CAM was applied to the proposed model to visualise image regions contributing to classification decisions [16].

III. RESULTS

➤ ResNet50 Baseline

ResNet50 achieved a best validation accuracy of 73.84% and a test accuracy of 72.92%. Weighted

precision, recall, and F1-score were 74.81%, 72.92%, and 72.66%, respectively. The training and validation curves showed stable optimisation, followed by a validation plateau below 75%, indicating limited additional generalisation from image-only learning.

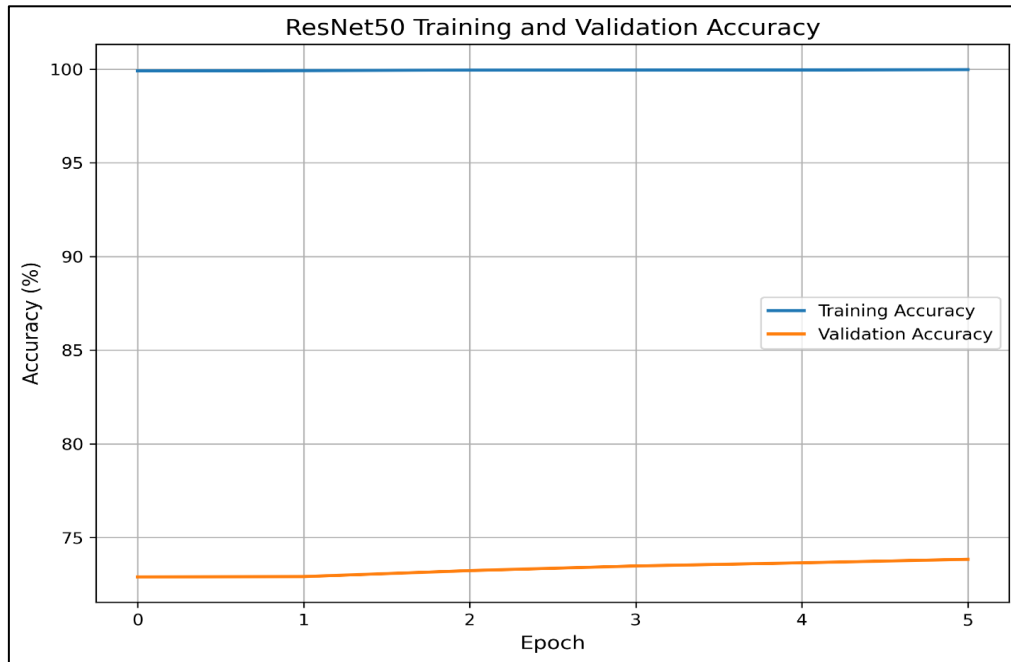


Fig 2 Training and Validation Accuracy Curve for ResNet50.

