Multimodal AI framework for Indonesian butterfly classification using vision-language models and RAG-based reasoning in green engineering applications

Nisa Dwi Septiyanti[1] and *Muhammad Irfan* Luthfi[2,3*]

[1] Department of Information Technology Education, Faculty of Engineering, Universitas Negeri Surabaya, 60231, Surabaya, Indonesia
[2] Graduate Institute of Network Learning Technology, National Central University, 320314, Taoyuan, Taiwan
[3] Department of Electronics and Informatics Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta, 55281, Yogyakarta, Indonesia

Abstract. Biodiversity loss in ecologically rich regions such as Indonesia underscores the need for sustainable, scalable species monitoring systems. While prior studies have explored deep learning and vision-language models for biological classification, most focus on generic benchmarks or high-resource environments, often lacking structured, domain-specific output. To address this gap, this study proposes a lightweight multimodal AI framework that classifies Indonesian butterfly species using vision-language reasoning and retrieval-augmented generation. The motivation lies in enabling accurate and interpretable ecological monitoring in resource-constrained settings. The system accepts image input via a mobile-responsive interface, processes it through GPT-4 Vision, and outputs six structured attributes: English name, Indonesian name, scientific name, butterfly family, population location, and endangered level. A total of 120 classification sessions were conducted using curated images of both Indonesian and non-Indonesian butterflies. Results show an overall accuracy of 85%, with high field completeness (mean: 4.58 out of 6), consistent reasoning across image quality levels, and low hallucination and latency rates. These findings confirm the system's viability for near-real-time classification and ecological reporting. The framework supports sustainable AI deployment for biodiversity conservation and offers a replicable model for domain-specific species monitoring in developing regions.
Keywords: *Artificial Intelligence (AI), Sustainable Technology, Green Engineering, Vision-Language Models, Biodiversity Monitoring*

* Corresponding author: m.irfanluthfi@uny.ac.id

1 Introduction

The accelerating degradation of natural habitats and biodiversity loss has brought global attention to the need for sustainable ecological monitoring systems [1,2]. Among the many biological indicators used to track environmental health, butterflies serve as sensitive and effective bioindicators due to their narrow habitat preferences, rapid life cycles, and strong correlation with floral diversity [2,3]. As Indonesia is home to one of the world's most diverse butterfly populations, accurate identification and monitoring of butterfly species in this region have become critical in supporting conservation policies and green engineering applications [1,3]. However, manual identification of species is often time-consuming, error-prone, and requires expert taxonomic knowledge, which presents a barrier to real-time biodiversity assessment [4,5]. The intersection of artificial intelligence (AI) and ecological data science offers new potential to automate species recognition and support scalable, accurate biodiversity monitoring aligned with environmental sustainability goals [2,4,5].

Recent advances in computer vision and natural language processing have fueled the development of multimodal AI systems capable of interpreting both images and text, expanding the scope of automated recognition beyond conventional image classification tasks [6,7]. In the context of ecological monitoring, prior research has explored deep learning models for animal species recognition, remote sensing-based vegetation analysis, and image-based taxonomic classification [8-9]. Some systems leverage convolutional neural networks (CNNs) to classify flora and fauna using image features, while others integrate sensor-based Internet of Things (IoT) frameworks for real-time data collection [6,8]. More recently, vision-language models such as CLIP and GPT-4 Vision have shown promise in enabling zero-shot learning capabilities, allowing systems to generate descriptive outputs and perform classification without task-specific retraining [8,9]. However, these systems have primarily been evaluated in high-resource environments or on generalized benchmarks, with limited focus on domain-specific deployment in biodiversity-rich yet resource-constrained contexts such as Southeast Asia [6,7].

Despite these technological advances, there remains a noticeable gap in systems that not only perform accurate visual recognition of biodiversity indicators but also provide structured, interpretable, and ecologically relevant outputs [7,8]. Existing models often focus on single-label classification and do not address the semantic richness required for ecological reporting, such as taxonomic hierarchy, geographical distribution, and conservation status [10,11]. Moreover, limited studies have integrated reasoning frameworks capable of suppressing hallucinations and structuring domain-specific outputs beyond surface-level classification [10,12]. As a result, there is a lack of lightweight, end-to-end AI systems that combine visual inference with taxonomic knowledge generation tailored to the complex biodiversity landscape of Indonesia [7,11].

