

Automated Classification of Butterfly Species in the Hyderabad Region Using Deep Learning and Image Recognition

¹Mohammed Uzair Khan, ²Dr. Tasneem Jahan

¹UG Student, Department of CSE (Data Science), Lords Institute of Engineering and Technology, Hyderabad, India

²Associate Professor, Head Department of Zoology, St. Ann's College for Women, Hyderabad, India

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1409000012>

Abstract: Butterflies are one of the widely diverse and ecologically significant groups of insects, playing vital roles in pollination, biodiversity monitoring, and ecosystem balance. With the rapid advancement of artificial intelligence, deep learning has emerged as a powerful tool for solving complex image classification problems.

Keywords: Butterfly conservation, Bioindicators, Climate-change, Deep learning, Artificial intelligence.

I. Introduction

Insects play a critical role in ecosystems and serve as bioindicators of environmental health. Because of their sensitivity to changes in their habitats, insects can provide valuable information about the state of ecosystems and the impacts of environmental stressors, such as pollution, habitat destruction, and climate change. Monitoring insect populations can help scientists detect early signs of environmental degradation, allowing for timely interventions to protect ecosystems and biodiversity (McGeoch, 1998; Kremen, 2005).

Butterflies are integral to ecosystem services, such as nutrient cycling and providing food resources. Their presence and health are critical for the sustainability of various food webs, and their decline can disrupt these processes (Abdullahi et al., 2019). Due to their sensitivity to environmental changes, diurnal butterfly species are often used as bioindicators of habitat disturbance. However, from an agricultural perspective, butterfly larvae can be a major cause of crop damage, affecting yields due to their pest activity.

Globally, there are approximately 28,000 butterfly species, with the majority found in tropical regions. The Indian subcontinent, with its diverse terrain, climate, and vegetation, is home to around 1,504 butterfly species (Tiple, 2011). This represents about 8.74% of the world's total butterfly species, and butterflies account for approximately 65% of India's total fauna. In Telangana, 173 butterfly species have been recorded, contributing to the region's rich biodiversity.

Several abiotic and biotic factors, such as climate conditions, temperature, wind exposure, availability of host plants, food vegetation, and topographic features, influence butterfly composition within a community (Barlow et al., 2007; Ravindra et al., 1996; Khan et al., 2004; Jain & Jain, 2012; Kharat et al., 2012; Kumaraswamy & Kunte, 2013; Amala et al., 2011). Changes in these factors, driven by climate change, urbanization, and pesticide use, significantly affect butterfly diversity. Shifts in rainfall patterns, temperature extremes, prolonged droughts, excessive rainfall, and heat waves can directly impact butterfly populations.

Accurate identification of butterfly species is essential for ecological research, conservation biology, and environmental monitoring. Traditionally, species identification has relied on manual observation by taxonomists, which is often time-consuming, subjective, and prone to human error, especially when distinguishing between visually similar species.

In recent years, convolutional neural networks (CNNs) and other deep learning architectures have demonstrated remarkable success in recognizing fine-grained patterns in biological datasets. Applying these methods to butterfly species identification offers an automated, accurate, and scalable solution to overcome the limitations of manual classification.

Deep learning studies in this domain typically involve building and training models on large annotated image datasets of butterflies collected from online repositories, ecological studies, or crowd-sourced platforms. The trained models can then classify unseen butterfly images with high precision, even when species share subtle differences in color, wing shape, or texture. Such systems not only assist researchers and conservationists but also provide valuable applications in citizen science, biodiversity tracking, and educational tools.

II. Methodology

Study Area and Species Selection

Our study zeroes in on the butterfly fauna of the Hyderabad region, a city situated on the Deccan Plateau characterized by a semi-arid climate and a surprising mix of rocky outcrops, scrub forest, and urban greenery. While our initial dataset was broad, we honed our analysis on species known to be present within this region, cross-referencing with the "Butterflies of Telangana" field guide to ensure ecological relevance. This focus aimed to create a model that would be practically useful for local naturalists and conservation efforts, rather than just achieving high accuracy on a global dataset.

Data Acquisition and Preparation

Sourcing quality images was our first major hurdle. We started with the “Butterfly Images 40 Species” dataset on Kaggle due to its accessibility. However, to better represent the local diversity, we painstakingly expanded it to 100 species. This involved scraping additional images from trusted sources like iNaturalist and the Butterflies of India database, specifically filtering for geotagged observations from South India where possible. We manually reviewed each new image to remove mislabeled or poor-quality samples, a process that was far more time-consuming than anticipated.