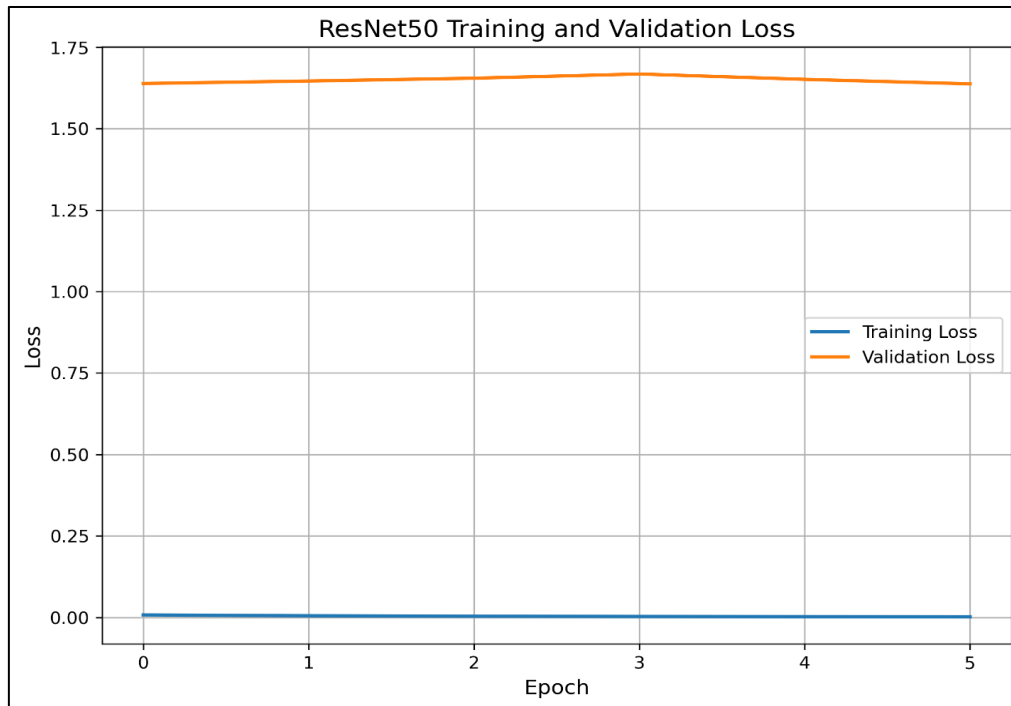


Fig 3 Training and Validation Loss Curve for ResNet50.

Table 4 Test Performance of ResNet50.

Metric	Value
Best validation accuracy	73.84%
Test accuracy	72.92%
Weighted precision	74.81%
Weighted recall	72.92%
Weighted F1-score	72.66%

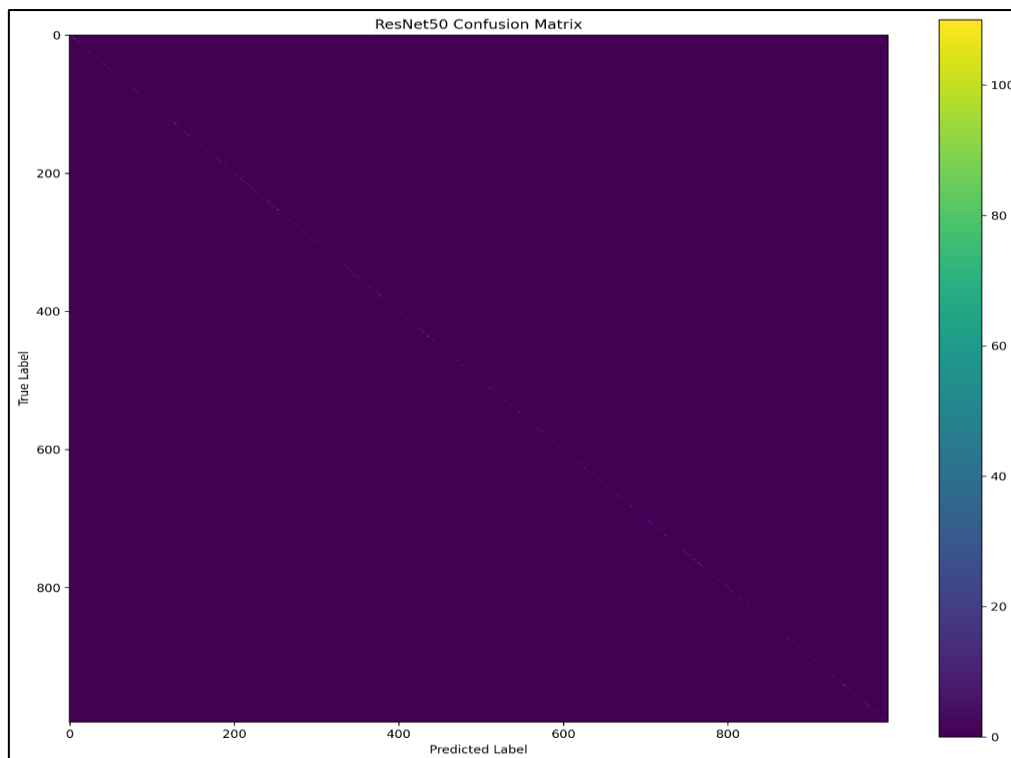


Fig 4 Confusion Matrix of ResNet50.

The ResNet50 confusion matrix showed correct predictions concentrated along the principal diagonal, but numerous off-diagonal entries remained. These errors indicate persistent confusion among species with similar visual morphology.

➤ *EfficientNet-B0 Baseline*

EfficientNet-B0 achieved a best validation accuracy of 70.11% and a test accuracy of 69.93%. Weighted precision, recall, and F1-score were 71.96%, 69.93%, and 69.65%, respectively. Its validation performance remained below that of ResNet50, indicating comparatively weaker generalization under the experimental conditions.

