Enhancing Fine-Grained Medical Image Classification Accuracy Via Test-Time Object Detection

Po-Chun Chuang
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
Kaohsiung, Taiwan
Department of Emergency Medicine
Kaohsiung Chang Gung Memorial Hospital
Kaohsiung, Taiwan
zhungboqun@gmail.com

Ye-In Chang
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
Kaohsiung, Taiwan
changyi@cse.nsysu.edu.tw

Abstract

Deep learning (DL) models often struggle to maintain accuracy when transitioning from validation datasets to real-world applications due to variations in image quality and object diversity. This study explores snake species identification as a case study, utilizing the Swin Transformer V2 model to address these challenges. The model, fine-tuned through transfer learning, achieved a validation accuracy of 96.29%. However, its accuracy declined to 83.29% when tested on user-submitted images collected via the LINE chatbot and social media platforms. To mitigate this issue, a Test-Time Object Detection and Cropping method was introduced, using the OWLv2 zero-shot object detection model to preprocess images by detecting and cropping snake regions. This approach improved the external test set accuracy to 89.75%, closely aligning with the human-annotated baseline accuracy of 90.25%. These findings underscore the significance of preprocessing techniques in enhancing the reliability and practical applicability of DL models. Future research should focus on expanding datasets and addressing challenges associated with underrepresented species to further improve performance.

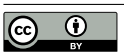
CCS Concepts

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision tasks; Biometrics.

Keywords

deep learning, snake species identification, object detection, preprocessing, real-world application

ACM Reference Format:
Po-Chun Chuang and Ye-In Chang. 2025. Enhancing Fine-Grained Medical Image Classification Accuracy Via Test-Time Object Detection. In *2025 9th International Conference on Medical and Health Informatics (ICMHI 2025), May 16–18, 2025, Kyoto, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3761712.3761746



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMHI 2025, Kyoto, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1514-3/2025/05
https://doi.org/10.1145/3761712.3761746

1 Introduction

Deep learning (DL) models often exhibit performance discrepancies between internal validation and external test datasets due to variations in population demographics and environmental settings [1]. This underscores the need for rigorous external testing to ensure model generalizability and reliability in real-world applications. For instance, models trained on a specific dataset may perform suboptimally when applied to different contexts. To address this challenge, this study explores the issue of snakebites as a use case, focusing on developing a DL model for snake species identification to minimize performance gaps across datasets.

Snakebites are a major global health concern, causing approximately 100,000 fatalities annually and affecting millions worldwide [2]. Accurate snake species identification is crucial for effective antivenom administration, which remains the cornerstone of snakebite management [3]. However, identifying snake species presents significant challenges for healthcare providers, with global data indicating that only 53% of snakebite cases are correctly identified [4]. Recent advancements in DL have shown promise, with some studies achieving accuracy rates of approximately 94% in snake species classification at the country level [5, 6].

This study aimed to develop a DL model designed for both public and clinical use, enabling accurate snake species identification from user-submitted images. By leveraging real-world images, we assessed the model's performance and proposed strategies to enhance its robustness and reliability, bridging the gap between experimental conditions and practical applications.

2 METHODS

2.1 Ethics Approval

This study was conducted in compliance with the ethical standards and approved by the Chang Gung Medical Foundation Institutional Review Board (IRB approval numbers: 202201210B0 and 202301246B0A3).

2.2 Classification of Snake Species in Taiwan

To ensure clinical relevance, snake species native to Taiwan were categorized into 11 classes. Ten of these classes represent venomous species with documented envenomation cases, while the "Others" category includes nonvenomous species or those without recorded envenomation incidents. The defined classes are as follows:

- Trimeresurus stejnegeri

Table 1: Image Collection and Sampling in Training, Validation, and External Test Sets

	Training and Validation Set				External test Set
	Collected	Sampled	Training	Validation	
Trimeresurus stejnegeri	3,166	1,000	793	207	230
Protobothrops mucrosquamatus	1,952	1,000	787	213	302
Naja atra	1,950	1,000	796	204	97
Bungarus multicinctus	1,160	1,000	810	190	131
Deinagkistrodon acutus	3,588	1,000	797	203	96
Daboia siamensis	1,891	1,000	804	196	43
Trimeresurus gracilis	960	960	771	189	30
Ovophis makazayazaya	1,394	1,000	803	197	80
Sinomicrurus spp.	677	500	399	101	31
Rhabdophis formosanus	418	418	350	68	23
Others	13,417	3,122	2,490	632	1,337
Total images	30,573	12,000	9,600	2,400	2,400

