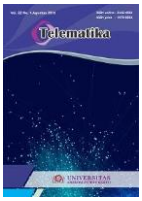

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

**Telematika**

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

# Comparative analysis of green snake identification using head structure and body patterns with vision transformer

**Eva Putriany[1], Dhani Ariatmanto[2]**

[1.2]Department Master of Informatics
[1.2]Universitas AMIKOM Yogyakarta, Indonesia
E-mail: evaputriany@students.amikom.ac.id[1], dhaniari@amikom.ac.id[2]

**ARTICLE INFO**

***History of the article:***
Received July 27, 2024
Revised February 17, 2025
Accepted February 28, 2025

**Keywords:**
Vision Transformer,
Image-based identification,
Snake classification,
Snake head structure

**Correspondence:**
E-mail:
evaputriany@students.amikom.ac.id

**ABSTRACT**

Snakebites remain a major global health concern, with over 4.5 million cases annually, primarily affecting rural populations in tropical regions. Accurate snake species identification is critical for proper treatment, yet challenges persist due to morphological similarities, particularly among visually similar green snake species. We test five Vision Transformer (ViT)-based models to see how well they can classify snakes based on pictures of their heads and bodies. The models are ViT-B16, DeiT, PoolFormer, Swin-T, and CaiT. Results indicate that head structure classification achieved higher accuracy than body pattern classification due to more distinct morphological features. CaiT outperformed other models, achieving 87% accuracy, particularly when trained on RGB images. These findings highlight the importance of model selection and dataset characteristics in improving snake species classification, especially for species with high visual similarity.

## INTRODUCTION

Snakebites have been deemed a Neglected Tropical Disease (NTD) by the World Health Organization (WHO) in the realm of global health (World Health Organization, 2023). These bites can result in severe health risks, often leading to life-threatening conditions (Afroz et al., 2023) With a call for a united global effort to mitigate fatalities and disabilities caused by snakebites health, WHO emphasizes the critical need for thorough attention and resolute action. Despite efforts to combat the issue, snakebite incidents persist every year with an estimated 4.5 to 5.4 million cases occurring annually. A vast majority of these cases, roughly 95%, take place in low- to middle-income countries and disproportionately affect impoverished populations, particularly those residing in rural areas of tropical nations. The frequency of bites is highest in South and Southeast Asia, with 2 million cases, followed by Sub-Saharan Africa with 420,000 cases and Latin America with 150,000 cases. Approximately 50%-55% of these bites are wet bites that inject venom. In 2016, snakebites resulted in 79,000 deaths and 400,000 permanent disabilities(World Health Organization, 2023).

Improper handling of snakebites can cause significant health risks, including harm, death, and lasting disabilities. Communities affected by limited access to snake identification and healthcare services and high treatment costs are exacerbated by this(Afroz et al., 2023). In a case in Chiang Mai, Thailand, a man was bitten by a venomous snake suspected to be a Malayan pit viper. After being given the wrong antivenom,

the patient's condition worsened, and Thai Red Cross hemato-polyvalent antivenom was used to correct the mistake (Tangtermpong et al., 2021).

Correctly identifying the snake species is crucial to determining the appropriate treatment to save lives and prevent complications or negative outcomes(Ralph et al., 2022).However, traditional methods of identifying snakes are difficult and have several limitations. Accurate snake identification requires experts who use morphological characteristics of the snake's body for identification. The current clinical standard for snakebite management is to have an expert identify the snake, usually a herpetologist (Bolon et al., 2020). Unfortunately, not all areas have easy access to herpetologists, and this process takes (Durso et al., 2021).

Automatically identifying snake images is vital, particularly for managing snakebites(Knudsen et al., 2021). Automated identification helps promote public safety by helping people avoid venomous snakes and enables healthcare providers to plan more effective treatments for snakebite victims. Additionally, this technology supports snake conservation efforts by providing accurate information about snake species distribution and behavior (Rajabizadeh & Rezghi, 2021).

Snake identification through automation involves the use of computer vision techniques, a subset of artificial intelligence that allows computers to analyze and interpret visual information (Matsuzaka & Yashiro, 2023). By mimicking human visual capabilities, this approach aims to provide computers with the ability to identify snakes accurately. One recent study employed the Vision Transformer (ViT) technique, achieving an F1 score of 92.2% at the species level with 772 classes. At the genus level, an F1 score of 96% was reached with 269 genera from 188 countries (Bolon et al., 2022), winning the SnakeSLEF 2021 challenge with an F1 score of 0.903 and an accuracy of 92.3%. Another study utilized the ResNet approach and achieved an accuracy of 91.6% as part of SnakeCLEF 2021(Chamidullin et al., 2021).