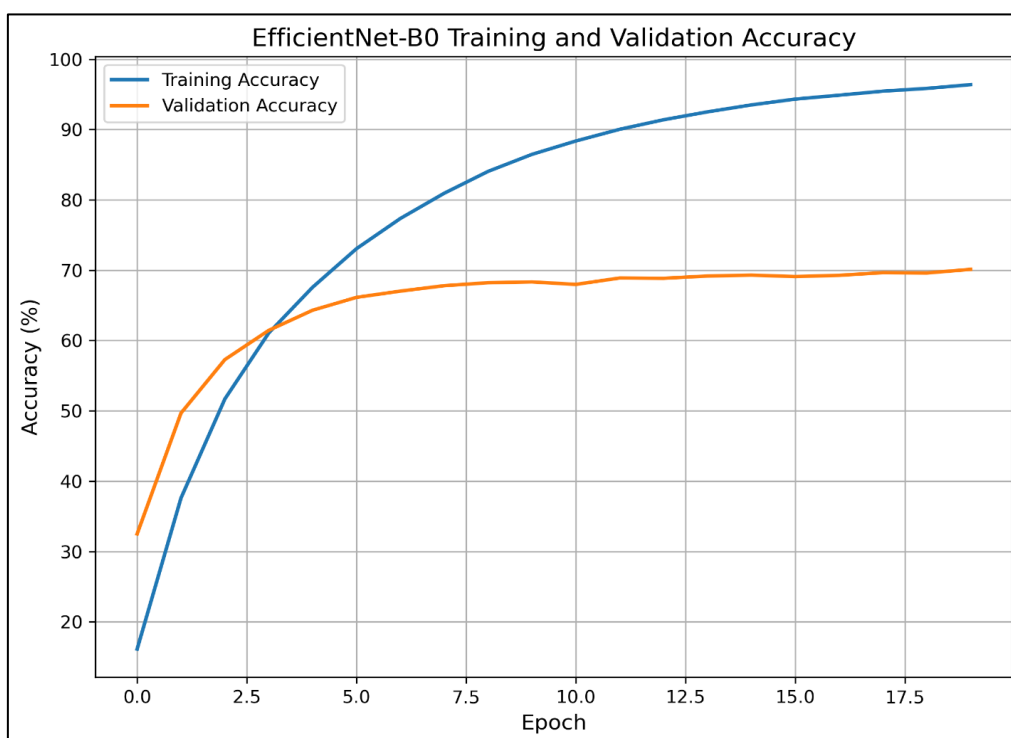


Fig 5 Training and validation accuracy curve for EfficientNet-B0.

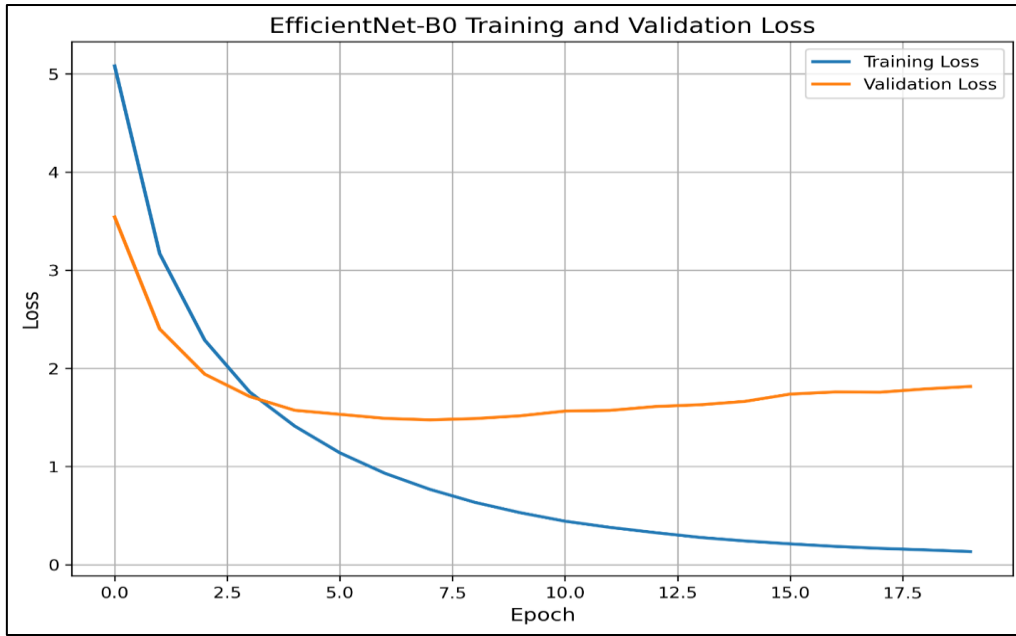


Fig 6 Training and validation loss curve for EfficientNet-B0.

Table 5 Test Performance of EfficientNet-B0.