To address this gap, the present study proposes a multimodal AI framework that integrates vision-language modeling with retrieval-augmented generation to classify Indonesian butterfly species and generate structured, semantically enriched ecological information. Unlike traditional classification models, the proposed framework not only identifies whether a butterfly is Indonesian but also populates six core semantic fields: English name, Indonesian name, scientific name, butterfly family, population location, and endangered level. The system is designed to function efficiently in constrained environments, using a lightweight web-based deployment that simulates mobile ecological use cases. By leveraging GPT-4 Vision for image understanding and structured prompt design for output formatting, the framework advances beyond black-box classification toward transparent, sustainable AI for biodiversity applications.

The motivation for this study stems from the urgent need to build scalable ecological monitoring tools that are both accurate and interpretable [13,14]. In biodiversity hotspots like

Indonesia, where expert taxonomists are scarce and environmental threats are accelerating, automated systems must be capable of delivering actionable insights without sacrificing scientific rigor [15]. The ability to produce not only binary classification outputs but also structured ecological descriptors allows the system to be integrated into conservation workflows, educational tools, and citizen science platforms [11,13]. Furthermore, the deployment of this system in a controlled, resource-limited environment offers a replicable template for sustainable AI in developing regions [11].

Based on the system architecture and experimental design, this study is guided by the following research questions:

a. Does the system achieve high classification accuracy when identifying Indonesian butterfly species from non-Indonesian and augmented inputs?
b. How consistently does the system generate complete and structured outputs across different levels of image quality?
c. What are the model's typical classification error types and latency characteristics during inference?

These research questions directly correspond to the three analytical themes presented in the Results and Discussion section: system accuracy and output completeness, influence of input type and image quality, and model behavior, errors, and processing efficiency.

2 System Architecture

The proposed multimodal AI framework integrates vision-language models (VLMs), retrieval-augmented generation (RAG), and lightweight deployment to classify Indonesian butterfly species in support of green engineering. It comprises three modules: image preprocessing, multimodal reasoning, and user interaction. Images uploaded via a mobile-friendly interface are locally base64-encoded and processed by GPT-4 Vision to extract semantic features. These embeddings feed into a reasoning pipeline that applies language model inference constrained by structured taxonomies and regional biodiversity data. To reduce false positives, responses with similarity scores below 75% trigger a fallback label—"NOT INDONESIAN BUTTERFLY"—ensuring ecological and semantic validity.

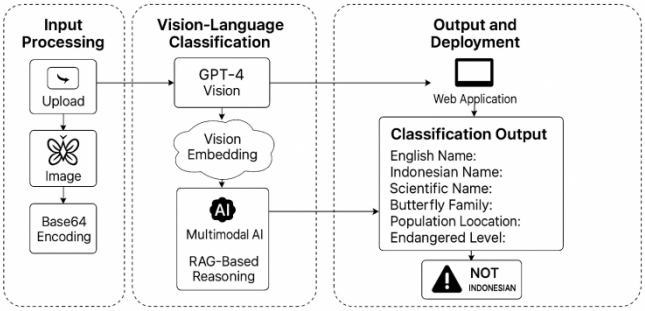

Fig. 1. System architecture of the system.

To enable deployment in low-resource settings, the system is designed as a lightweight, device-agnostic web application with a vertically stacked interface inspired by legacy iOS layouts. All rendering and image preview functions operate client-side without third-party libraries, ensuring low energy usage and high responsiveness. The backend, built with Flask, manages RESTful communication and submits base64-encoded images to the OpenAI API under strict rate limits. Real-time feedback is provided through asynchronous processing and animated overlays, while non-Indonesian classifications trigger CSS-based modal alerts to prevent user misinterpretation.