Despite significant advancements in automated snake identification using computer vision, existing research has primarily relied on dorsal pattern and coloration as the main visual features for classification (Rajabizadeh & Rezghi, 2021). However, this approach has limitations, particularly when identifying morphologically similar species with minimal color and pattern variation. Snakes with highly similar dorsal patterns and colors are frequently misclassified, suggesting that body pattern alone is not a sufficiently discriminative feature for distinguishing similar snake species. Meanwhile traditional snake identification can be performed based on morphological key identification features, such as the scale structure and shape of the head and tail (Sri Lanka Medical Association, 2022).While it is commonly assumed that triangular heads indicate venomous species and oval heads indicate non-venomous ones, these characteristics are not absolute identifiers, each snake species possesses unique morphological characteristics that differentiate them from others, making general assumptions unreliable(Wolfe et al., 2020).

Although prior studies have explored snake identification using deep learning models in similar snakes(de Solan et al., 2020), there has been no comparison of the effectiveness of body patterns versus head structure as key visual features for automated snake identification. This gap is particularly relevant for similar green snakes, which exhibit minimal variation in color and often lack distinctive dorsal patterns, making them difficult to differentiate based on body features alone. Both human observers and existing image-based algorithms struggle with these species due to their high morphological similarity. The lack of research explicitly affirming whether head structure provides superior classification accuracy compared to body pattern in such cases leaves a crucial unanswered question in the field of automated snake identification.

(a) (b) (c)

Figure 1. Green Snake (Javanese Gadung Snake)

There exist green snakes that share many similarities but come from different genera. These include the highly venomous Trimeresurus (a), the non-venomous Gonyosoma (b), and the low venomous Ahaetulla (c). Trimeresurus snakes are part of the medically significant Viperidae family, particularly in Southeast Asia (Ralph et al., 2022). Identifying Trimeresurus in Indonesia can be challenging because of their plain color and body features that resemble other local green snakes. Due to this difficulty, the general public often refers to all plain green snakes as "ular gadung" in the local community, as their color and shape resemble the tops of gadung plants (Rusli & Rini, 2020).

The purpose of this study is to addresses this challenge by conducting a comparative analysis of body pattern and head structure in snake classification using Transformers models. By evaluating the effectiveness of these distinct feature sets, this research aims to determine whether incorporating head structure into classification models enhances accuracy, particularly for visually similar green snake species. The accuracy of the results may vary due to a number of factors, such as dataset quality, algorithm selection, pre-processing, and optimizer usage, which can all impact the final outcome (Putriany & Ariatmanto, 2024). Further experimentation is required to determine the optimal combination for achieving maximum accuracy. This is essential for improving snakebite management, reducing misclassification risks, and advancing automated species identification methodologies.

## RESEARCH METHODS

This section explains about the materials of the research and the equipments used as the method. There is also explanations about the steps of solving the problem.

### 1. Datasets

Previous studies have highlighted several key factors that influence the accuracy of snake identification models. The dataset used in a model significantly impacts its results. Models trained on high-quality, well-annotated datasets tend to perform better, while those trained on datasets with poor-quality images or inconsistencies face challenges in distinguishing features accurately (Durso et al., 2021).. This highlights the importance of using a carefully curated dataset with consistent image quality, proper labeling, and balanced representation to enhance model performance.

To address these challenges, this study introduces a carefully curated dataset with controlled conditions. The dataset was compiled from a personal local snake collection, with expert-assisted labeling to ensure accuracy. It comprises three distinct snake classes at the genus level: Gonyosoma (a), Ahaetulla (b), and Trimeresurus (c). Each genus includes five individual snakes, with 100 images per individual for head structure and another 100 images for body patterns. This results in two datasets: one containing 300 images of head structures and another with 300 images of body patterns. The dataset used in this study can be accessed at www.kaggle.com/datasets/evaputriany/green-snake-datasets

(a) (b) (c)

Figure 2. Dataset 1 (*Head Structure*)

(a) (b) (c)

Figure 3. Dataset 2 (*Body Pattern*)

To ensure image consistency and minimize external distractions, all images were taken against a black background. This prevents potential biases caused by environmental elements and ensures that classification is based solely on morphological features. Additionally, the dataset is balanced, with an equal number of images per class, preventing skewed model performance due to class imbalance. Careful attention was given to lighting, angle variations, and image quality to provide a comprehensive dataset for evaluating the effectiveness of head structure versus body patterns in snake classification.

**2. Pre-Processing**

Prior to the classification stage (Ralph et al., 2022), the data preprocessing process serves as a critical initial step in image processing. Initially, image data normalization is performed to reduce differences in pixel intensity scale between images. This is achieved by scaling the pixel intensity to a smaller range between 0 and 1, thereby facilitating the learning process by the model.

The snake images dataset are converted from RGB to grayscale. This study employs grayscale conversion because the dataset consists entirely of green-colored snakes, meaning that color information is less discriminative in distinguishing between classes. In scenarios where color is not a strong distinguishing factor, grayscale-based descriptors that emphasize texture can enhance class separability, particularly for species with similar colors(Bhatta et al., 2023). By eliminating color, the model can focus more on morphological features such as shape, texture, and pattern, which are crucial for accurate classification. Afterward, the data is split into training, validation and testing data in 70:15:15 ratio. Then, data augmentation is applied to increase the variation of training data without adding new samples. Techniques such as rescaling, shearing, zooming, and horizontal flipping are applied to achieve this.