Metric	Value
Best validation accuracy	70.11%
Test accuracy	69.93%
Weighted precision	71.96%
Weighted recall	69.93%
Weighted F1-score	69.65%

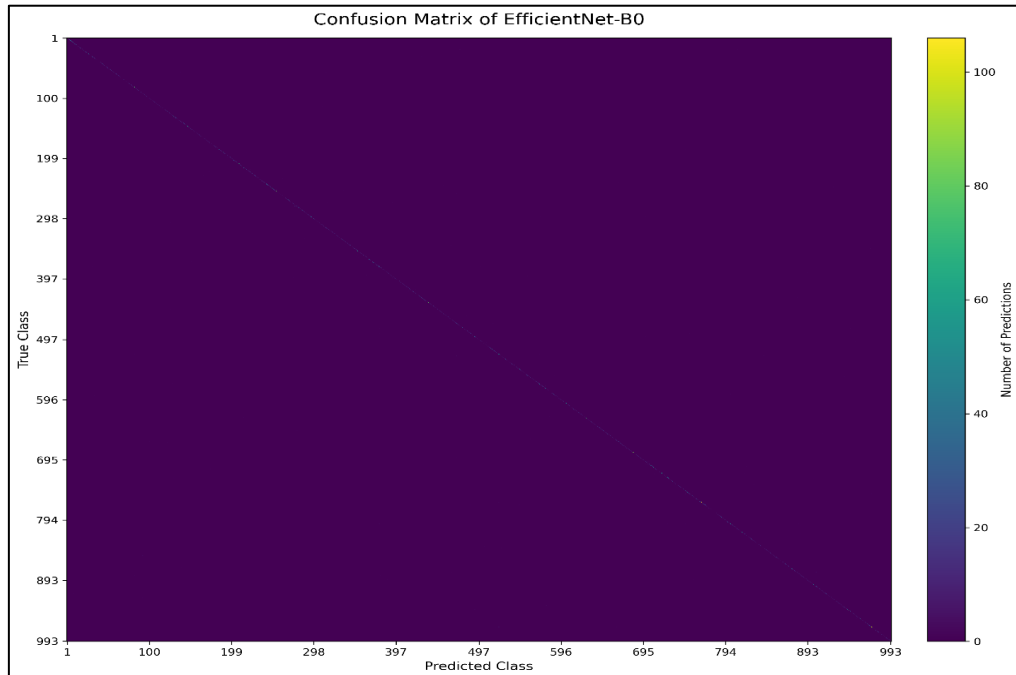


Fig 7 Confusion Matrix of EfficientNet-B0.

The EfficientNet-B0 confusion matrix similarly showed a diagonal concentration accompanied by substantial off-diagonal errors. The result reinforces the difficulty of distinguishing fine-grained plant categories from visual information alone.

➤ Proposed Enhanced Multimodal Swin Transformer

The proposed framework demonstrated substantially stronger learning behavior than both baselines. Training and validation accuracy increased toward a stable validation performance of approximately 97%, while the training and validation loss decreased during optimization. The best validation accuracy was 97.12%.