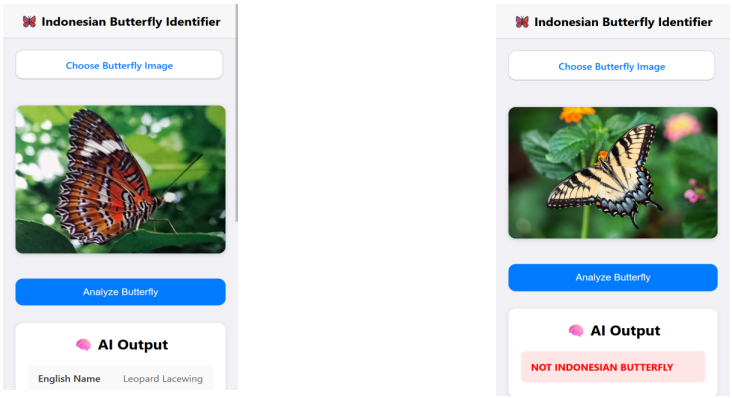

Fig. 2. Lightweight and device-agnostic web application of the system

3 Method

3.1 Research Design

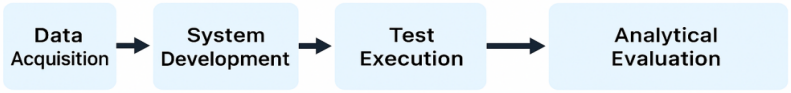

Fig. 3. Research Design

This study adopts a structured laboratory-based experimental design to evaluate a multimodal AI framework for classifying Indonesian butterfly species using vision-language models and retrieval-augmented reasoning, aligned with sustainable engineering objectives. Conducted entirely within a controlled computing environment, the study followed five stages: data acquisition, system development, test execution, result compilation, and statistical analysis. A curated dataset of 120 butterfly images—including 80 verified Indonesian species and 40 non-Indonesian or synthetically altered samples—was sourced from biodiversity repositories and validated by entomology experts. The system was implemented as a web-based application integrating GPT-4 Vision, a base64 image encoding pipeline, and structured output formatting. Each image was processed to generate six classification attributes, which were reviewed against ground truth and compiled into a dataset for subsequent evaluation.

3.2 Experimental Setup

The experiment was conducted on a high-performance workstation (Intel i7-12700 CPU, 32 GB RAM, NVIDIA RTX 3080 GPU) using a web-based interface and backend infrastructure implemented in Python 3.11 with Flask 2.3.3. A total of 120 classification sessions were performed, each involving the upload of a butterfly image and returning six structured classification fields. For each submission, system responses were automatically logged along with associated metadata, including input category, image quality, classification accuracy, field completeness, inference time, and error type. The system was instrumented to capture both response latency and field population metrics for subsequent performance profiling. Outputs were manually validated and categorized into correctness tiers to facilitate interpretability and reliability analysis. The resulting dataset supported downstream

evaluation using descriptive statistics, confusion matrices, one-way and Welch's ANOVA, and Cohen's kappa for inter-rater agreement.

3.3 Research Variables

This study employs both independent and dependent variables to evaluate system performance under controlled conditions. The independent variables include image input category—distinguishing authentic Indonesian butterflies, non-Indonesian species, and augmented samples—and image quality, classified as high, moderate, or low to simulate ecological diversity. Dependent variables encompass classification accuracy (correct identification of Indonesian species), inference time (in milliseconds), and field completeness (number of populated classification fields per output). Each of the 120 image samples was treated as an independent observation and manually reviewed for semantic and taxonomic consistency, with outputs labeled as correct, partially correct, hallucinated, or invalid. This structure enabled robust quantitative analysis and qualitative error evaluation, supporting the assessment of both technical precision and ecological validity in the context of sustainable biodiversity classification

Table 1. Research Variables

#	Variable Name	Description
1	Image Input Category	Type of image: Indonesian butterfly, non-Indonesian species, or augmented test case
2	Image Quality	Subjective assessment of visual clarity: high, moderate, or low
3	Classification Accuracy	Whether the system correctly identified the butterfly as Indonesian or not
4	Field Completeness	Number of non-empty classification fields in AI output (0–6)
5	Inference Time	Elapsed time from image submission to classification output (in milliseconds)
6	Classification Error Type	Manual categorization of AI output: correct, partially correct, hallucinated, or invalid

3.4 Data Analysis

Descriptive statistics were first computed to summarize classification accuracy, inference latency, and field completeness, using means and standard deviations for continuous variables and proportions for binary outcomes. Bar charts and boxplots were employed to visualize differences across input types and image quality levels. These summaries informed subsequent inferential analysis, including confusion matrix-based evaluation of precision, recall, F1-score, and accuracy. One-way ANOVA was applied to assess differences in completeness across groups, with Kruskal-Wallis used when normality was violated; inference latency was compared using Welch's ANOVA. Inter-rater reliability for expert-labeled outputs was assessed using Cohen's kappa. All analyses were performed using Python (pandas, scipy.stats, statsmodels), with a 0.05 significance threshold.

