A Mosquito is Worth 16x16 Larvae: Evaluation of Deep Learning Architectures for Mosquito Larvae Classification

Aswin Surya, David Backer Peral, Austin VanLoon, Akhila Rajesh

Abstract

Mosquito-borne diseases (MBDs) such as dengue virus, chikungunya virus, and West Nile virus, cause over one million deaths globally every year. Because such diseases are spread by the *Aedes* and *Culex* mosquitoes, tracking these larvae is critical to mitigating the spread of these diseases. Even as citizen science projects to obtain large mosquito image datasets continuously grow, the manual annotation of mosquito images becomes ever more time-consuming and inefficient. Previous research has seen computer vision utilized to identify mosquito species, and the Convolutional Neural Network (CNN) has become the de-facto for image classification, but these models typically require substantial computational resources. This research introduces the application of the Vision Transformer (ViT) in a comparative study to improve image classification on *Aedes* and *Culex* larvae. Through the utilization of mosquito larvae image data from the GLOBE Observer Mosquito Habitat Mapper, two ViT models, ViT-Base and CvT-13, and two CNN models, ResNet-18 and ConvNeXT, from the HuggingFace library were trained and compared to determine the most effective model to classify mosquito larvae as *Aedes, Culex,* or neither. Testing revealed that ConvNeXT obtained the greatest values across all four classification metrics. ConvNeXT is found to be a viable method for mosquito larvae image classification. Future avenues of research include creating and implementing a model specifically designed for mosquito larvae classification, combining elements of CNN and transformer architecture, based on the results of this research.

Research Question

Between vision transformers (ViTs) and convolutional neural networks (CNNs), which machine learning model is best to classify *Aedes* and *Culex* mosquito larvae to prevent mosquito-borne diseases?

Introduction

Among Earth's most deadly species, mosquitoes are among the deadliest. In fact, mosquito-borne diseases (MBDs) are responsible for at least 725,000 deaths annually (Barcelona Institute for Global Health, 2017). While MBDs have challenged humans for generations, factors like urbanization, climate change, and population growth have only exacerbated the issue (Sutherst, 2004).

MBDs are especially dangerous because mosquitoes transmit viruses easily. The three types of mosquitos are the *Aedes*, *Culex*, and *Anopheles* mosquitoes. *Aedes* and *Culex* mosquitoes are especially deadly because they can breed anywhere, not just in natural environments. For instance, the female *Aedes aegypti* mosquito can lay eggs in any moist and warm environment. In Brazil, the *Aedes aegypti* species alone started a Zika epidemic and caused 2,500 cases of microcephaly, a condition where a baby's brain has not developed properly (LaFrance, 2016). Additionally, *Culex* mosquitoes rapidly spread the West Nile virus, the leading cause of MBD in the United States (CDC, 2022).

The spread of MBDs can be prevented by classifying mosquito larvae, which are distinguishable by their siphon. Larvae classification allows health officials to track mosquito populations in an area, learn which species thrive in certain environments, and forecast the presence of invasive species to prevent outbreaks, as only certain mosquito species transmit certain viruses (Joshi, 2021). This preventative approach is important because many MBDs like dengue virus and West Nile Virus, which has become endemic in the U.S., have no vaccine or treatment (CDC, 2021; Petersen, 2017).

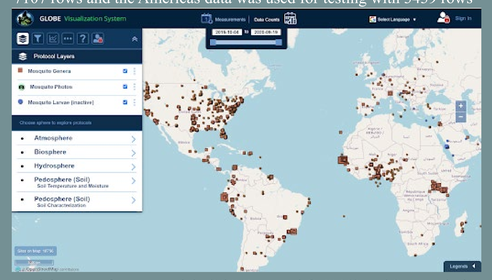
Recent research has focused on the use of artificial intelligence (AI) as an alternative to manual classification. For image classification, convolutional neural networks (CNNs) and vision transformers (ViTs) are the most common. CNNs perform convolution to locally extract features from an image and produce a feature map, from which the network can classify an image. Many past works have previously applied CNNs to identify mosquito-specific tasks. Goodwin et al. (2021) achieved an accuracy of 89.50% when identifying unknown mosquito species and 88.72% when identifying known species with CNNs. Elango et al. (2022) compared CNNs with the You-Only-Look-Once (YOLO) algorithm to predict mosquito habitats, and the YOLOv4 CNN worked best. The study concluded that CNNs were the most efficient and cost-effective approach to predict large-scale mosquito habitats, but were unable to identify small-scale habitats such as footprints, tires, or puddles.