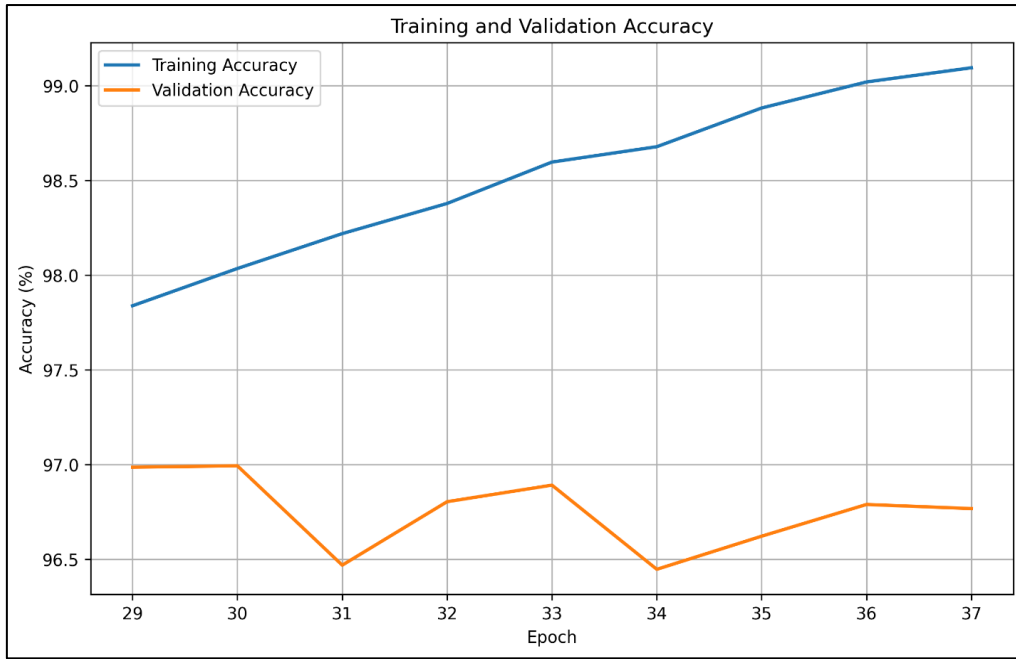


Fig 8 Training and Validation Accuracy Curve for the Enhanced Multimodal Swin Transformer.

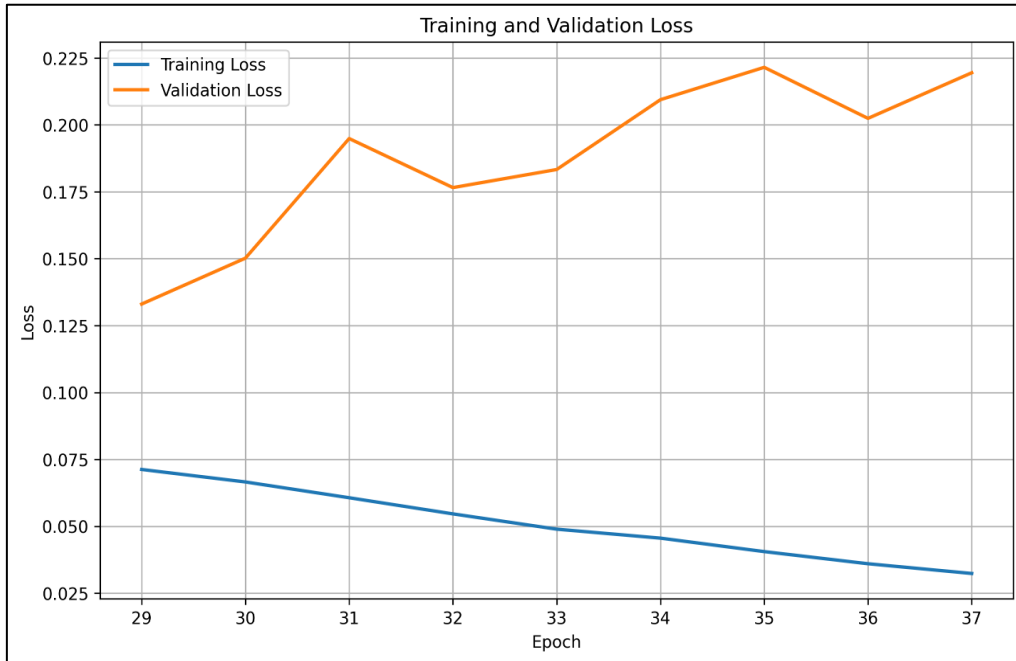


Fig 9 Training and Validation Loss Curve for the Enhanced Multimodal Swin Transformer.

Table 6 Performance of the Enhanced Multimodal Swin Transformer.

Metric	Value
Best validation accuracy	97.12%
Test accuracy	97.15%
Weighted precision	97.48%
Weighted recall	97.15%
Weighted F1-score	97.07%

The close agreement between validation accuracy (97.12%) and test accuracy (97.15%) indicates consistent performance on unseen test data. The model's confusion

matrix displayed a stronger diagonal structure and fewer off-diagonal errors than either image-only baseline.

