http://dx.doi.org/10.18623/rvd.v23.6532

IMAGE-BASED IDENTIFICATION OF SUMATRA BUTTERFLY SPECIES USING DEEP LEARNING

IDENTIFICAÇÃO DE ESPÉCIES DE BORBOLETAS DE SUMATRA BASEADA EM IMAGENS USANDO APRENDIZADO PROFUNDO

Article received on: 1/16/2026
Article accepted on: 4/15/2026

Rico Andrian*
*Doctoral Program of Mathematics and Natural Sciences, Faculty of Mathematics and Natural Sciences, Lampung University, Bandar Lampung, Lampung, Indonesia
Orcid: https://orcid.org/0000-0002-9512-8925
rico.andrian@fmipa.unila.ac.id

Admi Syarif*
*Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lampung University, Bandar Lampung, Lampung, Indonesia
Orcid: https://orcid.org/0000-0003-3316-0388
admi.syarif@fmipa.unila.ac.id

Favorisen R. Lumbanraja*
*Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lampung University, Bandar Lampung, Lampung, Indonesia
Orcid: https://orcid.org/0009-0007-8093-5891
favorisen.lumbanraja@fmipa.unila.ac.id

Emantis Rosa*
*Department of Biology, Faculty of Mathematics and Natural Sciences, Lampung University, Bandar Lampung, Lampung, Indonesia
Orcid: https://orcid.org/0000-0001-6348-2199
emantisrosa@gmail.com

The authors declare that there is no conflict of interest

Abstract
Deep learning has gained momentum in the last decade for image-based species identification. We examine the classification of eight Sumatra butterfly species with a dataset of 800 photos from Gita Persada Butterfly Park using several deep learning architectures. The dataset was separated into training, validation, and testing subsets under controlled experimental circumstances. We evaluated seven architectures employing transfer learning with ImageNet pretrained weights, including convolutional neural networks (CNNs) and a Vision Transformer (ViT). The DenseNet201 model obtained the highest classification accuracy (99.38%), followed by ResNet50 and Xception (98.75%), MobileNet (97.50%), InceptionV3 (95.63%), ViT (93.75%), and EfficientNetB0

Resumo
O aprendizado profundo ganhou impulso na última década para a identificação de espécies baseada em imagens. Examinamos a classificação de oito espécies de borboletas de Sumatra com um conjunto de dados de 800 fotos do Parque de Borboletas Gita Persada, utilizando várias arquiteturas de aprendizado profundo. O conjunto de dados foi separado em subconjuntos de treinamento, validação e teste sob circunstâncias experimentais controladas. Avaliamos sete arquiteturas empregando transferência de aprendizado com pesos pré-treinados do ImageNet, incluindo redes neurais convolucionais (CNNs) e um Vision Transformer (ViT). O modelo DenseNet201 obteve a maior precisão de classificação (99,38%), seguido pelo ResNet50 e Xception

(85.63%). The performances of CNN-based models were more stable under the conditions of the present dataset, while MobileNet reached a good trade-off between accuracy and computational efficiency. The results must be viewed from the perspective of a tiny, single-site data collection. This was a controlled study and not a deployable field application as a standard for the identification of butterfly species from images. There was no external validation or field-based assessment. The results, therefore, give initial support for the possibility of deep learning for automated identification of butterfly species, but further validation using bigger and more diverse datasets is needed to determine model generalizability.

Keywords: Deep Learning. Butterfly Species Identification. Convolutional Neural Networks. Vision Transformer. Transfer Learning.

(98,75%), MobileNet (97,50%), InceptionV3 (95,63%), ViT (93,75%) e EfficientNetB0 (85,63%). As performances dos modelos baseados em CNN foram mais estáveis nas condições do conjunto de dados atual, enquanto o MobileNet alcançou um bom equilíbrio entre precisão e eficiência computacional. Os resultados devem ser vistos na perspectiva de uma coleta de dados pequena e em um único local. Este foi um estudo controlado e não uma aplicação de campo implantável como padrão para a identificação de espécies de borboletas a partir de imagens. Não houve validação externa ou avaliação em campo. Os resultados, portanto, fornecem um suporte inicial para a possibilidade de aprendizado profundo na identificação automatizada de espécies de borboletas, mas é necessária uma validação adicional usando conjuntos de dados maiores e mais diversos para determinar a generalização do modelo.

Palavras-chave: *Aprendizado Profundo. Identificação de Espécies de Borboletas. Redes Neurais Convolucionais. Transformador de Visão. Aprendizado por Transferência.*

1 INTRODUCTION

Indonesia is recognized as one of the world's biodiversity hotspots, hosting a rich diversity of butterfly species that function as pollinators and bioindicators. Butterflies are commonly used in ecological studies due to their sensitivity to environmental changes, making them useful indicators of habitat quality and ecosystem stability. However, accurate species-level identification remains challenging, particularly in regions with high diversity and overlapping morphological characteristics.

Traditional identification methods rely on expert knowledge and manual observation, which can be time-consuming and subject to inconsistency. In Indonesia, the availability of well-annotated image datasets remains limited, constraining the development of automated recognition systems. Recent advances in deep learning have enabled progress in image-based species identification. Convolutional Neural Networks (CNNs) have been widely applied to fine-grained classification tasks by learning discriminative visual features such as wing patterns, venation, and coloration. Previous

studies have shown that transfer learning can improve recognition performance under limited data conditions (Adityawan *et al.*, 2023), while object detection approaches such as YOLO-based models have been used for butterfly detection in more complex image settings (Yasmin *et al.*, 2023). These developments indicate the potential of deep learning to support biodiversity informatics rather than replace traditional taxonomic practices.

In addition to CNN-based approaches, transformer-based models such as Vision Transformers (ViT) have been introduced for image classification tasks. These models rely on attention mechanisms rather than convolutional operations and have demonstrated strong performance in large-scale datasets. However, their effectiveness in fine-grained species identification under limited data conditions remains less well understood. In ecological applications with small and localized datasets, CNN-based models may provide more stable performance due to lower data requirements.