- Protobothrops mucrosquamatus
- Naja atra
- Bungarus multicinctus
- Deinagkistrodon acutus
- Daboia siamensis
- Trimeresurus gracilis
- Ovophis makazayazaya
- Sinomicrurus spp.
- Rhabdophis formosanus
- Others

2.3 Data Sources and Labeling

Training and validation datasets were compiled from publicly available platforms, including Flickr, iNaturalist, the Taiwan Reptile Report Program, and the Taiwan Roadkill Observation Network. External test images were sourced from a LINE chatbot and Facebook groups between November 2023 and April 2024 (Table 1). To ensure data quality, expert herpetologists from the National Pingtung University of Science and Technology Herpetology Laboratory meticulously labeled the images, excluding any that were unidentifiable.

2.4 Preprocessing Workflow

Data preprocessing varied by dataset type:

- Training Set: Extensive augmentation techniques, including random flipping, perspective transformations, rotation, and center cropping, were applied to improve model generalization.
- Validation Set: Minimal transformations were used to maintain consistency with real-world conditions.
- External Test Set: The Test-Time Object Detection and Cropping method was implemented to handle real-world image variability.

All images were resized to 224×224 pixels and normalized to the RGB channels for model input consistency.

2.5 Model Architecture and Training

2.5.1 Transfer Learning with Swin Transformer. Swin Transformer v2, specifically the *swinv2-base-patch4-window12-192-22k* variant pretrained on ImageNet-21k, served as the backbone for this study [7, 8]. This model leverages hierarchical vision prior to performing tasks such as classification and detection.

2.5.2 Deployment and User Interface: Line Chatbot. The trained model was deployed through a LINE chatbot to enhance real-world accessibility. The Django framework managed image submission and storage. The chatbot processed submitted images using the Swin Transformer model to predict snake species and delivered the results directly to users.

2.5.3 Test-Time Object Detection and Cropping. To address real-world challenges, images sourced from external platforms often exhibited reduced quality and poorly framed snake regions, leading to decreased classification accuracy. To mitigate this issue, this study introduces a Test-Time Object Detection and Cropping technique.

Among the various zero-shot object detection models tested, Google's *owlv2-large-patch14-finetuned variant* demonstrated superior performance (Table 2). The model identified snake regions using the query "snake," and the detected area was cropped for classification. If multiple regions were detected, the largest region was selected for analysis. When no region was identified, the original image was processed without modifications.

This method significantly improved the classification accuracy on external test datasets, bridging the gap between experimental and real-world performances.

2.5.4 Cropping Methods for Test-Time Image Processing. To improve classification performance and account for variability in real-world snake images, three distinct cropping methods were implemented during the preprocessing stage for the external test dataset:

- Top-Left Crop: In this approach, the bounding box coordinates obtained from the object detection model were utilized to crop a 224×224 pixel region starting from the top-left corner of the detected area. This straightforward method

Table 2: Accuracies for Different Cropping Methods in Validation and External Test Sets.

	Validation Set (n=2400)		
	Left top crop	Box-Center Crop	No Adjustment
no_object_detection	95.58%		
grounding_dino_tiny	81.54%	95.92%	92.58%
grounding_dino_base	83.75%	96.25%	92.67%
omdet_turbo_swin_tiny_hf	85.04%	96.04%	93.50%
owlvit_base_patch16	92.46%	95.92%	95.29%
owlvit_base_patch32	86.67%	95.88%	93.54%
owlvit_large_patch14	88.71%	95.88%	93.58%
owlv2_base_patch16	87.71%	95.92%	93.21%
owlv2_base_patch16_ensemble	88.04%	95.71%	92.71%
owlv2_base_patch16_finetuned	87.33%	96.08%	93.58%
owlv2_large_patch14	88.62%	95.46%	93.33%
owlv2_large_patch14_ensemble	88.38%	95.83%	93.33%
owlv2_large_patch14_finetuned	86.96%	96.29%	93.25%
detr_resnet_50	95.58%	95.58%	95.58%
	External Test Set (n=2400)		
	Left top crop	Box-Center Crop	No Adjustment
human_anotated	62.21%	72.38%	90.25%
grounding_dino_tiny	57.33%	70.08%	86.17%
grounding_dino_base	60.62%	70.29%	87.42%
omdet_turbo_swin_tiny_hf	63.17%	71.17%	88.17%
owlvit_base_patch16	77.96%	79.96%	83.92%
owlvit_base_patch32	67.17%	75.08%	84.88%
owlvit_large_patch14	61.79%	73.04%	88.67%
owlv2_base_patch16	64.12%	73.62%	88.29%
owlv2_base_patch16_ensemble	64.88%	73.83%	88.58%
owlv2_base_patch16_finetuned	66.12%	74.50%	87.12%
owlv2_large_patch14	64.21%	73.96%	89.00%
owlv2_large_patch14_ensemble	64.04%	73.12%	88.83%
owlv2_large_patch14_finetuned	64.33%	73.29%	89.75%
detr_resnet_50	83.29%	83.29%	83.29%