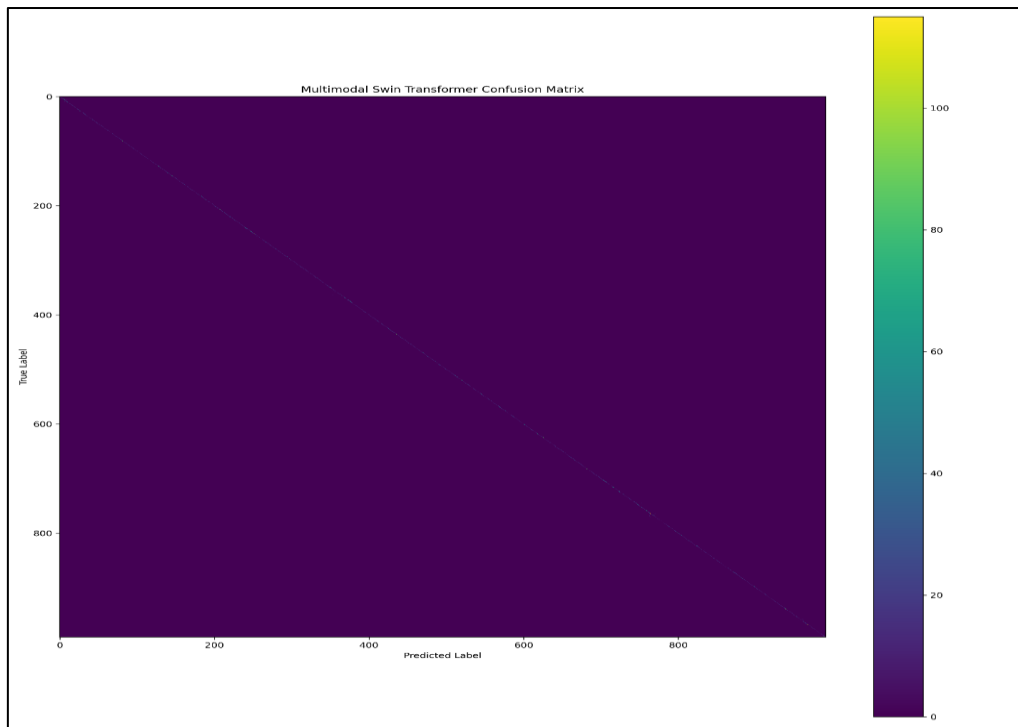


Fig 10 Confusion Matrix of the Enhanced Multimodal Swin Transformer.

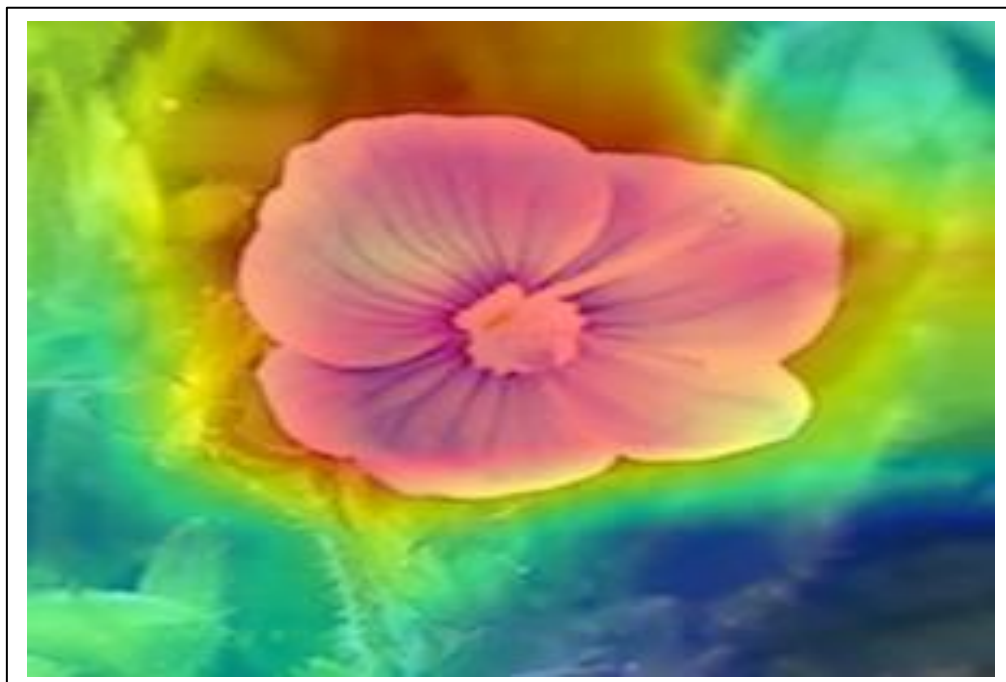


Fig 11 Grad-CAM Visualisation of the Enhanced Multimodal Swin Transformer, Showing Attention Concentrated on Biologically Meaningful Plant Structures.

➤ Comparative Performance

Table 7 Comparative Performance of the Evaluated Models.

Model	Input	Best validation accuracy (%)	Test accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNet-B0	Image only	70.11	69.93	71.96	69.93	69.65
ResNet50	Image only	73.84	72.92	74.81	72.92	72.66
Enhanced Multimodal Swin Transformer	Image + metadata	97.12	97.15	97.48	97.15	97.07

Table 8 Accuracy Improvement Achieved by the Proposed Model Relative to the Baselines.