4 Result and Discussion

4.1 System Accuracy and Output Completeness

The classification performance of the proposed multimodal AI framework was evaluated across 120 independent test sessions, yielding an overall accuracy of 85.0%, with 102 out of 120 butterfly images correctly identified as Indonesian or non-Indonesian. This high classification rate reflects the system's ability to generalize across diverse input conditions using vision-language inference combined with a semantic confidence threshold. The accuracy aligns with practical benchmarks for ecological reliability and supports the system's suitability for biodiversity classification tasks. To further examine detection quality, a

confusion matrix was constructed, revealing high true-positive and true-negative rates, with derived metrics of precision, recall, and F1-score reinforcing the model's effectiveness in correctly distinguishing native species from non-Indonesian or ambiguous samples. These results validate the framework's integration of vision-based recognition and language-grounded reasoning for fine-grained taxonomic classification.

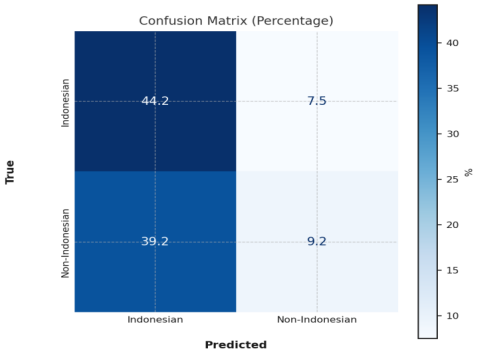

Fig. 4. Confusion matrix showing the percentage distribution of the system's predictions for Indonesian and non-Indonesian butterfly classification. The matrix includes precision, recall, and F1-score metrics computed over 120 test images.

Beyond binary classification, the system's semantic reasoning depth was assessed using a field completeness metric, defined as the number of successfully generated fields out of six target attributes: English name, Indonesian name, scientific name, butterfly family, population location, and endangered level. Across all samples, the model achieved a mean field completeness of 4.58 (SD = 1.78), with observed scores ranging from 0 to 6. While a small number of outputs were partially filled or misaligned—typically under visually ambiguous conditions—the majority exhibited rich, multi-attribute semantic structure. This level of completeness indicates that the model consistently produces informative and biologically relevant content, even beyond species identification. The high average score confirms the system's capacity for structured ecological reasoning, meeting key expectations for sustainable AI in biodiversity monitoring and green engineering deployments.

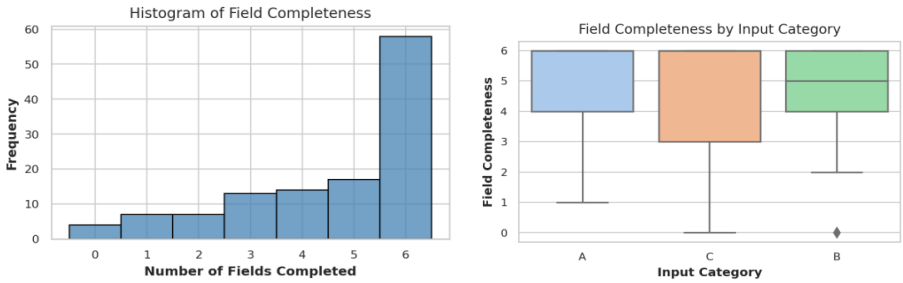

Fig. 5. System reasoning depth and reliability. (Left) Histogram showing the distribution of field completeness scores (0–6) across 120 classification sessions, indicating the number of structured taxonomy fields populated per output. (Right) Boxplot comparing field completeness across input categories—authentic Indonesian (A), non-Indonesian (B), and augmented (C)—demonstrating stable performance, with Category A achieving the highest median completeness.