**3. Training and Classification**

The Transformer architecture is utilized by the Vision Transformer, which is among the image classification models (Pangestu et al., 2024). When it comes to identifying snake species, the ViT algorithm is widely utilized. The Biomedical Computer Science Group combined the ViT Large model and the EfficientNet-B4 model by averaging the softmax-scaled prediction probabilities of both models in the SnakeCLEF 2021 competition (Bloch & Friedrich, 2021). This model fusion led to

better results compared to individual models, with an accuracy of 82.88% (Bloch & Friedrich, 2021). Furthermore, the highest accuracy achieved was 99% at the genus level, also obtained in the SnakeCLEF 2021 competition (Bolon et al., 2022). Accuracy and identification effectiveness are also influenced by the appropriate model architecture selection in this evaluation. In this research, the Vision Transformer (ViT), which is a state-of-the-art neural network architecture with excellent image classification performance, has helped to achieve high levels of accuracy (Bolon et al., 2022).

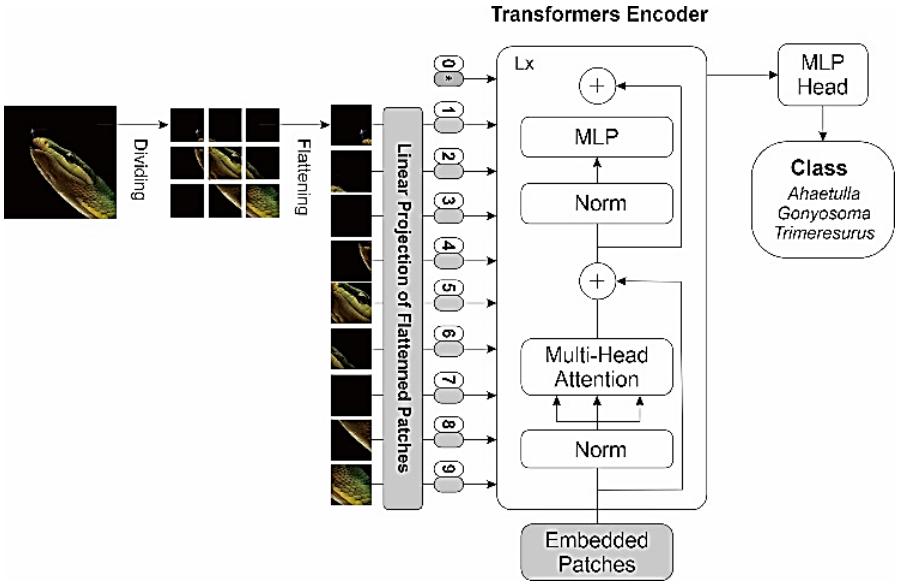


Figure 4. Vision transformer architecture

Essentially, ViT partitions the input image into smaller segments, or "patches" (Dosovitskiy et al., 2021). The number of patches is determined by dividing the image into as many segments as

$$N = (H/P) \times (W/P) \quad (1)$$

Where $P$ is the patch size, $H$ is the height, $W$ is the width, and $C$ represents the number of color channels (RGB). Each patch is then flattened, converting the $C$-D matrix into a 1-D vector. The flattened patches are transformed into embeddings, and a learnable token is added. In order to retain positional information from each patch, a positional embedding matrix **E** (known as positional embedding) is randomly generated with a size of $((N + 1) \times D)$ and added to the combined matrix of learnable class embeddings and patch embeddings. This enables the model to grasp the relative positions of each patch within the image (Dosovitskiy et al., 2021). In the process of training the neural network model, a learning rate of 0.001 was adopted alongside a batch size of 256. The model leveraged the softmax activation function in conjunction with the Adam optimization algorithm, across a total training duration of 30 epochs.

## 4. Evaluation

To systematically evaluate model performance in fine-grained snake classification, this study utilizes multiple performance metrics, including accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correctly classified images, providing a general assessment of model effectiveness. Accuracy is the fundamental metric that measures the proportion of correctly classified images relative to the total number of samples.

Precision and recall are particularly important for species classification, as they quantify the ability to correctly identify true positive cases while minimizing false positives and false negatives.

The F1-score, which balances precision and recall, ensures a more reliable evaluation, especially when handling class imbalances in the dataset.

## RESULTS AND DISCUSSION

The performance of five vision transformer-based models was evaluated for snake classification without pretraining. Those models are ViT-B16, DeiT, PoolFormer, Swin-T, and CaiT. that selected based on their architectural characteristics and potential for fine-grained image classification. ViT-B16, a standard Vision Transformer model with a patch size of 16, is known for its capacity to capture long-range dependencies in images. DeiT, a data-efficient variant of ViT, is optimized for improved training efficiency, reducing reliance on extensive datasets(Touvron et al., 2020). PoolFormer introduces a novel approach by substituting self-attention mechanisms with pooling operations, thereby reducing computational complexity(Weihao Yu et al., 2022). Swin-T employs a hierarchical structure with a shifted window mechanism, enhancing spatial locality and scalability(Dümen et al., 2024). Lastly, CaiT incorporates a class-attention mechanism to refine feature extraction, rendering it particularly effective for fine-grained classification tasks(Touvron et al., 2021). To systematically assess their performance, four key evaluation metrics were employed: accuracy, precision, recall, and F1-score. This comprehensive evaluation aims to determine the efficacy of each model across different datasets and classification tasks.