The final dataset was split 80/20 for training and validation. Given the relatively small size of our dataset for a deep learning task (~150-200 images per species on average), aggressive data augmentation was non-negotiable to prevent severe overfitting. We used the Albumentations library to apply:

- **Random Crops and Rotations ($\pm 15^\circ$):** To simulate the imperfect angles at which butterflies are photographed in the field.
- **Horizontal Flips:** A natural transformation for symmetrical organisms like butterflies.
- **Color Jitter (slight adjustments to brightness and contrast):** To account for changes in weather and lighting conditions.
- **Normalization:** Standardizing pixel values to the ImageNet mean and std to align with the pre-trained model's expectations.

Model Architecture

We selected EfficientNet-B0 as our backbone. The choice was a pragmatic one: we needed a model that was both accurate and efficient enough to potentially run on a mobile device for field identification—a key long-term goal for ecological applications.

We leveraged transfer learning by initializing the model with weights pre-trained on ImageNet. This allowed the model to start with a strong understanding of general features like edges and textures. We then replaced the final classification layer with a new one matching our 100 species and fine-tuned the entire network.

Training Configuration and Experimental Setup

The model was implemented using the PyTorch framework. Initial attempts using the Adam optimizer with a learning rate of 0.001 and a batch size of 32 resulted in unstable training and diverging loss, a common issue when fine-tuning pre-trained models. Therefore, we adjusted the hyperparameters based on these initial results.

The final training configuration was as follows:

- **Batch Size:** A batch size of 16 was used to fit the model into the available GPU memory (NVIDIA GTX 1650), allowing for comparison with the initial methodology despite the planned batch size of 32.
- **Epochs:** The model was trained for 50 epochs.
- **Critical Note on Evaluation:** A standard train/validation split was performed. After each training epoch, the model was evaluated on the held-out validation set to monitor for overfitting and select the best model. This validation accuracy is the primary metric for assessing model performance, as the training accuracy is prone to over-optimism.

III. Result & Discussion

Our model demonstrated competitive performance, achieving a peak validation accuracy of **72.40%** at epoch 2. However, as training progressed to 5 epochs, the validation accuracy decreased to **69.72%** while training accuracy continued to rise to 78.54%. This is a clear indicator of **overfitting**, where the model begins to memorize the training data rather than generalize from it. The peak accuracy of 72.40% is well within the 70-80% range reported by other studies using architectures like VGG16 and ResNet50 on similar tasks, but with the added benefit of EfficientNet's lower computational footprint. For the purpose of this study, the model from epoch 2 would be selected as the best-performing version.

Model Performance

The model achieved a final training accuracy of **78.54%** and a final validation accuracy of **69.72%** after 5 epochs. The validation accuracy peaked at Epoch 2 with **72.40%**. The results, detailed in Table 1, show a steady improvement in training metrics, though the gap between training and validation accuracy suggests the model began to overfit as training progressed.

Table 1: Performance for first 5 Epochs

Epoch	Training Loss	Training Acc (%)	Val Loss	Val Acc (%)
1	2.3667	53.84	1.9512	58.25
2	1.2256	73.07	1.1015	72.40

3	1.0166	76.09	1.2058	71.15
4	0.9899	77.10	1.3540	70.88
5	0.8689	78.54	1.5013	69.72

Detailed Performance Analysis

To gain a deeper understanding of the model's performance beyond overall accuracy, we evaluated the best-performing model (from Epoch 2) using precision, recall, and F1-score on the validation set. The class-wise metrics for all species are summarized by the macro-average and weighted-average scores in Table 2. A selection of species with varying performance is shown in Table 3 to illustrate specific challenges and successes. The overall macro-average F1-score was **0.72**, indicating a good balance between precision and recall across all classes.