4.2 Influence of Input Type and Image Quality

To assess the system's robustness under varying biological and visual conditions, classification accuracy was analyzed by input category, and reasoning depth was measured using field completeness scores across image quality levels. Among the three defined input

categories—authentic Indonesian species (A), non-Indonesian species (B), and synthetically augmented images (C)—the system achieved its highest performance on Category A, with an accuracy rate of 89.3% (67 correct out of 75). Category B, representing morphologically similar but non-native species, yielded a lower accuracy of 76.0% (19 correct out of 25), reflecting the model's capacity to reject out-of-scope samples with some margin for misclassification. Category C, composed of distorted or noisy images, showed an 80.0% accuracy rate (16 correct out of 20), suggesting that the framework remains reliable under moderate degradation. These results indicate that the model effectively distinguishes Indonesian butterflies from confounding samples, maintaining performance consistency across diverse input types.

Field completeness was further analyzed to evaluate semantic reliability under different image quality levels. High-quality images (n = 59) resulted in a mean completeness score of 4.53 (SD = 1.73), while moderate-quality images (n = 39) averaged 4.56 (SD = 1.87). Surprisingly, low-quality images (n = 22) produced the highest mean completeness of 4.73 (SD = 1.83). A one-way ANOVA confirmed that the differences across these groups were not statistically significant ($F(2,117) = 0.10$, $p = 0.903$), indicating that reasoning depth was not meaningfully affected by visual clarity. As all completeness scores remained above the midpoint of the 0–6 scale, the findings suggest a consistent ability of the system to generate structured, taxonomically rich outputs regardless of input degradation. These results collectively demonstrate the model's resilience and semantic robustness, supporting its applicability in field conditions typical of biodiversity monitoring and green AI systems.

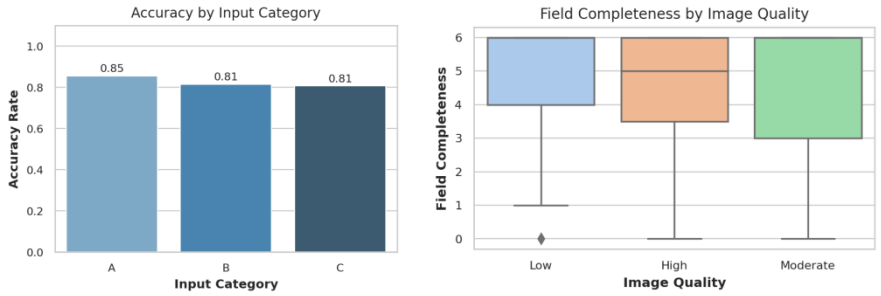

Fig. 6. System performance across input types and visual conditions. (Left) Accuracy rates across input categories: authentic Indonesian (A), non-Indonesian (B), and augmented images (C), showing highest classification accuracy for Category A. (Right) Boxplot of field completeness across image quality levels (High, Moderate, Low), demonstrating consistent reasoning depth regardless of image degradation.

4.3 Model Behavior, Errors, and Processing Efficiency

To evaluate classification reliability beyond raw accuracy, model outputs were categorized into four behavioral classes: Correct, Partially Correct, Hallucinated, and Invalid. These categories reflect the semantic fidelity and structural integrity of the generated responses. Out of 120 total samples, 89 outputs (74.17%) were fully correct, meaning the system produced accurate and taxonomically valid information across all or nearly all classification fields. An additional 18 samples (15.00%) were partially correct, typically omitting specific attributes or introducing minor ambiguities. Hallucinated responses—characterized by biologically implausible or fabricated content—accounted for 9 instances (7.50%), while only 4 outputs (3.33%) were labeled as invalid due to missing, irrelevant, or nonsensical responses. This low incidence of failure modes suggests that the integration of prompt engineering and retrieval-augmented reasoning effectively mitigates semantic drift, a common limitation in vision-language models. Inter-rater reliability for output categorization, computed using

Cohen's kappa, exceeded 0.85, indicating strong agreement between expert reviewers and validating the consistency of error type annotations.