A critical factor in model performance is the nature of input data, particularly the impact of color information on classification accuracy. The dataset was pre-processed in two different schemes, grayscale and RGB. The grayscale dataset was generated by converting the original images into single-channel representations, removing color information while retaining structural and textural features. Meanwhile, the RGB dataset remained in its original format, preserving full-color details that could contribute to species differentiation. By training models on both dataset variants, the study evaluates the extent to which color features influence classification performance.

The performance of the five transformer-based models varied significantly across classification tasks, particularly between grayscale and RGB image processing. A comparative analysis of head structure and body pattern classification highlights the relative effectiveness of each model, particularly when trained on grayscale and RGB images. The classification performance varied significantly across the five models, with noticeable differences between grayscale and RGB image processing. Table 1 presents the evaluation results for head structure classification, whereas Table 2 provides insights into body pattern classification.

Table 1. Evaluation of Dataset Head Structure

| *Model* | *Grayscale* | | | | *RGB* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *F1 Score* | *Precision* | *Recall* | *Accuracy* | *F1 Score* | *Precision* | *Recall* |
| ViT B16 | 0.67 | 0.63 | 0.68 | 0.67 | 0.76 | 0.76 | 0.78 | 0.76 |
| DeiT | 0.69 | 0.65 | 0.70 | 0.69 | 0.78 | 0.78 | 0.81 | 0.78 |
| Poolformer | 0.33 | 0.17 | 0.11 | 0.33 | 0.36 | 0.21 | 0.45 | 0.36 |
| Swin T | 0.58 | 0.51 | 0.49 | 0.58 | 0.67 | 0.63 | 0.71 | 0.67 |
| CaiT | 0.67 | 0.64 | 0.66 | 0.67 | 0.87 | 0.87 | 0.88 | 0.87 |

Table 1 shows that CaiT demonstrated the highest performance, achieving an accuracy of 87% and an F1-score of 0.87 when using RGB images. This result underscores the advantage of CaiT's class-attention mechanism in distinguishing head structures, setting it apart from other models. DeiT and ViT-

B16 also exhibited competitive performance, with accuracy values of 78% and 76%, respectively, suggesting that transformer-based models with efficient training strategies can achieve reliable classification. In contrast, PoolFormer performed poorly, with a grayscale accuracy of only 33% and an RGB accuracy of 36%, highlighting the limitations of its pooling-based feature extraction mechanism in capturing fine-grained details.

Table 2. Evaluation of Dataset Body Pattern

| *Model* | *Grayscale* | | | | *RGB* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Accuracy* | *F1 Score* | *Precision* | *Recall* | *Accuracy* | *F1 Score* | *Precision* | *Recall* |
| ViT B16 | 0.43 | 0.35 | 0.29 | 0.43 | 0.70 | 0.70 | 0.70 | 0.74 |
| DeiT | 0.43 | 0.36 | 0.31 | 0.48 | 0.63 | 0.61 | 0.66 | 0.63 |
| Poolformer | 0.27 | 0.14 | 0.09 | 0.33 | 0.23 | 0.13 | 0.08 | 0.33 |
| Swin T | 0.30 | 0.15 | 0.10 | 0.33 | 0.23 | 0.13 | 0.08 | 0.33 |
| CaiT | 0.37 | 0.34 | 0.39 | 0.38 | 0.77 | 0.75 | 0.75 | 0.75 |

A similar trend was observed in body pattern classification (Table 2), the models exhibited more pronounced differences in performance. CaiT again emerged as the best-performing model with an accuracy of 77% in the RGB setting, demonstrating its robustness in capturing complex texture patterns. However, DeiT showed a slight decline in performance compared to head structure classification, with an accuracy of 63%, potentially due to its reliance on distilled knowledge, which may be less effective for complex body patterns. PoolFormer and Swin-T performed the worst in this task, with accuracy values below 30% in grayscale and below 25% in RGB, reinforcing their limitations in processing fine-grained texture information.

ViT-B16 achieved 70% accuracy, indicating its ability to generalize well to body pattern recognition. While ViT utilizes standard self-attention across all tokens, CaiT introduces an additional class-attention mechanism that refines feature extraction in later stages of processing. This architectural distinction allows CaiT to focus more effectively on discriminative features, particularly for fine-grained classification tasks. However, despite this enhancement, the results indicate that CaiT experiences a notable performance drop when classifying body patterns compared to head structures. While the disrespancy excels at refining high-level morphological features such as head structure, it may not be as effective in capturing fine-grained textural variations required for body pattern recognition. In contrast, ViT, which applies global attention uniformly across the entire image, performed more consistently across both tasks, suggesting that its feature extraction process does not exhibit a strong bias toward structural over textural details. This discrepancy highlights a potential limitation in CaiT's attention mechanism when applied to species with subtle pattern variations, warranting further investigation into hybrid approaches that integrate both self-attention and convolution-based feature extraction to enhance texture classification.