Despite these developments, fine-grained classification of insects remains difficult. Many butterfly species share similar visual traits, particularly within the same genus, making them challenging to distinguish based on image data alone. These challenges are more pronounced when datasets are limited in size and diversity, which is often the case in ecological studies. In the Indonesian context, systematic evaluation of deep learning architectures for local butterfly species is still limited. Most existing studies focus either on general insect classification or on datasets collected from broader geographic regions. As a result, there is a need for controlled benchmarking studies using locally sourced datasets to better understand model behavior under constrained ecological conditions.

This study addresses this gap by conducting a comparative evaluation of several deep learning architectures for image-based identification of Sumatra butterfly species. The dataset consists of 800 images representing eight species collected from Gita Persada Butterfly Park, Lampung. The study is designed as a controlled experiment using a single-site dataset, enabling a consistent comparison across models under the same training protocol. The contribution of this study is threefold. First, it provides a curated local dataset of Sumatra butterfly species that can support future research in biodiversity informatics. Second, it presents a direct comparison between CNN-based models and a transformer-based model under the same experimental setup. Third, it evaluates the trade-off between classification accuracy and computational efficiency, which is relevant for

developing lightweight identification tools. In this research, we compare the CNN models and the Vision Transformer model for the identification of Sumatra Butterflies. For the experiments, we used the real Sumatra butterfly images taken from Gita Persada Butterfly Park in Bandar Lampung, Indonesia.

It is important to note that this study does not aim to present a deployable biodiversity monitoring system. Instead, the results should be interpreted as a proof-of-concept for automated species identification under controlled conditions. No external dataset validation or field-based evaluation was performed, and therefore, the findings are intended to support further research in digital taxonomy and automated recognition rather than immediate application in large-scale ecological monitoring.

2 LITERATURE REVIEW

Internationally, Indonesia is known as a major biodiversity hotspot with a huge variety of endemic butterfly species (Peggie *et al.*, 2025). These insects are crucial ecological actors as essential pollinators that support the stability of tropical ecosystems. In addition to their reproductive roles in plants, butterflies are good bioindicators for monitoring the health of the environment (Koneri *et al.*, 2020). Moreover, their population dynamics are often used as a criterion to rank conservation priorities in protected natural parks (Ilhamdi *et al.*, 2018)

Traditionally, taxonomic identification is mainly dependent on the competence of entomologists (Almryad & Kutucu, 2020). Butterfly species are generally time-consuming to observe manually(Liang *et al.*, 2020). The number of trained taxonomists has declined substantially (Almryad & Kutucu, 2020). This limited number has been a historical constraint on the rate of biodiversity assessment in the Indonesian context (Peggie *et al.*, 2025). These restrictions have resulted in the investigation of automated systems in the field of biodiversity informatics to deal with huge amounts of field data (Yasmin *et al.*, 2023). As a consequence, deep learning has become a revolutionary method for biological classification (Almryad & Kutucu, 2020). These algorithms can learn features directly from digital pictures (Fathimathul *et al.*, 2022).

Convolutional Neural Networks have transformed the field of biological identification by extracting fine-grained visual information (Fathimathul *et al.*, 2022).

These algorithms are adept at spotting discriminative patterns such as wing venation and intricate pigmentation, which are difficult for human observers to quantify (Sagar *et al.*, 2020). It has been shown that transfer learning approaches can substantially improve model performance when working with limited datasets (Adityawan *et al.*, 2023). This strategy allows researchers to exploit pre-trained weights to increase the accuracy of specific categorization tasks in ecology (Fathimathul *et al.*, 2022).

Moreover, object detection frameworks have been used to improve automated biodiversity monitoring systems (Liang *et al.*, 2020). Models based on the YOLO method allow for the simultaneous detection and classification of insects in unconstrained natural situations (Stark *et al.*, 2023). Recent research has shown the use of these frameworks for recognizing flower-visiting arthropods across different stances and backdrops(Stark *et al.*, 2023). This technical breakthrough permits the leap from simple laboratory-based categorization to real-time field application (Liang *et al.*, 2020).

Beyond CNN architectures, the advent of transformer-based models, such as Vision Transformers, has brought a new paradigm to image categorization (Pucci *et al.*, 2023). Different from local convolutional processes, they leverage self-attention mechanisms to capture global dependencies in a picture (Pucci *et al.*, 2023). However, the effectiveness of transformers for fine-grained species identification in data-limited settings is still under investigation (Pucci *et al.*, 2025). Results indicate that alternative designs generally require larger datasets than classic CNNs to attain similar stability in insect classification (Pucci *et al.*, 2023)

In practical conservation implementation, researchers are typically forced to balance the trade-off between classification accuracy and processing demand (Pradnyatama *et al.*, 2025). Compatibility of lightweight designs with field-based identification techniques on mobile platforms has been investigated (Liang *et al.*, 2020). Models have been compared, and MobileNetV2 has been found to provide a particularly efficient tradeoff for real-time processing on portable devices (Pradnyatama *et al.*, 2025). Such efficiency is important for the development of accessible citizen science programs that can be deployed in remote places (Peggie *et al.*, 2025).

The current study fills this gap by a controlled comparison of several deep learning architectures using a handpicked local dataset (Peggie *et al.*, 2025). The study offers an important insight into the possibilities of digital taxonomy in Indonesia by comparing the

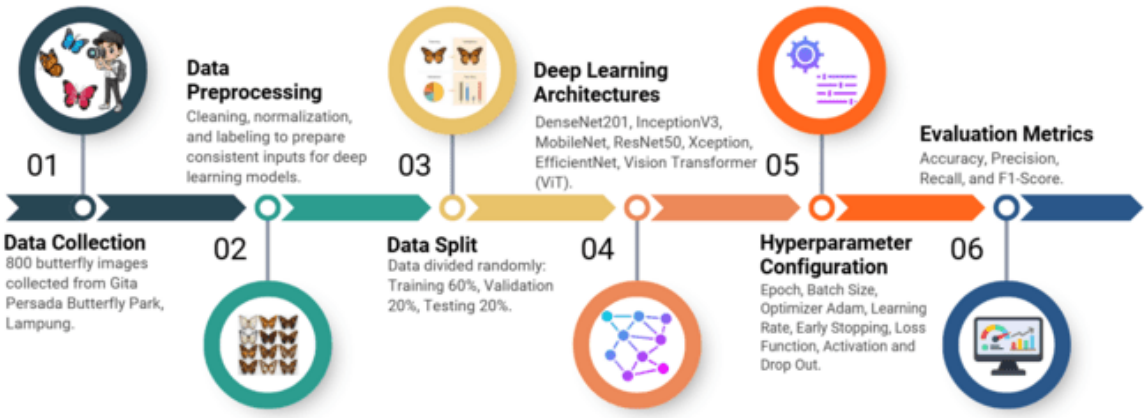
performance of CNN and Vision Transformers (Adityawan *et al.*, 2023). These results provide a proof-of-concept for automated species identification under controlled conditions (Peggie *et al.*, 2025). This effort provides a framework for future large-scale monitoring systems to help preserve Sumatra's unique natural heritage.