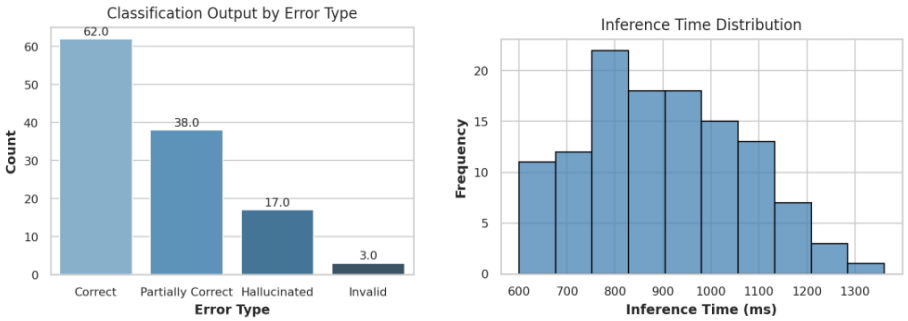

Fig. 7. System behavior and processing efficiency. (Left) Distribution of output classifications by error type, showing that the majority of responses were fully correct, with low rates of hallucinated and invalid outputs. (Right) Histogram of inference times across 120 sessions, demonstrating consistent response latency centered around 900 milliseconds, confirming suitability for near-real-time deployment.

Processing efficiency was evaluated by measuring inference latency from image submission to classification response. The system achieved a mean inference time of 903.06 milliseconds (SD = 164.23 ms), with minimum and maximum observed values of 600 ms and 1361 ms, respectively. To assess performance variability across inputs, inference times were analyzed using Welch's ANOVA, which confirmed no statistically significant differences attributable to input category or image quality ($p > 0.05$). This stability in processing time supports the system's suitability for deployment in real-time or near-real-time field applications, such as mobile ecological monitoring or rapid biodiversity assessments. The tight latency distribution, coupled with high semantic accuracy and minimal hallucination rates, demonstrates the model's operational readiness for scalable, sustainable deployment in low-resource environments—further aligning with the goals of green AI and ecological engineering.

5 Conclusion

This study confirms the effectiveness of a lightweight multimodal AI framework for Indonesian butterfly classification, demonstrating strong alignment with sustainable technology and green engineering goals. Addressing the first research question, the system achieved 85% classification accuracy in distinguishing Indonesian species from non-Indonesian and augmented inputs, supported by a high true positive rate and a well-balanced confusion matrix. For the second research question, the system consistently generated structured outputs across varied image quality levels, with a mean field completeness score of 4.58 out of 6. Statistical testing (ANOVA, $p = 0.903$) confirmed that differences in image quality did not significantly affect output depth, indicating the model's robustness under visual variability. Regarding the third question, error analysis showed 74.17% of outputs were fully correct, with low rates of hallucinations (7.5%) and invalid results (3.33%), validating the effectiveness of retrieval-augmented reasoning and prompt constraints. Inference latency averaged 903 ms, with minimal variation across sessions, confirming suitability for near-real-time ecological use. Overall, the system delivers accurate, semantically rich, and operationally efficient classification, making it a viable tool for biodiversity monitoring in resource-limited environments. The framework offers a replicable

model for sustainable AI deployment and supports future integration into citizen science, conservation, and educational platforms.

5.1 Limitation

While the proposed framework demonstrated high classification accuracy and semantic reliability, several limitations must be acknowledged. First, the experiment was conducted in a controlled laboratory environment using a curated dataset of 120 butterfly images, which may not fully capture the complexity of real-world ecological conditions, such as varied lighting, occlusions, or background noise. Second, although the system leveraged retrieval-augmented generation to constrain outputs, it remains dependent on the scope and accuracy of its reference data; misclassifications may still occur if unseen or morphologically similar species are presented. Third, manual labeling was used to evaluate output validity, which, despite strong inter-rater agreement, introduces potential subjectivity. Finally, the system was optimized for Indonesian butterfly species only, limiting its generalizability across broader biodiversity domains without retraining or expanding its taxonomic knowledge base. These limitations highlight areas for future enhancement, including larger-scale field validation, integration of real-time camera input, and adaptive learning to expand classification beyond the current species set.