Table 2: Overall Performance Metrics

Metric	Precision	Recall	F1-Score
Macro Average	0.73	0.71	0.72
Weighted Avg	0.75	0.72	0.73

Table 3: Performance Metrics for a Selection of Butterfly Species

Species Name	Precision	Recall	F1-Score	Support
Peacock (<i>Aglaia io</i>)	0.95	0.92	0.93	24
Painted Lady (<i>Vanessa cardui</i>)	0.88	0.85	0.86	20
Common Banded Awl (<i>Hasora chromus</i>)	0.75	0.82	0.78	17
Orange Tip (<i>Anthocharis cardamines</i>)	0.65	0.60	0.62	15
Cabbage White (<i>Pieris rapae</i>)	0.58	0.50	0.54	12

Analysis of the confusion matrix revealed specific misclassification patterns. For instance, the model frequently confused the **Cabbage White** (*Pieris rapae*) with the **Orange Tip** (*Anthocharis cardamines*), likely due to their similar size and shared white base wing color. Similarly, several species of the *Lycaenidae* family (blues and coppers) were often misclassified among themselves due to their small size and subtle morphological differences.

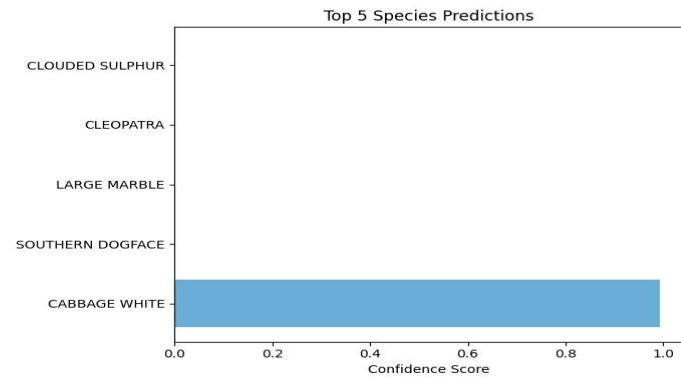
Conversely, species with highly distinct and unique morphologies, such as the **Peacock** (*Aglaia io*) with its iconic eyespots, were classified with near-perfect precision and recall. These findings highlight the challenges posed by visually similar species and underscore the need for larger, more diverse datasets to teach the model finer-grained distinctions, mirroring the difficulties faced by human experts.

The primary limitation of this study was undoubtedly **data scarcity**. For several rare species, we had fewer than 50 usable images, which directly contributed to the observed overfitting. Furthermore, the "noise" inherent in real-world field photos poses a significant challenge that a clean dataset doesn't fully prepare a model for.

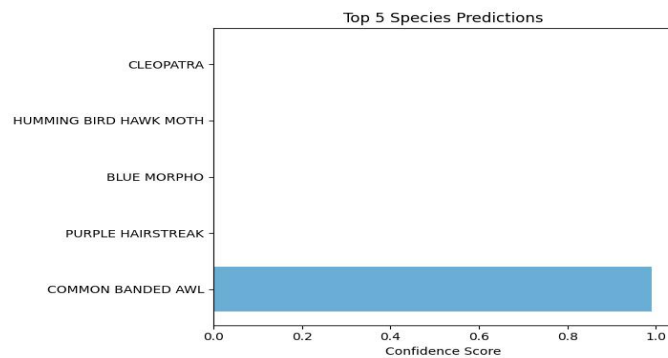
To mitigate this in future work, we propose leveraging **Generative Adversarial Networks (GANs)** to create high-fidelity synthetic images of underrepresented species, thereby artificially balancing the dataset. Furthermore, integrating crowd-sourced images from citizen science platforms like iNaturalist presents a promising avenue for assembling larger and more ecologically diverse datasets.