3 METHODOLOGY

3.1 Research stages

Figure 1

The research stages of the proposed approach.



The methodology adopted in this study comprises six primary stages that systematically guide the development, training, and evaluation of multiple deep learning architectures for the classification of Sumatra butterfly species. The workflow shown in Figure 1 is the basis for these steps: data collection, data preprocessing, dataset splitting, architectural selection, hyperparameter configuration, and evaluation. The following subsections describe each stage in detail to ensure transparency, reproducibility, and methodological rigour.

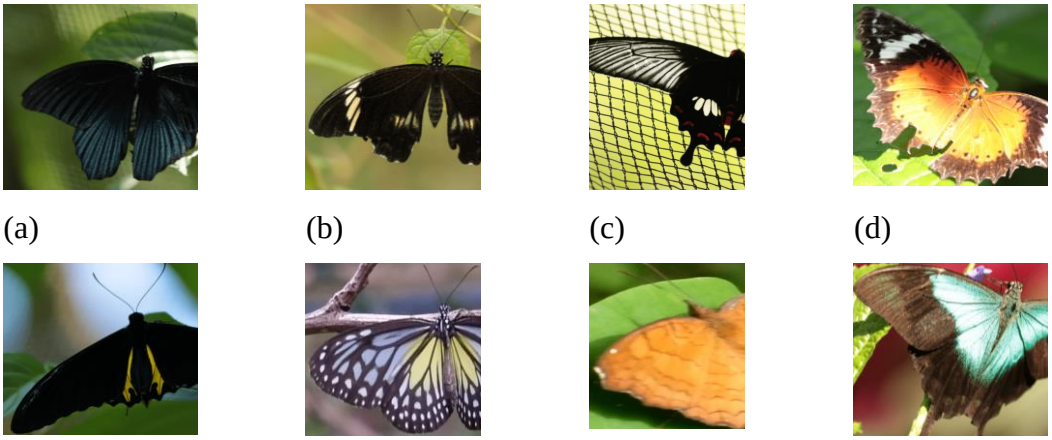
3.2 Data collection

The data of the current study take the form of a manually collected butterfly image dataset from Gita Persada Butterfly Park, Lampung Province, Indonesia. The dataset

consists of 800 high-resolution JPG images of eight butterfly species native to Sumatra: *Papilio memnon*, *Papilio nephelus*, *Pachliopta aristolochiae*, *Cethosia penthesilea*, *Troides helena*, *Papilio peranthus*, *Graphium doson*, and *Graphium sarpedon*. These images were taken at the Gita Persada Butterfly Park using digital cameras. All images were captured using digital cameras under natural lighting conditions. The photographs were taken within a curated butterfly park environment, where individuals were observed in semi-natural resting and flying conditions rather than in fully wild habitats or controlled laboratory settings. This setting allows the dataset to capture variations in wing posture, coloration, background, and partial occlusion while maintaining consistent acquisition conditions. Images were stored in JPG format to ensure compatibility and consistent quality for analysis.

Figure 2

The butterfly species used in this study: (a) Papilio memnon, (b) Papilio nephelus, (c) Pachliopta aristolochiae, (d) Cethosia penthesilea, (e) Troides helena, (f) Papilio peranthus, (g) Graphium doson, and (h) Graphium sarpedon.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2 presents representative images of the eight butterfly species included in this study. The subfigures (a–h) correspond to *Papilio memnon*, *Papilio nephelus*, *Pachliopta aristolochiae*, *Cethosia penthesilea*, *Troides helena*, *Papilio peranthus*, *Graphium doson*, and *Graphium sarpedon*, respectively. These images illustrate typical visual characteristics observed in the dataset, including variations in wing pattern,

coloration, and background conditions. All images were collected from Gita Persada Butterfly Park, Lampung, Indonesia, and form the basis of the dataset used for model evaluation. The dataset was preprocessed and directly used in the experimental procedures described in this study. The dataset used in this study has been made publicly available to support reproducibility and further research. It can be accessed via the following URL: https://ilkom.fmipa.unila.ac.id/research/dataset. The dataset, titled Sumatra Butterfly Species, was accessed on 26 December 2025. The dataset is organized into species-specific directories, where each directory contains high-resolution JPG images corresponding to a single butterfly species. While the dataset is currently hosted on an institutional repository, it does not yet have a DOI. Future work will consider depositing the dataset in a DOI-bearing public repository to ensure long-term accessibility and stable referencing. The dataset is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). In this study, machine learning classes correspond directly to biological species. The term “class” is used to denote the target category in the classification task and should not be confused with the higher-level taxonomic rank.

3.3 Data preprocessing

The preprocessing stage focused on preparing the image dataset for model training using a consistent and reproducible workflow. The dataset was first cleaned by removing duplicate, corrupted, and low-quality images to ensure data consistency. All images were resized to 224 × 224 pixels using nearest neighbor interpolation to match the input requirements of the evaluated models. Pixel values were rescaled from the range of 0–255 to 0–1 and normalized using ImageNet mean and standard deviation. All images were maintained in a three-channel RGB format. The dataset was split into training (60%), validation (20%), and testing (20%) subsets using ImageDataGenerator with ImageNet preprocessing. One-hot encoding was applied to represent class labels, and training data were shuffled to improve model generalization. No background removal or cropping was applied, allowing the models to learn features from the original image context. The preprocessing pipeline included image loading, resizing, normalization, label encoding,

and batch generation for training and evaluation. This standardized preprocessing workflow ensures consistency and reproducibility across all experiments.