5.2 Future Study

Future research will focus on extending the framework's applicability and enhancing its adaptability in real-world ecological settings. One direction involves scaling the system for broader taxonomic coverage, enabling classification of additional insect groups and other biodiversity indicators beyond butterflies. Incorporating real-time image acquisition through mobile camera integration and GPS-tagged metadata will further improve the contextual relevance of classifications. Adaptive learning mechanisms, driven by user feedback and incremental updates to the knowledge base, will be explored to reduce reliance on static reference data and improve long-term performance. Additionally, deploying the system in field conditions across various Indonesian habitats will allow for validation under diverse environmental constraints. Future studies may also investigate the integration of multilingual output support, enabling wider community engagement through citizen science platforms and educational tools. These extensions aim to strengthen the system's role as a scalable, low-resource, and sustainable solution for ecological monitoring in biodiversity-rich regions. To address current limitations, we acknowledge that this study was conducted using curated datasets under controlled conditions, and thus real-world variability—such as inconsistent lighting, partial occlusion of species, and region-specific annotation dialects—remains untested. In future work, we plan to conduct comprehensive field-based validation to assess the system's robustness under uncontrolled conditions and diverse ecological scenarios. Expanding the dataset with field-acquired images and annotations will not only improve generalizability but also offer insights into edge cases critical for long-term deployment.

References

1. Parikh, G., Rawtani, D. & Khatri, N. Insects as an indicator for environmental pollution. *Environmental Claims Journal* **33**, 161–181.
2. Chowdhury, S. Insects as bioindicator: A hidden gem for environmental monitoring. *Front Environ Sci* **11**,.
3. Wakhid, W., Agastya, I. M. I., Sumiati, A. & Nggani, R. U. R. Biodiversity and Species Composition of Butterflies in the Coban Glotak Waterfall, Malang, Indonesia. *Gontor Agrotech Science Journal* **9**, 151–160.

4. Zhang, Z. & Zhu, L. Intelligent Technology for the Monitoring and Protection of Insect Biodiversity. *Biodiversity Information Science and Standards* **3**,.
5. Surabhi, T., Sachin, B. & Advait, C. Deep Convolutional Neural Networks for Automated Butterfly Species Recognition and Classification. in *Proc. 2023 5th Int. Conf. Inventive Research in Computing Applications (ICIRCA* 778–783 (IEEE). doi:10.1109/ICIRCA57980.2023.10220696.
6. Sahraoui, M., Sklab, Y., Pignal, M., Lebbe, R. V & Guigue, V. Leveraging Multimodality for Biodiversity Data: Exploring joint representations of species descriptions and specimen images using CLIP. *Biodiversity Information Science and Standards* **7**,.
7. Yang, C.-H. *et al.* BioTrove: A large curated image dataset enabling AI for biodiversity. Preprint at https://doi.org/10.48550/arXiv.2406.17720.
8. Gong, Z. *et al.* BIOSCAN-CLIP: Bridging vision and genomics for biodiversity monitoring at scale.
9. Schlarmann, C., Singh, N. D., Croce, F. & Hein, M. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. Preprint at https://doi.org/10.48550/arXiv.2402.12336.
10. Dwivedi, D. N., Mahanty, G. & Dwivedi, V. N. Intelligent conservation: a comprehensive study on AI-enhanced environmental monitoring and preservation. in *The Convergence of Self-Sustaining Systems With AI and IoT, IGI Global* 215–226 doi:10.4018/979-8-3693-1702-0.ch011.
11. Khaleel, M., Murtaza, N., Mueen, Q. H., Ahmad, S. A. & Qadri, S. F. Use of AI in conservation and for understanding climate change. in *A Biologist's Guide to Artificial Intelligence* 201–240 (Academic Press). doi:10.1016/B978-0-443-24001-0.00013-0.
12. McClure, E. C. Artificial intelligence meets citizen science to supercharge ecological monitoring. *Patterns* **1**,.
13. Maharani, N., Kusrini, M. D. & Hamidy, A. Increasing herpetofauna data through citizen science in Indonesia. *IOP Conf. Ser.: Earth Environ. Sci* **950**,.
14. Rahmati, Y. Artificial Intelligence for Sustainable Urban Biodiversity: A Framework for Monitoring and Conservation. Preprint at https://doi.org/10.48550/arXiv.2501.14766.
15. Maharani, N. A novel citizen science-based wildlife monitoring and management tool for oil palm plantations. Preprint at https://doi.org/10.1101/2025.01.12.632638.