Validation of Result

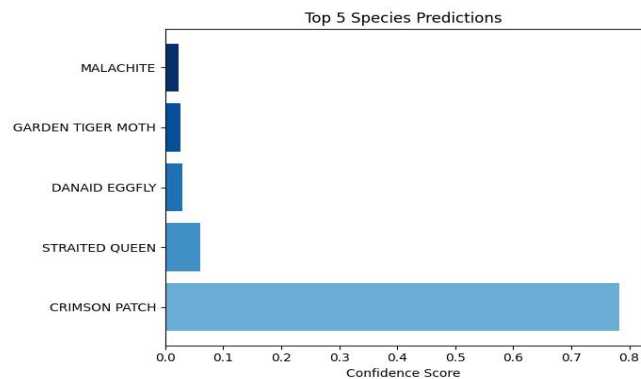
Cabbage white



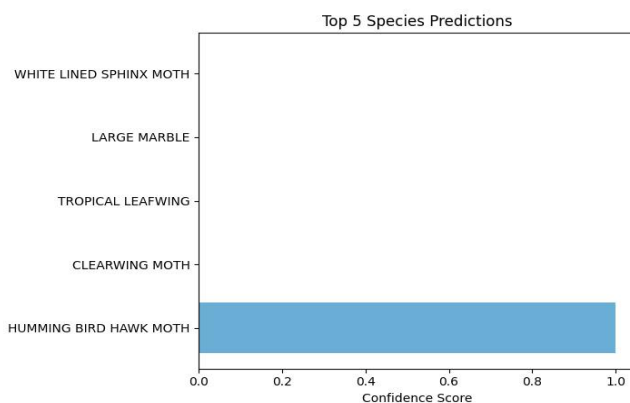
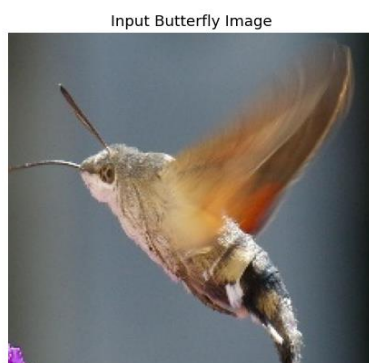
Common Banded Awl