Comparison	Validation improvement	Test improvement
Proposed vs. EfficientNet-B0	+27.01 percentage points	+27.22 percentage points
Proposed vs. ResNet50	+23.28 percentage points	+24.23 percentage points

The proposed model outperformed both baselines across every reported metric. The largest test-accuracy advantage was 27.22 percentage points over EfficientNet-B0 and 24.23 percentage points over ResNet50. The results therefore provide direct experimental evidence that the combined visual-contextual representation was more effective for this dataset than image-only classification.

IV. DISCUSSION

➤ *Effectiveness of Multimodal Learning*

The principal finding is the substantial improvement obtained by integrating contextual metadata with visual representations. The proposed model achieved 97.15% test accuracy, compared with 72.92% for ResNet50 and 69.93% for EfficientNet-B0. This result supports the premise that fine-grained species recognition benefits from information that complements morphology. Taxonomic, geographical, seasonal, and ecological attributes can provide discriminative cues when image appearance alone is ambiguous. Previous work on metadata fusion and plant recognition similarly reported benefits from combining contextual information with visual representations [7].

➤ *Comparison with Image-Only Models*

ResNet50 performed better than EfficientNet-B0 under the conditions of this study, but both models plateaued below 75% validation accuracy. This indicates that increasing the capacity or efficiency of image-only CNN feature extraction did not resolve the full ambiguity of the 1,000-class task. Fine-grained recognition methods have long demonstrated the value of attention and discriminative-part representations for visually similar categories [1,2,11], while biodiversity studies have highlighted the additional complexity introduced by ecological variability and species imbalance [8,9,12].

➤ *Transformer-Based Multimodal Representation*

The strong performance of the proposed framework is consistent with the broader development of transformer-based visual learning. Transformers can model relationships among image regions through attention mechanisms [5,6]. In the present study, this capability was extended by adding a contextual branch and an attention-based fusion mechanism. The resulting representation allowed visual and metadata features to interact before classification rather than treating metadata as an independent post-processing signal.

➤ *Generalization and Reliability*

The difference between the best validation accuracy (97.12%) and test accuracy (97.15%) was only 0.03 percentage points. Within the reported experimental setting, this close agreement is evidence of stable generalization to the held-out test set. The use of data augmentation, early stopping, gradient clipping, adaptive

learning-rate scheduling, and AMP supported stable optimisation. However, generalization beyond the evaluated PlantCLEF subset and geographic/ecological distributions was not directly tested and should not be assumed.

➤ *Explainability*

Grad-CAM visualizations showed that the proposed model concentrated on flowers, leaves, reproductive organs, and other relevant plant structures. This is important because high predictive performance alone does not establish whether a model is using biologically meaningful evidence. The visual explanations provide qualitative support that the learned image representation is linked to relevant morphology. Grad-CAM and related attribution methods are widely used to expose spatial evidence underlying deep-network predictions [16].

➤ *Practical Implications*

The results indicate potential applications in automated plant identification, biodiversity monitoring, ecological surveys, conservation programmes, invasive-species detection, and environmental assessment. A multimodal system can provide additional contextual evidence in cases where visual appearance is insufficient, potentially supporting botanists, conservation practitioners, environmental agencies, and citizen-science workflows.

V. LIMITATIONS

Several limitations should be considered. First, evaluation was performed on one large-scale PlantCLEF subset, so performance on independent datasets and other geographical regions remains to be established. Second, the framework depends on the availability and quality of contextual metadata; missing or inaccurate information could reduce the value of the metadata branch. Third, transformer-based models require substantial computational resources compared with lightweight CNNs. Finally, Grad-CAM primarily explains the visual pathway and does not fully reveal how individual metadata attributes influence a multimodal prediction.

VI. CONCLUSIONS

This study developed and evaluated an Enhanced Multimodal Swin Transformer for fine-grained plant species classification. Using 91,673 PlantCLEF images representing 1,000 species, the proposed framework combined hierarchical visual representations with taxonomic, geographical, seasonal, and ecological metadata. The model achieved 97.15% test accuracy, 97.48% weighted precision, 97.15% weighted recall, and 97.07% weighted F1-score, substantially outperforming the image-only ResNet50 and EfficientNet-B0 baselines. The findings demonstrate that contextual metadata can

provide complementary information for resolving ambiguities among visually similar plant species. The accompanying Grad-CAM analysis further indicated attention to biologically meaningful structures, supporting the interpretability of the visual component. Overall, the evidence supports multimodal transformer-based learning as a promising approach for automated fine-grained plant identification and biodiversity monitoring.