3.4 Data split

The dataset was divided into training (60%), validation (20%), and testing (20%) subsets using stratified per-class sampling to preserve class balance across all eight species. Images within each class were shuffled prior to splitting to ensure random distribution while maintaining proportional representation. The final dataset composition consists of 480 training images, 160 validation images, and 160 test images. To evaluate robustness under different partitioning schemes, an additional experiment was conducted using a 70% training, 10% validation, and 20% testing split. Both configurations were applied independently using the same stratification procedure, enabling comparison of model performance across different data distributions while preserving an independent test set. Deduplication was performed prior to data splitting using MD5 hashing to remove exact duplicate files, ensuring that identical images were not distributed across subsets. However, this method is limited to detecting only bit-for-bit identical files and cannot identify near-duplicate images of the same specimen.

3.5 Deep learning architectures

In this study, seven deep learning architectures were selected to evaluate their performance for image-based identification of Sumatra butterfly species: DenseNet201, InceptionV3, MobileNet, ResNet50, Xception, EfficientNetB0, and Vision Transformer (ViT). These architectures were chosen to represent a range of widely used model families with different design characteristics, including residual connections (ResNet50), dense connectivity (DenseNet201), depthwise separable convolutions (MobileNet and Xception), compound scaling (EfficientNetB0), and attention-based mechanisms (ViT). The selection was intended to provide a controlled comparison between convolutional neural network (CNN)-based models and a transformer-based model under a consistent experimental setting, rather than to identify a universally optimal architecture.

DenseNet201 promotes feature reuse via dense connections, facilitating information and gradient flow for more compact models and improved performance. DenseNet201 enhances feature reuse via dense connections, facilitating information and gradient flow for more compact models and improved performance (Yasmin *et al.* 2023). InceptionV3 employs factorised convolutions and auxiliary classifiers to enhance computational efficiency and mitigate overfitting, making it well-suited to diverse image recognition tasks (Spiesman *et al.* 2021). MobileNet, designed for mobile and embedded vision applications, uses depthwise separable convolutions to cut computational costs and model size while retaining competitive accuracy (Adityawan *et al.* 2023). ResNet50 addresses the vanishing gradient problem in deep networks through residual connections, enabling the training of deeper, higher-accuracy models (Karim *et al.* 2024). Xception is characterised by its use of depth-wise separable convolutions and demonstrated strong performance, achieving the highest validation accuracy among the tested models for image classification (Hasan *et al.* 2024). EfficientNetB0 is renowned for its compound scaling method, which uses fewer parameters to achieve high performance by consistently scaling network depth, width, and resolution. The ViT, a newer architecture, processes images by dividing them into fixed-size patches, treating them as sequences, and using self-attention to capture long-range dependencies, unlike CNNs' local operations (Pucci *et al.* 2025).

3.6 Hyperparameter configuration

For the convolutional neural network architectures (InceptionV3, Xception, and DenseNet201), the pre-trained convolutional base (ImageNet weights) was frozen during training, and only the classification head layers were updated. For the Vision Transformer (ViT), a partial freezing strategy was applied, where the patch embedding and initial transformer layers were kept frozen, and the final layers were fine-tuned. Hyperparameter values were determined through a limited empirical tuning process involving five preliminary training runs per architecture. The tuning focused on key parameters, particularly learning rate and dropout, using manual adjustment based on validation performance. This should be interpreted as a constrained tuning strategy rather than an exhaustive search.

The final hyperparameter settings are summarized in Table 1. The AdamW optimizer was used for all models, with learning rates set to 0.00001 for most architectures, 0.00003 for DenseNet201, and 0.0001 for ViT. A ReduceLROnPlateau scheduler (patience = 3) and early stopping (patience = 5) were applied. The batch size was fixed at 32, and training was performed for a maximum of 20 epochs. Dropout was applied in the classification layers, with values of 0.3 for MobileNet, ResNet50, and EfficientNetB0; 0.5 for InceptionV3 and Xception; and 0.2 for ViT. The loss function used was categorical cross-entropy, and all models were trained with a fixed random seed of 42. Experiments were conducted on Google Colab using an NVIDIA Tesla T4 GPU. TensorFlow was used for CNN-based models, while PyTorch was used for ViT.

Table 1

Standardised Training Configuration.

Hyperparameter	Configuration Detail
Optimizer	AdamW
Learning Rate	0.00001 (majority) / 0.00003 (DenseNet201) / 0.0001 (ViT)
LR Scheduler	ReduceLROnPlateau (patience 3)
Early Stopping	Yes (patience 5)
Batch Size	32
Maximum Epochs	20
Loss Function	Categorical Cross-Entropy
Dropout Rate	0.3 (majority) / 0.5 (InceptionV3 and Xception) / 0.2 (ViT)
Random Seed	42
Framework	TensorFlow / PyTorch (specifically for ViT)
Hardware	NVIDIA Tesla T4 GPU

3.7 Evaluation metrics

The concluding phase assesses the trained models using four standard performance metrics: Accuracy, Precision, Recall, and F1-score (Ali *et al.* 2022). These metrics are derived from the confusion matrix based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy measures the proportion of correctly classified images relative to the total number of samples. Precision reflects the proportion of correct positive predictions, while Recall measures the proportion of actual positive cases correctly (Ong & Hamid, 2022). The F1-score combines Precision and Recall into a single metric (Alzubaidi *et al.* 2021). The Accuracy metric is defined in Equation (1), the Precision metric is presented in Equation

(2), the Recall metric is shown in Equation (3), and the F1-score metric is described in Equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

4 RESULTS AND DISCUSSION

4.1 Design experiment

This section outlines the experimental protocol and methodological framework used to assess the relative performance of diverse deep learning architectures for classifying Sumatra butterfly species. The design not only incorporates hyperparameter configurations but also systematically quantifies computational metrics, including training time, inference time, total parameters, and model size, essential for evaluating practical deployability. The judiciously selected hyperparameters and experimental configurations, uniformly applied across all models, are summarised in Table 1, providing a clear outline of the methodology.