The superior performance of CaiT can be attributed to its underlying architectural advantages. CaiT demonstrated superior performance across all classification tasks due to its advanced architectural design, particularly its class-attention mechanism and deep feature extraction capabilities. Unlike other transformer-based models, which apply self-attention across all image tokens equally, CaiT introduces specialized class-attention layers that enhance feature refinement. This allows the model to focus more effectively on key morphological details, such as head shape and body patterns, making it particularly well-suited for fine-grained classification. Compared to PoolFormer, which struggles with feature retention due

to its pooling-based structure, CaiT maintains high-resolution spatial information, leading to significantly better accuracy.

Furthermore, the effective utilization of color information plays a crucial role in CaiT's success. Another key factor contributing to CaiT's success is its ability to leverage color information more effectively than other models. The performance gap between grayscale and RGB images was more pronounced in CaiT, suggesting that it utilizes color-based features to distinguish species more accurately. This advantage is especially important in cases of similar snakes. While models like ViT-B16 and DeiT also benefited from color information, their reliance on standard self-attention mechanisms made them slightly less effective in capturing complex texture and color variations.

To further analyze CaiT's performance, class-wise evaluations were conducted to assess its effectiveness across different snake species. Table 3 presents the results. The model's higher accuracy in head structure classification suggests that it was more effective at learning morphological features than body patterns, likely due to the inherent complexity of texture-based classification. The variation in class-wise performance further highlights CaiT's strengths and areas for improvement.

Table 3. Evaluation of CaiT

| *Dataset* | *Accuracy* | *Class* | *Evaluation of CaiT* | | |
|---|---|---|---|---|---|
| | | | *F1 Score* | *Precision* | *Recall* |
| Body Pattern | 0.77 | Ahaetulla | 0.59 | 0.56 | 0.62 |
| | | Gonyosoma | 0.96 | 1.00 | 0.92 |
| | | Trimeresurus | 0.70 | 0.70 | 0.70 |
| Head Structure | 0.87 | Ahaetulla | 0.85 | 0.78 | 0.93 |
| | | Gonyosoma | 0.83 | 0.86 | 0.80 |
| | | Trimeresurus | 0.93 | 1.00 | 0.87 |

CaiT demonstrated strong performance across both head structure and body pattern classification tasks, achieving an accuracy of 87% and 77%, respectively. The model's higher accuracy in head structure classification suggests that it was more effective at learning morphological features than body patterns, likely due to the inherent complexity of texture-based classification. This discrepancy in accuracy highlights the model's strengths in structural recognition while revealing challenges in texture differentiation. The variation in class-wise performance further highlights CaiT's strengths and areas for improvement.

For head structure classification, *Trimeresurus* exhibited the highest F1-score (0.93), with a perfect precision score of 1.00 but a slightly lower recall of 0.87. This indicates that the model made no false positive predictions for *Trimeresurus*, but some instances were misclassified as other species. *Ahaetulla* had the highest recall (0.93), meaning the model correctly identified most samples, but a lower precision of 0.78 suggests that some misclassifications occurred. Meanwhile, *Gonyosoma* had a balanced performance with an F1-score of 0.83, precision of 0.86, and recall of 0.80, indicating that the model was generally reliable but occasionally confused this class with others. These findings suggest that while the model is adept at recognizing certain species, there are still inconsistencies that need to be addressed, particularly in handling species with overlapping features.

In body pattern classification, the performance was less consistent, reflecting the increased difficulty of distinguishing species based on texture alone. *Gonyosoma* was the easiest class to identify, with a perfect precision score of 1.00 and an F1-score of 0.96, signifying that the model rarely misclassified this species.

This result suggests that *Gonyosoma* possesses distinct texture features that CaiT could learn effectively. *Trimeresurus*, on the other hand, had a moderate F1-score of 0.70, with balanced precision and recall, indicating that the model struggled to capture fine-grained variations in its body pattern. The lowest performance was observed in *Ahaetulla*, with an F1-score of 0.59, precision of 0.56, and recall of 0.62, showing that the model frequently misclassified this class, possibly due to similarities with other species or limitations in feature extraction for texture-based discrimination. The contrast in performance between different species further supports the notion that structural features are easier to learn than texture, reinforcing the challenges of fine-grained classification in natural settings.

The results indicate that structural features, particularly those related to the head, are generally easier for transformer-based models to learn compared to texture-based body patterns. The more stable and distinct geometric characteristics of head structures play a crucial role in identification. Head structures offer more consistent and distinct geometric features, such as snout shape, eye placement, and scale arrangement, which remain stable across different individuals and lighting conditions. These structural features tend to stay the same across different images, which makes them more reliable and a solid foundation for classification. They have clear edges and well-defined geometric features that Transformers' self-attention mechanisms can effectively pick out and emphasize.