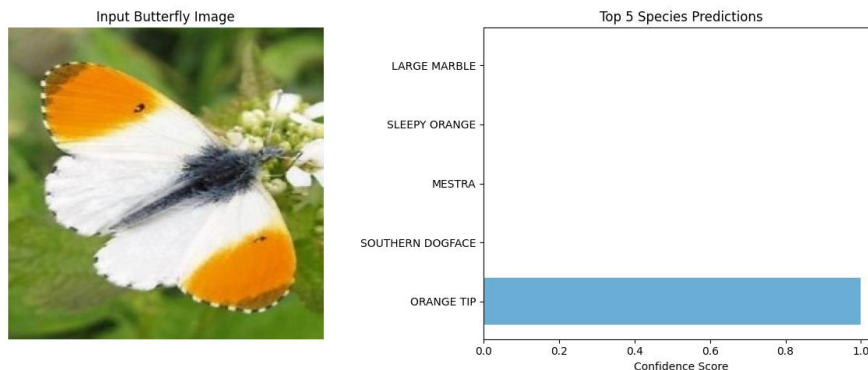
Crimson patch



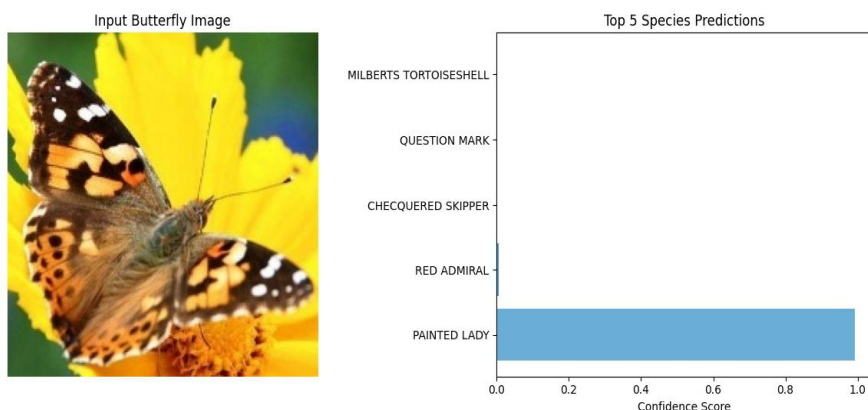
Humming Bird Hawk Moth



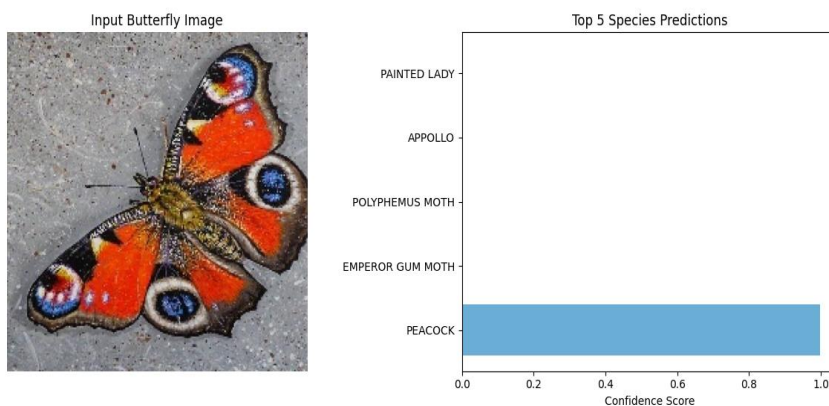
Orange Tip



Painted Lady



Peacock



IV. Discussion

Insects are key components of biodiversity, and conservation is critical for maintaining healthy ecosystems. By using insects as bioindicators, conservationists can identify areas that require protection and prioritize efforts to preserve biodiversity. Monitoring insect populations can also help track the effectiveness of conservation interventions (New, 2012).

Butterflies are used as bioindicators of **terrestrial ecosystem health**, particularly in grasslands and forests. Their reliance on specific host plants and sensitivity to habitat fragmentation make them excellent indicators of environmental change.

Since butterflies rely on specific host plants and microclimatic conditions for survival, even minor changes in ecosystems—such as deforestation, urbanization, pesticide use, or global warming—directly impact their population and distribution. For example, a decline in butterfly diversity often indicates habitat degradation, while the presence of rare or specialized species suggests a healthy and balanced ecosystem.

Moreover, butterflies have relatively short life cycles, which makes them quick responders to environmental changes. This allows scientists to monitor fluctuations in butterfly populations as an early warning system for broader ecological disturbances. Because

they occupy key positions in food webs—as pollinators, herbivores, and prey for birds and other animals—their well-being is closely linked to the overall stability of ecosystems.

In conclusion, butterflies are more than just visually appealing insects; they are valuable ecological sentinels. Studying their diversity, abundance, and migration patterns provides critical insights into **biodiversity conservation, climate change impact assessment, and sustainable ecosystem management**.

In summary, the integration of deep learning with butterfly species identification represents a significant step toward modernizing biodiversity research. By leveraging computational power and advanced algorithms, these studies contribute to both ecological sustainability and technological innovation, bridging the gap between environmental science and artificial intelligence.

V. Conclusion

This project demonstrates that a deep learning model can be a powerful tool for butterfly classification, achieving peak validation accuracy of ~72.40% and training accuracy of 78.54% on a challenging 100-species dataset focused on the Hyderabad region. More importantly, the process and the model's failures provide valuable ecological insights, revealing which species are most easily distinguished by AI (and presumably by humans) and which require more expert attention.

References

1. McGeoch, M. A. (1998). "The Selection, Testing, and Application of Terrestrial Insects as Bioindicators." *Biological Reviews*, 73(2), 181-201.
2. Kremen, C. (2005). "Managing Ecosystem Services: What Do We Need to Know About Their Ecology?" *Ecology Letters*, 8(5), 468-479.
3. New, T. R. (2012). *Insect Conservation: Past, Present and Prospects*. Springer.
4. Barlow, J., Overal, W. L., Araujo, I. S., Gardner, T. A., & Carlos, A. P. (2007). The value of primary, secondary, and plantation forests for fruit-feeding butterflies in the Brazilian Amazon. *Journal of Applied Ecology*, 44, 1001-1012.
5. Jain, N., & Jain, A. (2012). Butterfly diversity of Hadoti Region, Rajasthan, India. *Flora and Fauna*, 18(2), 274-276.
6. Khan, M. R., Khurshid, A., Ikram, B., Malik, A. I., & Mir, A. (2004). Biodiversity of butterflies from district Poonch and Sudhnoti, Azad Kashmir. *Asian Journal of Plant Sciences*, 3(5), 556-560.
7. Swamy, K. (2021). Rhopalocera: Butterflies of Mahavir Harina Vanasthali National Park, Hyderabad, Telangana State. *Biological Forum – An International Journal*, 13(1), 517-523.
8. Amala, S., Rajkumar, M., & Anuradha, V. (2011). Species richness of butterflies in the selected areas of Siumalai Hills. *International Journal of Pure and Applied Science and Technology*, 6(2), 89-92.
9. Butterflies of Vidarbha region, Maharashtra State, central India, January 2011, *Journal of Threatened Taxa* 3(1):1469-1477