[a] The training and validation sets originally contained 30,573 images, of which 12,000 were randomly sampled within each group to ensure even distribution. The external set included 2,400 images, all of which were used without adjusting for group proportions.

ensures that the detected object remains within the cropped region; however, it may occasionally miss key features if the object is not optimally positioned.

- Box-Center Crop: A custom proportional cropping method was developed to dynamically adjust the crop size based on the image dimensions. Using the Center Crop Proportional class, the image was cropped around its center in a randomly selected proportion (e.g., 40% to 60% of the original size). This approach provides flexibility and preserves more contextual information when a snake occupies a smaller portion of the image. The implementation leverages the Python Imaging Library (PIL) to calculate and apply the crop based on the specified proportions.
- No Adjustment: When object detection was unavailable or failed to identify a region, the original image was directly resized to 224×224 pixels without additional cropping. Although this method simplifies preprocessing, it risks irrelevant background information, potentially affecting classification accuracy.

2.6 Article Writing

This manuscript was initially written in a mix of Chinese and English (especially for technical terms) and was later translated using ChatGPT-4. The first author promptly reviewed the translation for accuracy and clarity, making necessary revisions.

3 RESULTS

3.1 Model Training Performance

The swine transformer V2 model was fine-tuned using transfer learning on the training dataset over 20 epochs. The highest validation accuracy of 95.58% was achieved at epoch 7, corresponding to

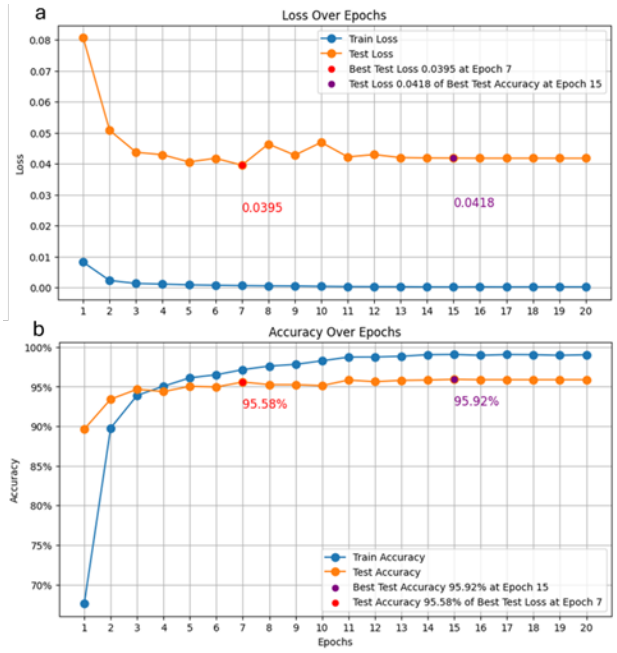


Figure 1: Transfer Learning with the Swin Transformer: (a) Loss and (b) Accuracy.

the lowest test loss of 0.0395 (Fig. 1(a)). This suggests efficient optimization in the early stages of training while maintaining strong generalization to the validation dataset.

3.2 Loss and Accuracy Trends

The loss trends during training and validation, as shown in Fig. 1(a), revealed a steady decrease in training loss, which stabilized after the initial epochs. In contrast, the test loss exhibited minor fluctuations, reflecting the model's adaptation to the validation dataset, with the lowest test loss observed at epoch 7. Notably, at epoch 15, the model achieved its highest test accuracy of 95.92% despite a slightly higher test loss of 0.0418.