Conversely, body pattern classification poses several challenges that may contribute to model misclassification, especially when using datasets with visually similar species. body pattern classification presents additional challenges due to the greater variability in texture and color patterns among individuals of the same species. The high degree of interspecies similarity in body coloration among green snake species may have further complicated the classification task. Additionally, occlusions caused by the snake's natural posture, such as coiling or partial concealment by vegetation, may obscure crucial body markings. These issues are exacerbated in species with near-identical body patterns, where classification must rely on subtle differences that are not always consistently captured across samples. Moreover, body patterns may exhibit strong within-class diversity due to age, sex, or regional variations. Some snake species display polymorphic coloration, where individuals of the same species can have slightly different hues or markings, further complicating the classification process. The use of multiple datasets with similar-looking green snakes increases the likelihood of misclassification, as the model struggles to differentiate species based on minor textural differences alone. In contrast, head structures show less within-species variation, allowing models to capture more stable representations.

While CaiT's self-attention mechanism should be capable of capturing such complex features, the results indicate that its performance is still influenced by the relative stability and distinctiveness of the features it processes. Transformers rely entirely on token-based representations, which may struggle to capture the subtle, non-uniform patterns found in snake body markings. The lack of explicit spatial anchoring in pure self-attention models may limit their ability to distinguish fine-grained, high-frequency textures (Dosovitskiy et al., 2021). Another potential factor is the importance of global versus local features in classification. Head structures contain salient, species-specific morphological traits that are recognizable even at lower resolutions. This aligns well with the way Transformers process images, as they can attend to global structures while filtering out irrelevant background noise. However, body pattern classification requires fine-scale texture differentiation, which involves learning minute pixel-level differences rather than broader geometric patterns. Since self-attention mechanisms are not inherently designed to focus on fine-grained textures, models like CaiT may require additional architectural modifications—such as

convolutional embeddings or multi-scale feature processing—to improve their ability to distinguish subtle texture variations (Xie et al., 2022).

The discrepancy between head structure and body pattern classification performance also suggests that CaiT leveraged shape-based features more effectively than color and texture. This aligns with the model's attention-based architecture, which prioritizes structural details over fine-grained textures. This architectural bias towards structure over texture may explain why the model performed well in head classification but struggled with body patterns. Additionally, the lower recall in body pattern classification for *Ahaetulla* and *Trimeresurus* indicates that these species share visual similarities with others, making classification more challenging.

The observed differences in model performance can be attributed to multiple factors, including the effectiveness of attention mechanisms, architectural variations, and the reliance on color information in classification. Transformer-based architectures, particularly those employing global attention mechanisms, generally demonstrated superior performance compared to models relying on pooling-based feature extraction. These insights suggest that the choice of architecture plays a crucial role in determining classification accuracy, particularly for tasks requiring high-resolution feature extraction. Additionally, the disparity in performance between grayscale and RGB datasets highlights the critical role of color information in species identification, particularly in distinguishing visually similar snake species.

Attention mechanisms also played a key role in determining the effectiveness of different models. CaiT, which incorporates a class-attention mechanism, consistently outperformed other models, demonstrating higher accuracy and F1-scores across both head structure and body pattern classification. This suggests that class-attention helps enhance feature extraction by focusing on the most relevant regions of the image. Similarly, DeiT and ViT-B16, both of which employ global attention, showed competitive performance, indicating that transformers with well-optimized attention strategies are well-suited for fine-grained image classification tasks. In contrast, Swin-T, which uses a hierarchical attention mechanism, exhibited lower performance, particularly in body pattern classification, suggesting that its window-based attention approach may be less effective for capturing detailed texture patterns necessary for distinguishing snake species. These findings emphasize the importance of optimizing attention strategies to enhance model performance, particularly for species with subtle morphological differences.

Architectural differences also contributed significantly to the varying performance levels. Models with global attention mechanisms, such as CaiT, DeiT, and ViT-B16, demonstrated superior performance compared to hierarchical transformers like Swin-T and pooling-based models like PoolFormer. PoolFormer, in particular, struggled to capture fine-grained details, as indicated by its consistently low accuracy in both grayscale and RGB settings. This can be attributed to its reliance on pooling operations rather than attention mechanisms, which limits its ability to retain crucial morphological and textural details necessary for accurate classification. The poor performance of PoolFormer highlights the importance of feature-rich representations in snake classification, where subtle differences in scale patterns and head structures play a significant role in species identification. This limitation underscores the necessity of using feature-rich architectures in tasks that require precise differentiation of species with similar visual characteristics.

The reliance on color information was another major factor influencing model performance. Across all models, RGB images consistently led to higher classification accuracy compared to grayscale images, reinforcing the importance of color-based features in distinguishing snake species. The impact of color

information was particularly evident in models such as CaiT and DeiT, where RGB-based training resulted in at least a 10% improvement in accuracy compared to grayscale. This suggests that many species-specific features, such as scale coloration and pattern distribution, are more effectively captured when color information is available. The significant performance gap between grayscale and RGB datasets also highlights the challenges posed by similar snake in batesian mimicry, where non-venomous species closely resemble venomous ones in terms of color patterns. Without color information, models struggled to differentiate between these species, leading to increased misclassification rates.