While CNNs were the longtime state-of-the-art machine learning for image classification, Dosovitskiy et al. (2021) proposed a novel architecture that outperformed CNNs: the ViT. Rather than convolutions, the ViT uses self-attention to integrate all features of the data. Unlike CNNs, ViTs have not been extensively applied to mosquito-related tasks. Thus far, Sengar et al. (2022) reached an accuracy of 90.03% in using ViTs to predict malaria using thin blood smear microscopic images.

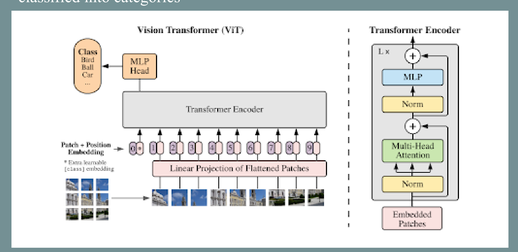
This work seeks to compare the CNN to the ViT for mosquito larvae classification to proactively prevent the spread of MBDs. This study compares four machine learning models to classify larvae as *Aedes*, *Culex*, or neither, a distinction from many previous works.

Methodology

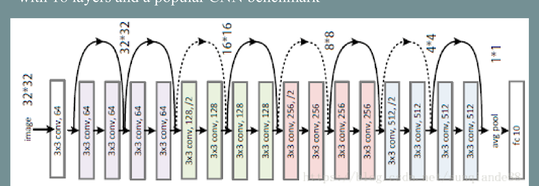
- Data was acquired through the GLOBE Mosquito Habitat Mapper database. (data collection dates ranged from May 31, 2017 to July 7, 2022, and data from Africa, North America, and Latin America was collected)
- After downloading the data using a spreadsheet builder, each image was carefully classified as *Aedes, Culex*, or neither, based on the length and shape of the siphon, color of the larvae, and amount of hair
- The spreadsheets were converted into CSV files and extraneous commas were removed to avoid blank columns
- Image links that were no longer supported were also removed using a separate script that identified incorrect HTTP codes, and null values were reviewed and validated
- After preprocessing, the Africa data was used for training and consisted of 7107 rows and the Americas data was used for testing with 3439 rows



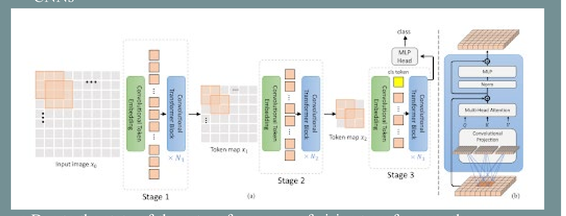
- 2 vision transformer models (ViT-Base and CvT-13) and 2 convolutional neural network (ResNet-18 and ConvNeXT) models were finetuned using the mosquito larvae image data
- The ViT-Base model works by breaking down an input image into equal-sized batches and using a multi-attention layer to capture features of the image in parallel in the relative position of each feature
- Then the image is passed into an MLP Head to understand specific details of each feature
- Finally the output is passed into the classification layer where the image is classified into categories