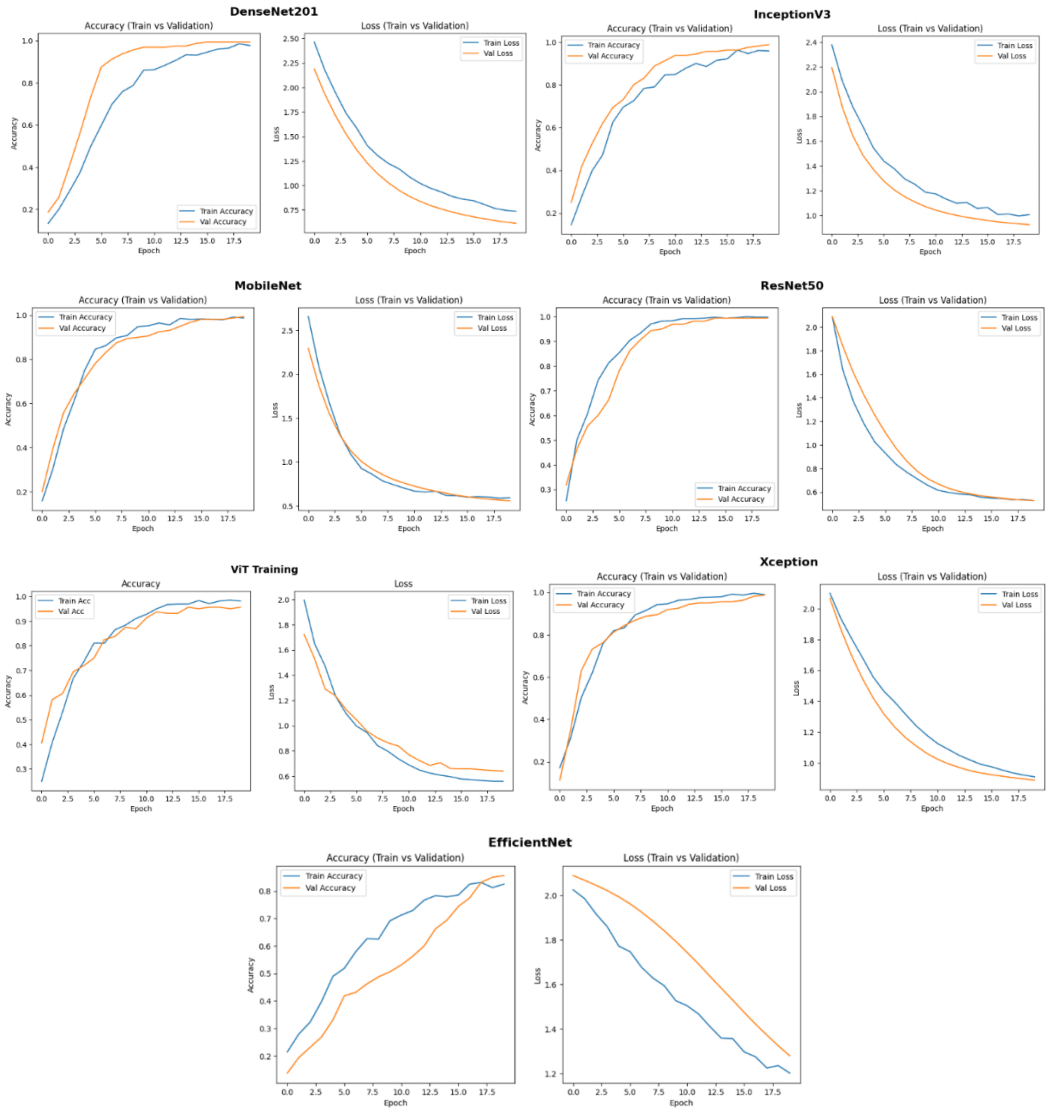
4.2 Evaluation results

The training and validation accuracy and loss curves (Figure 3) were analyzed to assess learning behavior and convergence across the evaluated models. Overfitting was examined based on the divergence between training and validation curves, rather than on absolute performance values. Among the CNN-based models, DenseNet201, MobileNet,

and Xception showed stable learning patterns. For these models, training and validation accuracy curves increased in parallel, with only minor gaps observed between them. The corresponding loss curves decreased consistently without significant divergence, indicating stable convergence within the training epochs.

Figure 3

The accuracy and loss curves across the deep learning model.

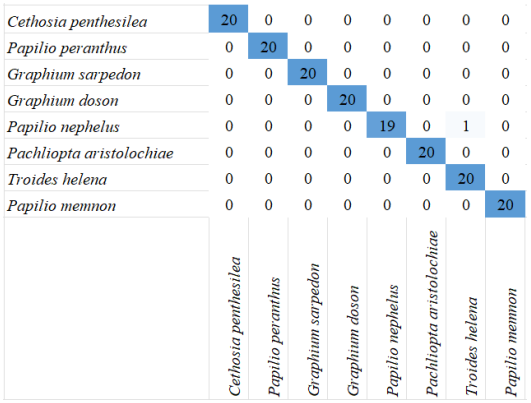


ResNet50 and InceptionV3 exhibited similar trends. While both models achieved high accuracy, a slightly larger gap between training and validation loss was observed in InceptionV3 during later epochs, suggesting a moderate degree of overfitting.