PoolFormer exhibited the lowest performance across both head structure and body pattern classification tasks, primarily due to its pooling-based mechanism, which lacks the dynamic feature extraction capabilities of attention-based models. Unlike transformer architectures that refine features through self-attention, PoolFormer relies on a non-parametric pooling operation to aggregate spatial information. While this design reduces computational complexity, it significantly limits the model's ability to capture fine-grained details essential for distinguishing snake species. As a result, PoolFormer struggled to differentiate subtle morphological variations, leading to poor accuracy and F1-scores, particularly in body pattern classification.

When compared to attention-based models like CaiT, ViT-B16, and DeiT, PoolFormer's limitations become even more evident. Transformers utilize self-attention to dynamically weigh important regions within an image, enabling better feature refinement and contextual understanding. In contrast, PoolFormer's fixed pooling operations discard crucial high-frequency details, making it less effective at recognizing intricate patterns such as texture and scale arrangements. This was particularly evident in the body pattern classification task, where PoolFormer's accuracy fell below 30% in grayscale and 25% in RGB, significantly lower than its transformer-based counterparts. This comparison further highlights the importance of selecting models with adaptive feature extraction mechanisms for fine-grained classification tasks.

While CaiT demonstrated strong performance, certain limitations may affect its generalization and practical application. One notable constraint is its reliance on structural features, which may contribute to inconsistencies in body pattern classification. The lower accuracy in distinguishing species based on texture suggests that CaiT's attention mechanisms prioritize high-level structural details while potentially overlooking fine-grained pattern variations. Future research could explore hybrid architectures that integrate both attention-based and convolutional mechanisms to enhance texture recognition while maintaining CaiT's capacity for structural feature extraction. Expanding the dataset with more diverse and balanced samples might reduce bias and improve generalization across species. Additionally, incorporating domain adaptation techniques, such as self-supervised learning or contrastive learning, could enhance the model's ability to recognize snakes under varying environmental conditions. The potential integration of CaiT into real-time applications, such as mobile-based snake identification tools, may necessitate further optimization to balance computational efficiency and accuracy. These advancements could contribute to improving the model's applicability in ecological research, conservation efforts, and biodiversity monitoring.

## CONCLUSIONS AND RECOMMENDATIONS

Transformer-based models are evaluated for fine-grained snake classification using head structure and body pattern datasets. CaiT achieved the highest accuracy, reaching 0.87 particularly with RGB images,

due to its class-attention mechanism, which effectively captured fine-grained morphological details. ViT-B16 and DeiT performed competitively, benefiting from global attention but were slightly less effective in capturing texture and color variations. Swin-T and PoolFormer exhibited lower accuracy, with PoolFormer having the lowest at 0.36, as its pooling-based approach limited morphological detail retention. The results highlight the importance of color in classification, as RGB images consistently outperformed grayscale, emphasizing challenges posed by similar snake. Despite its strengths, CaiT showed limitations in body pattern classification, suggesting the need for refinements in texture-based feature extraction. Future research should explore hybrid architectures combining attention with convolutional layers, dataset expansion for better generalization, domain adaptation techniques, enhanced texture-based feature extraction, real-world deployment strategies, and comparisons with CNNs and hybrid models to improve fine-grained snake classification.

## ACKNOWLEDGEMENT

The authors sincerely thank Thurgana Noviantoro Wardiyan of Jogja Exo Dungeon, a dedicated snake expert and experienced breeder of endemic snakes, for his invaluable contributions to this study. His expertise in snake identification and handling was instrumental in ensuring the accuracy of the dataset. We are deeply grateful for his efforts in capturing high-quality photographs and meticulously labeling the dataset, which greatly enhanced the reliability of our research. His unwavering support and commitment have been essential in advancing our study on snake classification using computer vision.

## REFERENCES

Afroz, A., Siddiquea, B. N., Shetty, A. N., Jackson, T. N. W., & Watt, A. D. (2023). Assessing knowledge and awareness regarding snakebite and management of snakebite envenoming in healthcare workers and the general population: A systematic review and meta-analysis. *PLOS Neglected Tropical Diseases*, *17*(2), e0011048. https://doi.org/10.1371/journal.pntd.0011048

Alexey Dosovitskiy, Lucas Beyer, Dirk Weissenborn, Alexander Kolesnikov, Xiaohua Zhai, & Thomas Unterthiner. (n.d.). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

Bhatta, A., Mery, D., Wu, H., Annan, J., King, M. C., & Bowyer, K. W. (2023). *What's color got to do with it? Face recognition in grayscale*. http://arxiv.org/abs/2309.05180

Bloch, L., & Friedrich, C. M. (2021). *EfficientNets and Vision Transformers for Snake Species Identification Using Image and Location Information*. https://www.fh-dortmund.de/personen/Christoph-Friedrich/index.php