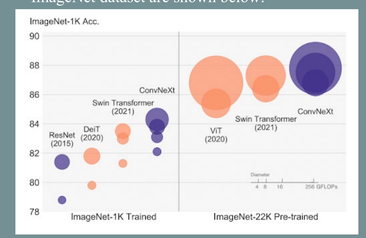
- The Residual Network (ResNet) was created to solve the issue of the vanishing gradient in conventional CNNs. This occurred when more layers were added to make the network deeper, causing the network to essentially "forget" what it was learning
- To solve this ResNets introduced skip connections between layers that allowed gradients to flow from the final layers to the initial filters, retaining information
- The ResNet model tested in this project was the ResNet-18, a ResNet with 18 layers and a popular CNN benchmark



- The Convolutional Vision Transformer (CvT) introduced convolutions to the vision transformer architecture by implementing depthwise convolutions in the transformer block, instead of the normal linear projection found in ViTs.
- A Convolutional Embedding Token also applies convolutions on patches of the image and reshapes them, increasing their depth, similar to conventional CNNs



- Due to the state-of-the-art performances of vision transformers, the ConvNeXT was created to prove that a pure convolution model could compete and outperform the latest attention-based architectures
- It was created by gradually modifying a ResNet-50 through replicating training techniques from transformers such as increasing the number of epochs, using the AdamW Optimizer, and replacing the ReLu activation function with the GELU activation
- Depthwise convolutions were also utilized in the training
- The results of the ConvNeXT compared to the ViT and ResNet on the ImageNet dataset are shown below:



- The evaluation metrics were accuracy, precision, recall, and F1 score.
- Each metric relies on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- The equations for each metric are listed below:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad Recall = \frac{TP}{TP+FN}$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad Precision = \frac{TP}{TP+FP}$$

- Each model was imported and trained using the HuggingFace Library
- A feature extractor was applied to each model and the image data was augmented with transformations
- Each classification head was altered to 3 to account for the three types of classification

```
In [7]: def transform(example_batch):
            # Take a list of PIL images and turn them to pixel values
            inputs = feature_extractor([x for x in example_batch['Photo']], return_tensors='pt')

            # Don't forget to include the labels!
            inputs['labels'] = example_batch['labels']
            return inputs

In [8]: prepared_ds = dataset.with_transform(transform)

In [9]: prepared_ds['test'].features

Out[9]: {'MHN Id': Value(dtype='float64', id=None),
         'labels': ClassLabel(num_classes=3, names=['aedes', 'culex', 'neither'], id=None),
         'Container': Value(dtype='string', id=None),
         'UserId': Value(dtype='int64', id=None),
         'latitude': Value(dtype='float64', id=None),
         'longitude': Value(dtype='float64', id=None),
         'date': Value(dtype='string', id=None),
         'Photo': Image(decode=True, id=None)}

In [10]: from transformers import default_data_collator

         data_collator = default_data_collator

In [33]: from transformers import ViTForImageClassification
         labels = dataset['train'].features['labels'].names
         #print(len(labels))
         model = ViTForImageClassification.from_pretrained(
             model_name_or_path,
             num_labels=len(labels)
         )

         model.train()
```

IVSS Badges

Data Science: This badge is being applied for due to the large amount of data collected from GLOBE Observer from 2017 to 2022 in North America, South America, and Africa. Several data preparation techniques were utilized to ensure high-quality data of mosquito larvae. We employed this data to compare state-of-the-art vision transformer models and CNN models on how well they did in classifying each image as *Aedes*, *Culex*, or neither. We also released this data to the public as a database of mosquito larvae images, which is available at https://huggingface.co/datasets/TheNoob3131/mosquito-data.

Engineer: This badge is being applied because we utilized feature extraction and data augmentation to improve the performance of our ViT-base model, ConvNeXT model, ResNet-18 model, and CvT model. This, in turn, allows us to determine which model is most capable to classify mosquito larvae species, providing a solution to scientists and researchers out in the field to quickly identify mosquito larvae and prevent potentially dangerous diseases.