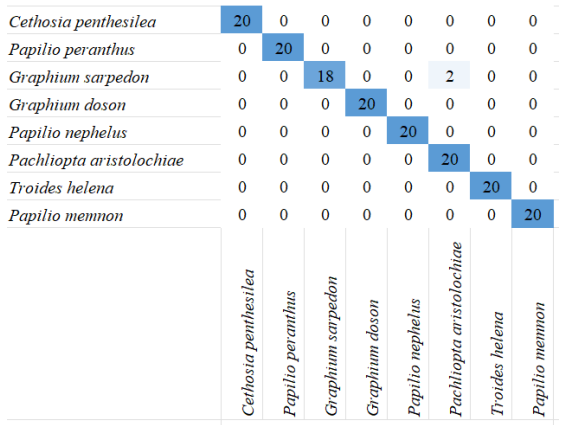
EfficientNetB0 showed slower convergence, with validation accuracy lagging behind training accuracy during early and mid-training phases. The Vision Transformer (ViT) demonstrated a different pattern. A consistent gap between training and validation accuracy curves was observed, with validation accuracy remaining lower throughout training. This pattern, together with a similar divergence in loss curves, indicates a higher sensitivity to overfitting under the given dataset size. Given the limited test set (20 images per class), high accuracy values should be interpreted with caution, as small absolute errors can lead to noticeable percentage changes. Early stopping (patience = 5) limited further divergence between training and validation performance across models, contributing to stable training behavior under a fixed number of epochs.

Figure 4

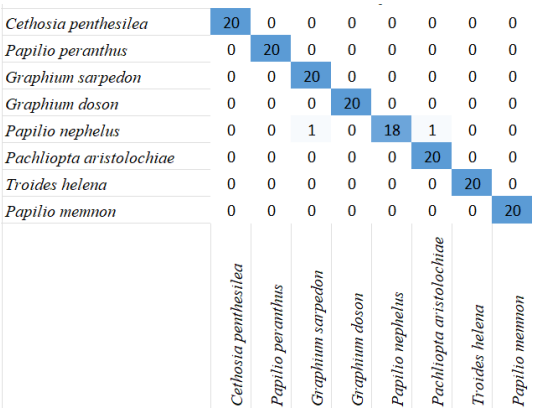
The confusion matrix shows the top three models (DenseNet201, ResNet50, and Xception) and the lowest-performing model (EfficientNetB0).



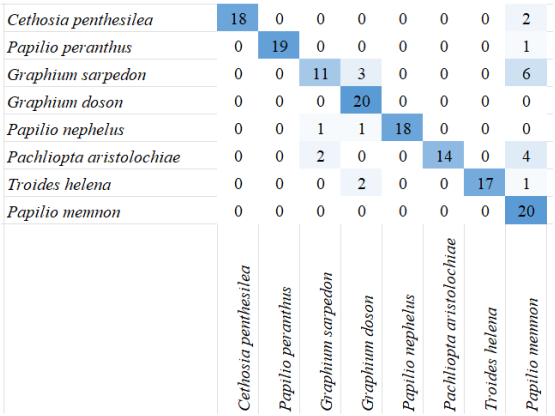
(a) The confusion matrix of DenseNet201



(b) The confusion matrix of ResNet50



(c) The confusion matrix of Xception



(d) The confusion matrix of EfficientNetB0

The classification performance of the evaluated deep learning architectures is visually summarised through the confusion matrices illustrated in Figure 4. These matrices provide a granular view of each model's ability to discriminate between the eight Sumatra butterfly species, where rows represent true labels and columns denote predicted labels. By synthesising results from the top-performing architectures DenseNet201, ResNet50, and Xception alongside EfficientNetB0, a clearer interpretation of model reliability and common misclassification patterns is established.

Among the superior architectures, DenseNet201 (Figure 4a) showed strong performance. As detailed in Table 2, it achieved a perfect F1-Score of 1.00 for nearly all species, including *Cethosia penthesilea, Papilio peranthus, Graphium sarpedon, Graphium doson, Pachliopta aristolochiae*, and *Papilio memnon*. Its only deviation was observed in *Papilio nephelus*, where a single instance was misclassified as *Troides helena*, resulting in a recall of 0.95 for the former and a slight dip in precision for the latter. This near-flawless performance suggests that DenseNet201's dense connectivity is effective at capturing the subtle, fine-grained phenotypic markers required for high-accuracy identification.

ResNet50 (Figure 4b) and Xception (Figure 4c) also maintained high performance, though they exhibited specific localised confusion. ResNet50 encountered difficulty with *Graphium sarpedon*, misidentifying two instances as *Pachliopta aristolochiae*, which led to a recall of 0.90. Xception followed a similar pattern of minor errors, primarily with *Papilio nephelus*, where specimens were erroneously assigned to *Graphium sarpedon* and *Pachliopta aristolochiae*. Despite these slight variances, both architectures resolved the majority of morphological traits successfully, maintaining F1-Scores above 0.94 across all classes.

In contrast, EfficientNetB0 (Figure 4d) showed weaker performance in distinguishing species with similar visual patterns. The model struggled particularly with

Table 2

Class-wise precision, recall, F1-Score, and support obtained using EfficientNet-B0.

No.	Class	Precision	Recall	F-Score	Support
1	*Cethosia penthesilea*	1.00	0.90	0.94	20
2	*Papilio peranthus*	1.00	0.95	0.97	20
3	*Graphium sarpedon*	0.78	0.55	0.64	20

4	*Graphium doson*	0.76	1.00	0.86	20
5	*Papilio nephelus*	1.00	0.90	0.94	20
6	*Pachliopta aristolochiae*	1.00	0.70	0.82	20
7	*Troides helena*	1.00	0.85	0.91	20
8	*Papilio memnon*	0.58	1.00	0.74	20

Graphium sarpedon, which obtained a recall of only 0.55. In addition, *Papilio memnon* became the dominant false-positive class, with multiple samples from different species incorrectly assigned to it, reducing its precision to 0.58. This suggests that EfficientNetB0 was less effective in extracting discriminative features for species with similar wing coloration and patterns.

Table 3 presents the computational metrics of all evaluated architectures, including training time, inference time, parameter count, and model size, enabling direct comparison of computational efficiency. Among the evaluated models, MobileNet demonstrates the most efficient configuration, with the lowest parameter count (3.37 million) and the smallest model size (12.84 MB). It also achieves the fastest inference time (0.025 s), indicating its suitability for resource- constrained environments.

Table 3

The computational metrics of seven deep learning architectures.