Bolon, I., Durso, A. M., Botero Mesa, S., Ray, N., Alcoba, G., Chappuis, F., & Ruiz de Castañeda, R. (2020). Identifying the snake: First scoping review on practices of communities and healthcare providers confronted with snakebite across the world. *PLOS ONE*, *15*(3), e0229989. https://doi.org/10.1371/journal.pone.0229989

Bolon, I., Picek, L., Durso, A. M., Alcoba, G., Chappuis, F., & Ruiz de Castañeda, R. (2022). An artificial intelligence model to identify snakes from across the world: Opportunities and challenges for global health and herpetology. *PLOS Neglected Tropical Diseases*, *16*(8), e0010647. https://doi.org/10.1371/journal.pntd.0010647

Chamidullin, R., Šulc, M., Matas, J., & Picek, L. (2021). *A Deep Learning Method for Visual Recognition of Snake Species*. http://ceur-ws.org

de Solan, T., Renoult, J. P., Geniez, P., David, P., & Crochet, P.-A. (2020). Looking for Mimicry in a Snake Assemblage Using Deep Learning. *The American Naturalist*, *196*(1), 74–86. https://doi.org/10.1086/708763

Dümen, S., Kavalcı Yılmaz, E., Adem, K., & Avaroglu, E. (2024). Performance of vision transformer and swin transformer models for lemon quality classification in fruit juice factories. *European Food Research and Technology*, *250*(9), 2291–2302. https://doi.org/10.1007/s00217-024-04537-5

Durso, A. M., Moorthy, G. K., Mohanty, S. P., Bolon, I., Salathé, M., & Ruiz de Castañeda, R. (2021). Supervised Learning Computer Vision Benchmark for Snake Species Identification From Photographs: Implications for Herpetology and Global Health. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.582110

Eva Putriany, & Dhani Ariatmanto. (2024). Literatur Reviu Sistematis: Identifikasi Jenis Ular Berbasis Computer Vision. *JNANALOKA*, 43. https://doi.org/10.36802/jnanaloka.2024.v5-no01-43-50

Knudsen, C., Jürgensen, J. A., Føns, S., Haack, A. M., Friis, R. U. W., Dam, S. H., Bush, S. P., White, J., & Laustsen, A. H. (2021). Snakebite Envenoming Diagnosis and Diagnostics. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.661457

Matsuzaka, Y., & Yashiro, R. (2023). AI-Based Computer Vision Techniques and Expert Systems. *AI*, *4*(1), 289–302. https://doi.org/10.3390/ai4010013

Pangestu, A., Purnama, B., & Risnandar, R. (2024). Vision Transformer untuk Klasifikasi Kematangan Pisang. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(1), 75–84. https://doi.org/10.25126/jtiik.20241117389

Rajabizadeh, M., & Rezghi, M. (2021). A comparative study on image-based snake identification using machine learning. *Scientific Reports*, *11*(1), 19142. https://doi.org/10.1038/s41598-021-96031-1

Ralph, R., Faiz, M. A., Sharma, S. K., Ribeiro, I., & Chappuis, F. (2022). Managing snakebite. *BMJ*, e057926. https://doi.org/10.1136/bmj-2020-057926

Rusli, N., & Rini, C. P. (2020). *Ular di Sekitar Kita: Pulau Jawa*. Indonesia Herpetofauna Foundation.

Sri Lanka Medical Association. (2022, April 5). *Identification of Snakes*.

Tangtermpong, A., Pinyopornpanish, K., Vasaruchapong, T., Chenthanakij, B., & Pinyopornpanish, K. (2021). The Treatment of Unidentified Hematotoxic Snake Envenomation and the Clinical Manifestations of a *Protobothrops kelomohy* Bite. *Wilderness & Environmental Medicine*, *32*(1), 83–87. https://doi.org/10.1016/j.wem.2020.11.001

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). *Training data-efficient image transformers & distillation through attention*. http://arxiv.org/abs/2012.12877

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). *Going deeper with Image Transformers*. http://arxiv.org/abs/2103.17239

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, & Shuicheng Yan. (2022). MetaFormer Is Actually What You Need for Vision. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*.

Wolfe, A. K., Fleming, P. A., & Bateman, P. W. (2020). What snake is that? Common Australian snake species are frequently misidentified or unidentified. *Human Dimensions of Wildlife*, *25*(6), 517–530. https://doi.org/10.1080/10871209.2020.1769778

World Health Organization. (2023, September 12). *Snakebite Envenoming*. Https://Www.Who.Int/News-Room/Fact-Sheets/Detail/Snakebite-Envenoming.

Xie, J., Zhang, J., Sun, J., Ma, Z., Qin, L., Li, G., Zhou, H., & Zhan, Y. (2022). A Transformer-Based Approach Combining Deep Learning Network and Spatial-Temporal Information for Raw EEG Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 2126–2136. https://doi.org/10.1109/TNSRE.2022.3194600