Impact: This badge is being applied as a result of our models being the first to classify *Aedes*, *Culex*, or other species as larvae, as opposed to adult classification. Therefore, our research enables scientists to identify disease-carrying mosquitoes before they fully develop into adults and contributes to the active prevention of mosquito epidemics worldwide.

Results and Discussion

Model	Accuracy	Precision	Recall	F1 Score
ConvNeXT (CNN)	**0.6563**	**0.6386**	**0.6563**	**0.6355**
ResNet-18 (CNN)	0.5967	0.6034	0.5967	0.5756
CvT-13 (ViT)	0.6400	0.6292	0.6400	0.6209
ViT-Base (ViT)	0.6374	0.6061	0.6374	0.5868

The ConvNeXT scored the highest on all four classification metrics. This was likely due to the combination of standard transformer techniques with state-of-the-art CNN models like ResNet-50. Depthwise convolutions would also play a role in the robustness of the network by providing additional network width for the data to be more integrated than the other models.

All four models had similar performances, with all around 60%, but had much lower values than expected. This general finding could be due to the regional difference between the training and testing data. In the Africa train data, there were 3917 rows of *Aedes*, 2405 rows of *Culex*, and 785 rows of "neither." In the Americas test data, there were 2062 rows of *Aedes*, 1253 rows of *Culex*, and only 124 rows of "neither." The percentage of "neither" rows in the Africa data was 11.05% while the percentage in the Americas data was 3.61%. This regional difference in the number of mosquito larvae that were not *Aedes* or *Culex* might have contributed to the lack of accuracy from training to testing.

Another issue that was faced during the training of the models was overfitting. Overfitting is when a model fits exactly to its training data to the point that it almost "memorizes" the data. As a result, the model is unable to classify on new, unseen test data, resulting in poor accuracy. One reason for overfitting could be the size of the models. The ViT-Base model had 86 million trainable parameters, the CvT-13 had 19.98 million parameters, the ConvNeXT had 89 million trainable parameters, and the ResNet-18 had 11 million parameters. Adding parameters to a model increases the likelihood of overfitting because more parameters means the more closely the model will fit to the training data, as opposed to recognizing broader trends that could be used on the testing data. Additionally, overfitting could also be observed during training, as the training loss kept decreasing at a steady rate, while the evaluation loss decreased for a short amount of time before rising up and continually increasing.

Conclusion

Although all four models performed similarly for classifying mosquito larvae images as *Aedes*, *Culex*, or neither, the metric values were much lower than expected. It is likely that the models yielded lower accuracies because of the variation of regional data from the training and testing dataset. From this result, it seems that CNNs with depthwise separable convolutions perform better in complex image classification tasks than purely attention-based models. Future research could include the integration of mosquito species as opposed to genera to narrow down the features of each specific mosquito. Mapping could also be performed by correlating the mosquito species to the most likely disease it spreads. The dataset could be further expanded to higher-quality images that reveal the whole larvae body as opposed to a single part.. A novel model could also be created with the sole purpose of mosquito larvae image classification by utilizing depthwise convolutions and aspects of the transformer architecture, creating a hybrid CNN and ViT model like the CvT-13 and the ConvNeXT. During the training process of said models, a finer, more thorough analysis could be conducted using methods such as increasing the number of epochs, raising the probability of dropout, and adding specific data preparation techniques. Through this approach, epidemics can be controlled efficiently and at a rapid pace by locating and identifying potentially dangerous mosquito larvae. There are several cases where rapid mosquito larvae identification and classification are necessary for public health control. Extrapolation of key features from vision transformers and convolutional neural networks to create a more efficient model would prove as a viable, cost-effective, and autonomous approach to controlling the spread of mosquito-borne diseases.

References

1

Acknowledgements

The material contained in this poster is based upon work supported by the National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.

Special thanks to SEES Earth System Explorer mentors: Ria Jain, Dr. Russanne Low, Dr. Cassie Soeffing, Dr. Peder Nelson, Matteo Kimura, Dr. Erika Podest