Model	Training Time (s)	Inference Time (s)	Parameter (Million)	Size (MB)
DenseNet201	1435	0.145	18.34	69.95
InceptionV3	1656	0.040	21.82	83.23
MobileNet	1243	0.025	3.37	12.84
ResNet50	1685	0.037	23.60	90.04
ViT	1941	0.001	5.53	21.95
Xception	1378	0.039	20.88	79.64
EfficientNetB0	1416	0.049	4.06	15.49

In contrast, deeper architectures such as ResNet50 and DenseNet201 require substantially higher computational resources. ResNet50 has the largest parameter count (23.60 million) and model size (90.04 MB), while DenseNet201 also shows high computational demand with a training time of 1435 seconds and a model size of 69.95 MB. Although these larger models achieve slightly higher classification accuracy, their increased computational cost highlights a clear trade-off between predictive performance and efficiency. The Vision Transformer (ViT) presents a different profile, with relatively low parameter count (5.53 million) but the longest training time (1941 seconds),

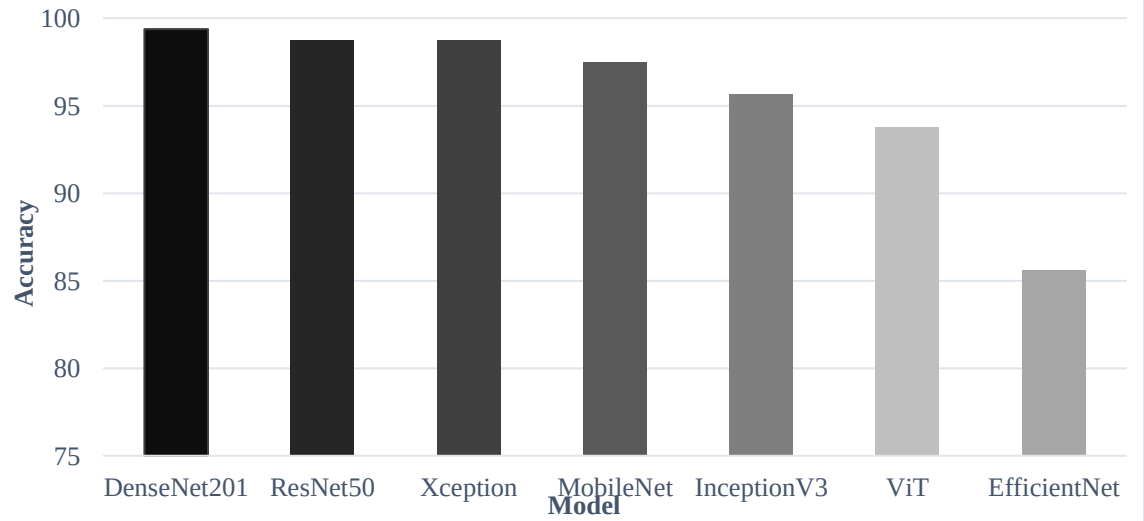
reflecting its higher computational complexity during training despite fast inference. Overall, these results support the observation that MobileNet provides a favorable balance between classification performance and computational efficiency under the experimental conditions of this study.

The results show that the evaluated models achieved varying levels of performance. As shown in Figure 5, DenseNet201 achieved the highest accuracy (99.38%), followed by

ResNet50 and Xception (98.75%), and MobileNet (97.50%), while EfficientNetB0 showed the lowest performance (85.63%). These differences should be interpreted within the constraints of the dataset size and experimental setting. CNN-based architectures performed well under the conditions of this study. In contrast, EfficientNetB0 achieved lower accuracy (85.63%), which may be related to the limited dataset size and the absence of data augmentation. Under these conditions, the model may not fully capture fine-grained visual differences between species. This suggests that, under constrained data conditions, simpler CNN architectures can perform more consistently than more complex models. To ensure the validity and robustness of the reported performance metrics, this study conducted repeated experimental trials and implemented rigorous data integrity controls.

Figure 5

The accuracy achieved by the deep learning architectures.



These measures are essential to mitigate the risk of overfitting and to confirm that the high accuracy achieved is not a result of statistical anomalies. As detailed in the methodology section, data partitioning was conducted using a leakage-resistant splitting strategy, such as stratified sampling (or subject-level separation). This approach ensures that data from the same individual or source does not appear in both the training and testing sets simultaneously. By maintaining this strict independence, the model is evaluated on genuinely unseen data, thereby addressing concerns regarding potential data leakage.

4.3 Discussion

We intentionally did not apply background removal or cropping to retain the visual context in which butterflies naturally occur. In field conditions, butterflies are observed against heterogeneous backgrounds, under varying lighting, and often with partial occlusion. Preserving these characteristics allows the models to learn from more realistic visual scenarios. However, it is important to note that all evaluations were conducted on curated images from a butterfly park rather than independent field data. Therefore, this setting should be considered a controlled experimental environment rather than a direct representation of natural conditions.

The comparative evaluation shows that convolutional neural network (CNN) architectures, particularly DenseNet201, ResNet50, Xception, and MobileNet, achieved higher classification performance under the current dataset conditions. DenseNet201 reached the highest accuracy (99.38%), followed by ResNet50 and Xception (98.75%). MobileN*et al*so performed well (97.50%) while maintaining a substantially smaller model size and lower computational cost. These results should be interpreted with caution. The dataset is relatively small, limited to a single location, and contains only eight species. Under such controlled conditions, high accuracy does not necessarily indicate equivalent performance in more complex ecological settings. The results are therefore best understood as a benchmark within a constrained experimental scenario.

The observed performance differences align with known architectural characteristics. CNN-based models benefit from locality and hierarchical feature extraction, which are advantageous for fine-grained visual discrimination such as

butterfly wing patterns. In contrast, the Vision Transformer (ViT) relies on global attention mechanisms and typically requires larger datasets to learn robust representations. The lower performance of ViT in this study is therefore likely related to dataset size and training configuration, rather than indicating a general limitation of transformer-based models.

From a computational perspective, MobileNet offers a favorable balance between accuracy and efficiency. It has the smallest number of parameters (3.37 million) and the lowest model size (12.84 MB), while maintaining competitive performance. This suggests potential suitability for resource-constrained environments. However, this should not be interpreted as evidence of deployment readiness, as no evaluation on mobile devices or field conditions was conducted.