The accuracy trends over epochs (Fig. 1(b)) demonstrated a rapid increase in both training and validation accuracy during the early epochs, converging after epoch 10. From epoch 7 onward, the model consistently maintained a test accuracy above 95%, highlighting its reliability and strong performance on the validation dataset.

3.3 Real-World Deployment and Dataset Creation

To evaluate the real-world applicability of the model, it was deployed through a LINE chatbot, which attracted more than 4,000 users. Between November 2023 and April 2024, 730 unique images.

Identifiable images were submitted after filtering out duplicates and unidentifiable images. Additionally, 1,670 images were collected from Facebook groups, resulting in a 2,400-image external test set representing real-world scenarios. The initial evaluation of this dataset revealed an accuracy of 83.29%, which was significantly lower than the validation accuracy.

3.4 Analysis of Misclassifications

Two main factors contributed to the reduced accuracy of the external test set:

3.4.1 Underrepresentation of Rare Variants: Some species, such as the cyan-blue variants of *Trimeresurus stejnegeri*, were not included in the training or validation datasets. These rare variants differ significantly from the typical green-bodied, red-tailed specimens, posing challenges for accurate classification.

3.4.2 Small Snake Proportions: Many images featured snakes occupying only a small portion of the frame, which lowered the model's classification accuracy (Fig. 2a).

3.5 Preprocessing with Test-Time Object Detection and Cropping

To address these challenges, a Test-Time Object Detection and Cropping method was implemented. The OWLv2 zero-shot object detection model, which required no prior training on snake-specific datasets, was used to identify snake regions in the images. Of the 2,400 external test images, 2,289 (95.38%) successfully had snake regions detected, which were then cropped and passed to the classification model. For the remaining images, in which no snake regions were detected, the original images were used directly for classification.

This preprocessing method improved the external test set accuracy from 83.29% to 89.75%, while maintaining the validation set accuracy at 96.29%. Table 2 highlights the performance improvement achieved by applying this method, with accuracy approaching the human-annotated cropping baseline of 90.25%. Fig. 2b presents examples of the detected and cropped images used in this process.

3.6 Cropping Methods and Model Performance

Table 2 presents the accuracy results of various cropping methods applied during test-time processing, including the top-left crop, box-center crop, and no adjustment. These accuracies were evaluated across different object detection models used to identify and crop snake regions.

The external dataset processed with the *owlv2_large_patch14* fine-tuned model achieved the highest overall performance, with accuracies of 64.33% for top-left crops, 73.29% for box-center crops, and 89.75% for no adjustment. This performance closely matches the human-annotated cropping baseline of 90.25%, demonstrating its robustness and practical utility for real-world image processing.

Some models, such as *owlvit_base_patch16*, performed well in specific scenarios but were less consistent across diverse test cases. These findings highlight the importance of selecting appropriate object detection models and cropping methods to maximize classification accuracy.

3.7 Impact Assessment of Preprocessing

To further assess the effectiveness of Test-Time Object Detection and Cropping, confusion matrices were generated to compare classification accuracy before and after preprocessing (Fig. 3). The results demonstrated significant improvements across nearly all snake species, confirming the proposed method's ability to enhance

Figure 2: Examples of Cropping Preprocessing. (a) Original images from the external test set; (b) images after object detection, cropping, and resizing.