RECOMMENDATIONS

Evaluate the proposed framework on additional plant datasets spanning different geographical regions, climates, and ecological communities. Incorporate additional contextual variables such as soil properties, climate, altitude, satellite observations, and environmental sensor data.

Compare the proposed architecture with alternative transformer and hybrid architectures, including Vision Transformer, ConvNeXt, MaxViT, and hierarchical multimodal transformers.

Develop lightweight versions suitable for mobile devices, unmanned aerial vehicles, and edge-computing biodiversity monitoring platforms.

Extend explainability to multimodal attribution methods, SHAP, Integrated Gradients, and attention visualisation so that the contributions of both visual and contextual inputs can be examined.

ACKNOWLEDGMENT

The author acknowledges the data and computational resources used in this research, including the PlantCLEF dataset, GBIF biodiversity information services, and the Kaggle computing environment. Specific institutional or funding acknowledgments should be added if applicable.

➤ Conflict of Interest

The author declares no conflict of interest.

REFERENCES

- [1]. Guo C, Xie J, Liang K, Sun X, Ma Z. Cross-layer navigation convolutional neural network for fine-grained visual classification. Proceedings of the ACM Multimedia Asia (MMA Asia '21); 2021. doi:10.1145/3469877.3490579.
- [2]. Chai Y, Lempitsky V, Zisserman A. Symbiotic segmentation and part localisation for fine-grained categorisation. Proceedings of the IEEE International Conference on Computer Vision; 2013. p. 321–328. doi:10.1109/ICCV.2013.47.
- [3]. Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for fine-grained category detection. arXiv. 2014; arXiv:1407.3867.
- [4]. Rodner E, Simon M, Brehm G, Pietsch S, Wägele JW, Denzler J. Fine-grained recognition datasets for biodiversity analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern

- Recognition Workshops; 2015. doi:10.48550/arXiv.1507.00913.
- [5]. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv. 2020; arXiv:2010.11929.
- [6]. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems. 2017.
- [7]. Javed H, Fahad LG, Tahir SF, Hassan M, Alquhayz H. Enhanced plant species identification through metadata fusion and vision transformer integration. Computers, Materials & Continua. 2025;85(2):3981–3996. doi:10.32604/cmc.2025.064359.
- [8]. Pettorelli N. Deep learning and satellite remote sensing for biodiversity monitoring and conservation. Remote Sensing in Ecology and Conservation. 2024;10(2):178–195. doi:10.1002/rse2.379.
- [9]. Mai G, Jin J, Wu Y, Cai Y, Liu X, Chen C, et al. BioScan-1M: a step towards worldwide biodiversity assessment. arXiv. 2023; arXiv:2307.10455.
- [10]. Wang Q, Song Y, Du Y, Yang Z, Cui P, Luo B. Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species. Ecological Informatics. 2024; 80:102538. doi: 10.1016/j.ecoinf.2024.102538.
- [11]. Xiao T, Xu Y, Yang K, Zhang J, Peng Y, Zhang Z. Two-level attention models in deep convolutional neural networks for fine-grained image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 842–850. doi:10.1109/CVPR.2015.7298685.
- [12]. Zhai L, Wang Y, Huang J, Xu H, Wang C. Few-shot fine-grained fish species classification via contrastive learning and attention mechanisms. Frontiers in Marine Science. 2023; 10:1149186. doi:10.3389/fmars.2023.1149186.
- [13]. Keaton MR, Zaveri RJ, Kovur M, Henderson C, Adjero DA, Doretto G. Fine-grained visual classification of plant species in the wild. arXiv. 2021; arXiv:2106.02141. doi:10.48550/arXiv.2106.02141.
- [14]. Patil S, Sasikala M. Accurate plant species analysis for plant classification using convolutional neural network architecture. International Journal of Reconfigurable and Embedded Systems. 2024;13(1):160–170. doi:10.11591/ijres. v13.i1. pp160-170.
- [15]. Sulc M, Matas J. Fine-grained recognition of plants from images. Plant Methods. 2017; 13:115. doi:10.1186/s13007-017-0265-4.
- [16]. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalised gradient-based visual explanations for deep convolutional networks. Proceedings of the IEEE Winter Conference on Applications of Computer Vision; 2018. doi:10.1109/WACV.2018.00097.