Differences in training efficiency were also observed. MobileNet required the shortest training time (1243 s), while the Vision Transformer required the longest (1941 s), despite having fewer parameters than some CNN models. This indicates that transformer-based architectures may incur higher computational cost during training in limited-data settings. Several limitations should be considered. First, all images were collected from a single site (Gita Persada Butterfly Park), which may introduce bias in background, lighting, and environmental conditions. Second, the dataset includes only eight species, limiting taxonomic diversity. Third, although specimen-level separation was applied, multiple images were captured under similar conditions, which may reduce intra-class variability. Fourth, no external dataset or field-based validation was performed. These factors limit the generalizability of the results.

Data augmentation was not applied and was treated as a baseline design choice to enable controlled model comparison, although this may limit generalizability under diverse visual conditions. Model performance may also decrease in real-world environments due to variations in lighting, camera devices, backgrounds, and phenotypic differences. Therefore, the reported results should not be directly generalized to field applications. Future studies should investigate augmentation strategies, external validation using datasets from multiple locations, and evaluation on mobile or edge devices to assess model robustness and practical feasibility for automated butterfly species identification.

5 CONCLUSIONS

The study evaluated several deep learning architectures for image-based identification of eight Sumatra butterfly species using a controlled dataset from Gita Persada Butterfly Park. The results showed that transfer learning with pretrained models achieved high classification performance, with DenseNet201 showing the best overall performance, while MobileNet provided a good balance between classification performance and computational efficiency. CNN-based architectures performed more consistently than the Vision Transformer model under the current dataset conditions. However, the study was limited to a small single-site dataset without external or field-based validation. Therefore, the findings should be considered preliminary benchmarks, and further studies using larger and more diverse datasets are needed to improve model robustness and generalizability.

REFERENCES

Adityawan, H. T., Farroq, O., Santosa, S., Islam, H. M. M., Sarker, M. K., & Setiadi, D. R. I. M. (2023). Butterflies Recognition using Enhanced Transfer Learning and Data Augmentation. *Journal of Computing Theories and Applications*, *1*(2), 115–128.

Ali, K., Shaikh, Z. A., Khan, A. A., & Laghari, A. A. (2022). Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer. In *Neuroscience Informatics*, *2*(4).

Almryad, A. S., & Kutucu, H. (2020). Automatic identification for field butterflies by convolutional neural networks. *Engineering Science and Technology, an International Journal*, *23*(1), 189–195.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53.

Fathimathul, R. P. P., Orban, R., Vadivel, K. S., Subramanian, M., Muthusamy, S., Elminaam, D. S. A., Nabil, A., Abulaigh, L., Ahmadi, M., & Ali, M. A. S. (2022). A Novel Method for the Classification of Butterfly Species Using Pre-Trained CNN Models. *Electronics (Switzerland)*, *11*(13).

Hasan, M. R., Rahman, M. M., Shahriar, F., Khan, M. S. I., Mohi Uddin, K. M., & Hasan, M. M. (2024). Smart farming: Leveraging IoT and deep learning for sustainable tomato cultivation and pest management. *Crop Design*, *3*(4).

Ilhamdi, M. L., Al Idrus, A., & Santoso, D. (2018). Diversity of Species and Conservation Priority of Butterfly at Suranadi Natural Park of West Lombok, Indonesia. *Biosaintifika*, *10*(1), 48–55.

Karim, A. A. J., Mahmud, M. Z., & Khan, R. (2024). Advanced vision transformers and open-set learning for robust mosquito classification: A novel approach to entomological studies. *PLOS Computational Biology*, *20*(12).

Koneri, R., Maabuat, P. V., & Nangoy, M. J. (2020). The distribution and diversity of butterflies (Lepidoptera: Rhopalocera) in various urban forests in north Minahasa regency, north Sulawesi province, Indonesia. *Applied Ecology and Environmental Research*, *18*(2), 2295–2314.

Liang, B., Wu, S., Xu, K., & Hao, J. (2020). Butterfly Detection and Classification Based on Integrated YOLO Algorithm, *Genetic and Evolutionary Computing* (500–512).

Ong, S. Q., & Hamid, S. A. (2022). Next generation insect taxonomic classification by comparing different deep learning algorithms. *PLoS ONE*, *17*(12).

Peggie, D., Prabowo, S. W. B., Shahroni, A. M., Shidiq, F. I. A., Irwansyah, L., Soenarko, S., Rahma, N., & Wafa, I. Y. (2025). Kupunesia App For Citizen Science: New Way of Mainstreaming Interest and Study of Indonesian Butterflies. *Treubia*, *49*(2), 137–148.

Pradnyatama, M., Sari, C. A., Rachmawanto, E. H., & Islam, H. M. M. (2025). A Comparative Analysis of Convolutional Neural Network (CNN): MobileNetV2 and Xception for Butterfly Species Classification. *Jurnal Masyarakat Informatika*, *16*(1), 69–90.

Pucci, R., Kalkman, V. J., & Stowell, D. (2023). *Comparison between transformers and convolutional models for fine-grained classification of insects*.

Pucci, R., Kalkman, V. J., & Stowell, D. (2025). Performance of Computer Vision Algorithms for Fine-Grained Classification Using Crowdsourced Insect Images. *IET Computer Vision*, *19*(1).

Sagar, V., Sachin, R., Chandrashekara, K., & Ganeshaiah, K. (2020). Identification of Indian Butterflies and Moths with Deep Convolutional Neural Networks x. *Current Science*, *118*, 1456.

Spiesman, B. J., Gratton, C., Hatfield, R. G., Hsu, W. H., Jepsen, S., McCornack, B., Patel, K., & Wang, G. (2021). Assessing the potential for deep learning and computer vision to identify bumble bee species from images. *Scientific Reports*, *11*(1).

Stark, T., Ştefan, V., Wurm, M., Spanier, R., Taubenböck, H., & Knight, T. M. (2023). YOLO object detection models can locate and classify broad groups of flower-visiting arthropods in images. *Scientific Reports*, *13*(1).

Yasmin, R., Das, A., Rozario, L. J., & Islam, M. E. (2023). Butterfly detection and classification techniques: A review. In *Intelligent Systems with Applications* (18).