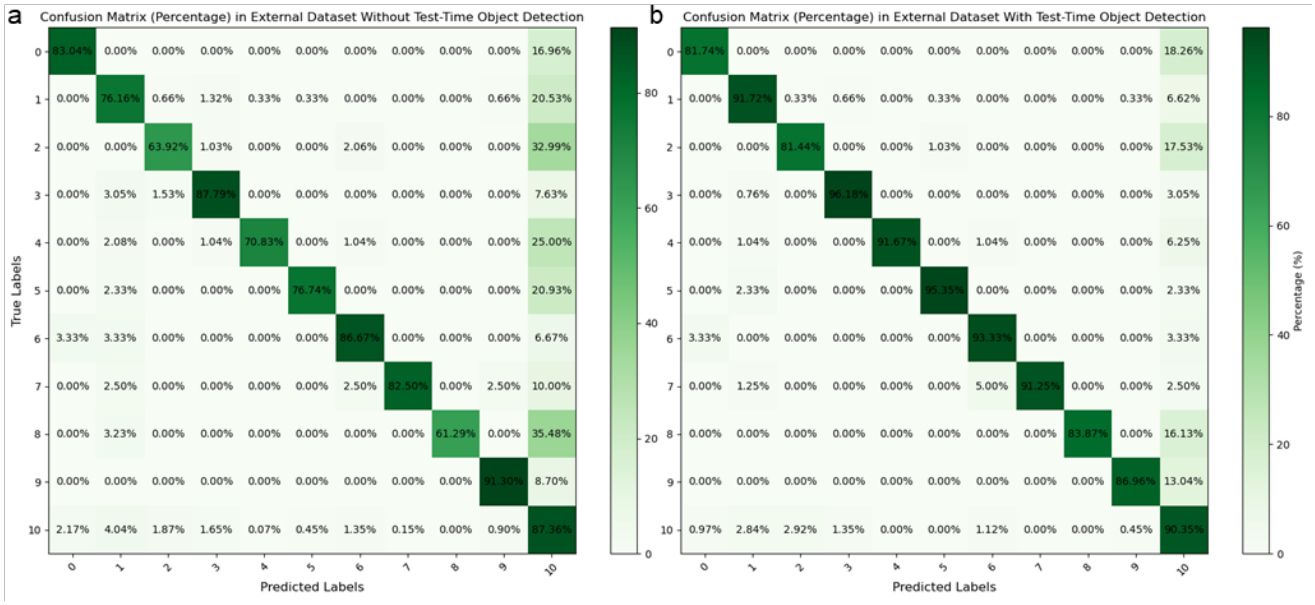


Figure 3: Confusion Matrices Showing Classification Accuracy in the External Test Set. (a) Without Test-Time Object Detection and Cropping, and (b) With Test-Time Object Detection and Cropping. Each matrix shows the percentage of correct and incorrect classifications. Species labels: 0, *Trimeresurus stejnegeri*; 1, *Protobothrops mucrosquamatus*; 2, *Naja atra*; 3, *Bungarus multicinctus*; 4, *Deinagkistrodon acutus*; 5, *Daboia siamensis*; 6, *Trimeresurus gracilis*; 7, *Ovophis makazayazaya*; 8, *Sinomicrurus* spp.; 9, *Rhabdophis formosanus*; 10, other.

model performance on challenging real-world datasets. This highlights the potential of preprocessing techniques in bridging the gap between controlled experimental conditions and real-world applications.

4 DISCUSSION

4.1 Impact of Test-Time Object Detection and Cropping

This study underscores the importance of preprocessing techniques in enhancing model performance for real-world applications. While traditional training and validation workflows prioritize achieving high accuracy within controlled datasets, real-world deployment often exposes performance gaps due to variations in image quality, framing, and object proportions. By introducing Test-Time Object Detection and Cropping, this study demonstrated a tangible improvement in classification accuracy for real-world images, increasing from 83.29% to 89.75%. This advancement highlights the potential of object detection and preprocessing methods in bridging the gap between experimental conditions and practical challenges.

The integration of the OWLv2 zero-shot object detection model enabled effective identification and cropping of snake regions without requiring additional training on snake-specific datasets. This adaptability makes the approach suitable for diverse scenarios and datasets. By refining the input to the classification model, the

preprocessing method not only improves accuracy but also mitigates the influence of irrelevant image features, such as background clutter or poorly framed subjects.

These findings establish Test-Time Object Detection and Cropping as a valuable tool for overcoming the challenges associated with real-world deployment, particularly in applications where user-submitted images exhibit significant variability in quality and framing. This approach sets the study apart from others that primarily emphasize achieving high validation accuracy within controlled datasets, demonstrating its broader applicability in practical settings.

4.2 Clinical and Real-World Applicability

A key strength of this study lies in its emphasis on clinical and real-world applicability rather than solely optimizing validation accuracy. Deploying the model on a LINE chatbot allowed for performance evaluation on user-submitted images, which often presented challenging conditions, such as uncommon snake variants and poorly framed subjects. By addressing these limitations through preprocessing, this study demonstrated a practical approach to enhancing the utility of DL models in real-world applications.

Although the model's validation accuracy of 96.29% across 11 categories is comparable to existing benchmarks such as SnakeCLEF competitions, the true value of this study lies in its emphasis on improving the performance in realistic scenarios [9, 10]. This approach aligns with the broader goal of enhancing clinical decision making, particularly in identifying medically significant venomous snakes.

4.3 Limitations and Future Directions

Although the Test-Time Object Detection and Cropping methods showed promising results, several limitations remain. First, the study relied on a single external test set sourced from the LINE chatbot and Facebook groups. To validate the broader applicability of this method, additional datasets from other real-world or open sources are required. Future research should focus on collecting diverse real-world datasets to compare performance across various scenarios and further establish the generalizability of the method.

Secondly, as shown in Fig. 3, the model performance remained suboptimal for specific species, including T. *stejnegeri*, *N. atra*, *Sinomicrurus* spp., and *R. formosanus*. The reduced accuracy for *T. stejnegeri* and *N. atra* likely reflects the significant variability in coloration and patterns, which were underrepresented in the training data. For *Sinomicrurus* spp. and *R. formosanus*, limited training images contributed to the lower accuracy. Addressing these issues requires expanding the dataset to include a more diverse range of images of these species.

Lastly, a notable limitation of snakebite envenomation cases is the frequent absence of images of the causative snake. In such situations, alternative methods, such as wound image classification, may be critical. However, the collection and accurate labeling of wound images present significant logistical challenges, highlighting an important direction for future research.

5 CONCLUSION

This study highlighted the importance of addressing real-world challenges when deploying DL models for practical applications. Although achieving high validation accuracy remains essential, bridging the performance gap between experimental datasets and real-world scenarios is critical for clinical and public utility. By introducing the Test-Time Object Detection and Cropping method, we demonstrated a significant improvement in model accuracy on user-submitted images, increasing from 83.29% to 89.75%. This approach effectively mitigated issues related to image quality, framing, and underrepresented snake variants, demonstrating its potential as a robust preprocessing solution. Despite its limitations, including the need for broader dataset validation and challenges with specific species, this study provides a foundation for future research aimed at enhancing the reliability and applicability of AI models in real-world and clinical settings.

Acknowledgments

Po-Chun Chuang thanks Chih-Hsiang Hung for assistance with computer hardware issues. This work was supported by the National Science and Technology Council (grant numbers NSTC-113-2222-E-182A-001-MY2).

References

[1] G. S. Collins, P. Dhiman, J. Ma, *et al.* 2024. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 384 (Jan. 2024), e074819. https://doi.org/10.1136/bmj-2023-074819

[2] A. Kasturiratne, A. R. Wickremasinghe, N. de Silva, *et al.* 2008. The global burden of snakebite: a literature analysis and modelling based on regional estimates of envenoming and deaths. *PLoS Med.* 5, 11 (Nov. 2008), e218. https://doi.org/10.1371/journal.pmed.0050218

[3] World Health Organization. 2019. *Snakebite envenoming: a strategy for prevention and control.* World Health Organization, Geneva, Switzerland. Licence: CC BY-NC-SA 3.0 IGO.

[4] I. Bolon, A. M. Durso, S. Botero Mesa, *et al.* 2020. Identifying the snake: First scoping review on practices of communities and healthcare providers confronted with snakebite across the world. *PLoS One* 15, 3 (2020), e0229989. https://doi.org/10.1371/journal.pone.0229989

[5] I. Bolon, L. Picek, A. M. Durso, G. Alcoba, F. Chappuis, and R. Ruiz de Castaneda. 2022. An artificial intelligence model to identify snakes from across the world: Opportunities and challenges for global health and herpetology. *PLoS Negl Trop Dis.* 16, 8 (Aug. 2022), e0010647. https://doi.org/10.1371/journal.pntd.0010647

[6] J. Zhang, X. Chen, A. Song, and X. Li. 2023. Artificial intelligence-based snakebite identification using snake images, snakebite wound images, and other modalities of information: A systematic review. *Int. J. Med. Inform.* 173 (May 2023), 105024. https://doi.org/10.1016/j.ijmedinf.2023.105024

[7] Z. Liu, Y. Lin, Y. Cao, *et al.* 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '21).* IEEE, 10012–10022.

[8] Z. Liu, H. Hu, Y. Lin, *et al.* 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22).* IEEE, 12009–12019.

[9] F. Hu, P. Wang, Y. Li, *et al.* 2023. Watch out venomous snake species: A solution to snakeclef2023. *arXiv preprint* arXiv:2307.09748.

[10] A. Joly, L. Picek, S. Kahl, *et al.* 2024. Overview of LifeCLEF 2024: Challenges on species distribution prediction and identification. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2024).* Springer, 183–207.